PROBABILISTIC LINGUISTICS

EDITED BY **RENS BOD**,
JENNIFER HAY,
AND **STEFANIE JANNEDY**

Probabilistic Linguistics

This page intentionally left blank

Probabilistic Linguistics edited by Rens Bod, Jennifer Hay, and Stefanie Jannedy

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times New Roman on 3B2 by Asco Typesetters, Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Probabilistic linguistics / editors: Rens Bod, Jennifer Hay, Stefanie Jannedy.
p. cm.
"... originated as a symposium on 'Probability theory in linguistics' held in Washington, D.C. as part of the Linguistic Society of America meeting in January 2001"—Preface.
"Bradford books."
Includes bibliographical references and index.
ISBN 0-262-025360-1 (hc. : alk. paper)—ISBN 0-262-52338-8 (pbk. : alk. paper)
1. Linguistic analysis (Linguistics) 2. Linguistics—Statistical methods.
3. Probabilities. I. Bod, Rens, 1965– II. Hay, Jennifer. III. Jannedy, Stefanie.
P128.P73 P76 2003
410′.1′5192—dc21 2002032165

10 9 8 7 6 5 4 3 2 1

Contents

Preface

A wide variety of evidence suggests that language is probabilistic. In language comprehension and production, probabilities play a role in access, disambiguation, and generation. In learning, probability plays a role in segmentation and generalization. In phonology and morphology, probabilities play a role in acceptability judgments and alternations. And in syntax and semantics, probabilities play a role in the gradience of categories, syntactic well-formedness judgments, and interpretation. Moreover, probabilities play a key role in modeling language change and language variation.

This volume systematically investigates the probabilistic nature of language for a range of subfields of linguistics (phonology, morphology, syntax, semantics, psycholinguistics, historical linguistics, and sociolinguistics), each covered by a specialist. The probabilistic approach to the study of language may seem opposed to the categorical approach, which has dominated linguistics for over 40 years. Yet one thesis of this book is that the two apparently opposing views may in fact go very well together: while categorical approaches focus on the endpoints of distributions of linguistic phenomena, probabilistic approaches focus on the gradient middle ground.

This book originated as the symposium "Probability Theory in Linguistics," held in Washington, D.C., as part of the Linguistic Society of America meeting in January 2001. One outcome of the symposium was the observation that probability theory allows researchers to change the level of magnification when exploring theoretical and practical problems in linguistics. Another was the sense that a handbook on probabilistic linguistics, providing necessary background knowledge and covering the various subfields of language, was badly needed. We hope this book will fill that need.

We expect the book to be of interest to all students and researchers of language, whether theoretical linguists, psycholinguists, historical linguists, sociolinguists, or computational linguists. Because probability theory has not formed part of the traditional linguistics curriculum, we have included a tutorial on elementary probability theory and probabilistic grammars, which provides the background knowledge for understanding the rest of the book. In addition, a glossary of probabilistic terms is given at the end of the book.

We are most grateful to the authors, who have given maximal effort to write the overview chapters on probabilistic approaches to the various subfields of linguistics. We also thank the authors for their contribution to the review process. We are grateful to Michael Brent for his contribution to the original symposium and to Anne Mark for her excellent editorial work. Finally, we would like to thank the editor, Thomas Stone, for his encouragement and help during the processing of this book.

The editors of this book worked on three different continents (with the South Pole equidistant from us all). We recommend this as a fabulously efficient way to work. The book never slept.

Contributors

R. Harald Baayen R. Harald Baayen studied linguistics at the Free University of Amsterdam. In 1989, he completed his Ph.D. thesis on statistical and psychological aspects of morphological productivity. Since 1989, he has held postdoctoral positions at the Free University of Amsterdam and at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. He is now affiliated with the University of Nijmegen. His research interests include lexical statistics in literary and linguistic corpus-based computing, general linguistics, morphological theory, and the psycholinguistics of morphological processing in language comprehension and speech production. He has published in a variety of international journals, including *Language*, *Linguistics*, *Computers and the Humanities*, *Computational Linguistics*, *Literary and Linguistic Computing*, *Journal of Quantitative Linguistics*, *Journal of Experimental Psychology*, and *Journal of Memory and Language*.

Rens Bod Rens Bod received his Ph.D. from the University of Amsterdam. He is one of the principal architects of the Data-Oriented Parsing model, which provides a general framework for probabilistic natural language processing and which has also been applied to other perceptual modalities. He published his first scientific paper at the age of 15 and is the author of three books, including *Beyond Grammar: An Experience-Based Theory of Language*. He has also published in the fields of computational musicology, vision science, aesthetics, and philosophy of science. He is affiliated with the University of Amsterdam and the University of Leeds, where he works on spoken language processing and on unified models of linguistic, musical, and visual perception.

Ariel Cohen Ariel Cohen received his Ph.D. in computational linguistics from Carnegie Mellon University. In 1996, he joined the Department of

Foreign Literatures and Linguistics at Ben-Gurion University of the Negev, Beer Sheva, Israel. His main research interest is in formal semantics, especially the study of generics. He has also investigated adverbs of quantification, the meaning of topic and focus, plurals, coordination, default reasoning, and, of course, probability. His dissertation, ''Think Generic!'', was published (1999) by the Center for the Study of Language and Information at Stanford University.

Jennifer Hay Jennifer Hay received her Ph.D. from Northwestern University and is currently a lecturer in the Department of Linguistics, University of Canterbury, New Zealand. One strand of her current research investigates how speech-processing strategies shape linguistic (particularly morphological) representation and structure. A second focuses on sociophonetics, with special attention to New Zealand English. She has published work on language and gender, proper names, phonotactics, lexical frequency, sociophonetics, lexical semantics, morphological productivity, and humor.

Stefanie Jannedy Stefanie Jannedy received her Ph.D. from the Department of Linguistics at The Ohio State University. She currently holds a position at Lucent Technologies/Bell Laboratories working on linguistic issues as they relate to the development of multilingual text-to-speech systems. Her research interests include phonetic work on connected speech processes in Turkish, German, and English, the acquisition of contrastive emphasis in childhood, the interpretation of intonation contours in context, and topics in sociophonetics. She is an editor (with Robert Poletto and Tracey Weldon) of the sixth edition of *Language Files*, an introductory textbook on linguistics.

Dan Jurafsky Dan Jurafsky is an associate professor in the Departments of Linguistics and Computer Science and the Institute of Cognitive Science at the University of Colorado at Boulder. His research focuses on statistical models of human and machine language processing, especially automatic speech recognition and understanding, computational psycholinguistics, and natural language processing. He received the National Science Foundation CAREER award in 1998 and serves on various boards, including the editorial boards of *Computational Linguistics* and *Computer Speech and Language* and the Technical Advisory Board of Ask Jeeves, Inc. His most recent book (with James H. Martin) is the widely used textbook *Speech and Language Processing*.

Christopher Manning Christopher Manning is an assistant professor of computer science and linguistics at Stanford University. He received his Ph.D. from Stanford University in 1995 and served on the faculties of the Computational Linguistics Program at Carnegie Mellon University and the Linguistics Department at the University of Sydney before returning to Stanford. His research interests include probabilistic models of language, statistical natural language processing, constraint-based linguistic theories, syntactic typology, information extraction, text mining, and computational lexicography. He is the author of three books, including *Foundations of Statistical Natural Language Processing* (MIT Press, 1999, with Hinrich Schütze).

Norma Mendoza-Denton Norma Mendoza-Denton received her Ph.D. in linguistics from Stanford University in 1997. She is an assistant professor of linguistic anthropology at the University of Arizona in Tucson and has also taught in the Departments of Spanish and Linguistics at The Ohio State University. With primary linguistic interests and publications in sociophonetic, syntactic, and discourse variation, she works to understand the relationship between socially constructed identities (ethnicity, gender, class) and their symbolic implementation, not only in language but also in other modes of semiotic practice (such as clothing or makeup). She has done fieldwork among Japanese-American internment camp survivors, Latino youth involved in gangs in California, and American politicians in the House of Representatives in Arizona and in Washington, D.C. She is currently working on a book entitled *Homegirls: Symbolic Practices in the Articulation of Latina Youth Styles*, forthcoming from Blackwell.

Janet Pierrehumbert Janet Pierrehumbert started out studying syntax, but switched to phonology and phonetics during a summer internship at AT&T Bell Laboratories. Her MIT Ph.D. dissertation developed a model of the phonology and phonetics of English intonation. After two postdoctoral years in the Center for Cognitive Science at MIT, she joined the staff of the Bell Laboratories Department of Linguistics and Artificial Intelligence Research. She moved to Northwestern University in 1989, where she established a phonetics laboratory and started additional lines of research on segmental phonetics, lexical representation, and probabilistic models of phonology. She is now a professor of linguistics at Northwestern, and she served as chair of the Department of Linguistics from 1993 to 1996. She co-organized the Fifth Conference on Laboratory

Phonology in 1996. She has also held visiting appointments at Kungl Tekniska Högskolan in Stockholm (1987–1988) and Ecole Nationale Supérieure des Télécommunications in Paris (1996–1997, as a Fellow of the John Simon Guggenheim Foundation).

Kie Zuraw Kie Zuraw is an assistant professor of linguistics at the University of Southern California and, in 2001–2002, a visiting assistant professor of linguistics at MIT. She received a Ph.D. from the University of California, Los Angeles, in 2000. Her dissertation proposed, using detailed case studies of lexical patterns in Tagalog, a model of the learnability, representation, and use (in speaking and listening) of lexically probabilistic phonology and of how lexical patterns are perpetuated historically. Her other interests include variable and probabilistic grammars, learning algorithms, contact and loanword phonology, the selection and assimilation of new words into the lexicon, computational models of the speech community, reduplication and pseudoreduplication, and Philippine and other Austronesian languages.

Chapter 1

Introduction

Rens Bod, Jennifer Hay, and Stefanie Jannedy

1.1 Probabilistic Linguistics

One of the foundations of modern linguistics is the maxim of categoricity: language is categorical. Numbers play no role, or, where they do, they are artifacts of nonlinguistic performance factors. Thus, while it is widely recognized that real language can be highly variable, gradient, and rich in continua, many linguists would argue that the competence that underlies such ''performance factors'' consists of well-defined discrete categories and categorical grammaticality criteria. Performance may be full of fuzziness, gradience, and continua, but linguistic competence is not.

However, a groundswell of recent results challenge the idea that linguistic competence is categorical and discrete. While linguistic phenomena such as phonological and morphological alternations and syntactic well-formedness judgments tend to be modeled as categorical, it has become increasingly clear that alternations and judgments display properties of continua and show markedly gradient behavior. Moreover, psycholinguistic experiments demonstrate that speakers' well-formedness judgments of words and sentences are extremely well predicted by the combined probabilities of their subparts.

While generative approaches to linguistics have evolved to capture the endpoints of such distributions, there is growing interest in the relatively unexplored gradient middle ground, and a growing realization that concentrating on the extremes of continua leaves half the phenomena unexplored and unexplained. The chapters in this book illustrate that one need not discard the many insights of modern linguistics in order to insightfully model this middle ground. On the contrary, a probabilistic approach can push the boundaries of linguistic theory forward, by substantially enriching the current state of knowledge. Probabilistic linguistics

increases the range of data for which a theory can account, and for which it must be accountable.

1.2 Motivating Probabilities

In recent years, a strong consensus has emerged that human cognition is based on probabilistic processing. Jurafsky (this volume) outlines some recent literature, and papers documenting the probabilistic underpinnings of a wide range of cognitive processes appear in Rao, Olshausen, and Lewicki 2002. The editors of that book praise the probabilistic approach for its promise in modeling brain functioning and its ability to accurately model phenomena "from psychophysics to neurophysiology."

However, the fact that probability theory is an increasingly useful and important tool in cognitive science does not make it automatically suitable for modeling language. To be convinced of its suitability, readers should rightly demand evidence that the language faculty itself displays probabilistic properties. We briefly outline the nature of this evidence below.

1.2.1 Variation

Language changes over time—a process that is usually echoed synchronically across age groups. Zuraw provides evidence that language change can result from probabilistic inference on the part of listeners, and she argues that probabilistic reasoning "could explain the maintenance of lexical regularities over historical time" (sec. 5.5.1).

It is well accepted that language does not just vary across time—it is inherently variable. There is no known case, for example, where analogous phonemes have exactly the same implementation across two languages (Pierrehumbert).

Acquiring a language or dialect, then, involves not just identifying its phonemes, but also learning the extremely subtle patterns of production and allophony relevant to each phoneme in that language. Within a particular language, production patterns differ across individuals, depending on aspects of identity (Mendoza-Denton, Hay, and Jannedy). Within individuals, production patterns differ on the basis of stylistic factors such as addressee, context, and topic, and this stylistic variation to a large degree echoes the variation present across members of society. Knowledge of variation, then, must form part of linguistic competence, since individuals can manipulate their implementation of phonetic variants to

portray linguistic and extralinguistic information. And individuals differ not only in the specific variants they use in different contexts, but also in the frequency with which they use them. Knowledge of variation must involve knowledge of frequencies (Mendoza-Denton, Hay, and Jannedy). And this, as it turns out, does not set it apart from other types of linguistic knowledge.

1.2.2 Frequency

One striking clue to the importance of probabilities in language comes from the wealth of frequency effects that pervade language representation, processing, and language change.

The chapters in this book document many ways in which frequency permeates language. Frequent words are recognized faster than infrequent words, and there is a bias toward interpreting ambiguous words in terms of their more frequent meanings (Jurafsky). Frequent words lead leniting changes (Zuraw) and are more prone to reduction in speech (Jurafsky; Mendoza-Denton, Hay, and Jannedy). Frequent combinations of phonemes (Pierrehumbert) and structures (Manning) are perceived as more grammatical, or well formed, than infrequent combinations. The relative frequency of derived words and their bases affects the morphological decomposability of complex words (Baayen). These are just a few of the many frequency effects discussed in this book that influence language perception, production, and representation.

Frequency affects language processes, and so it must be represented somewhere. The language-processing system tracks, records, and exploits frequencies of various kinds of events.

We can best model many of these effects by making explicit the link between frequency and probability. Probability theory provides well-articulated methods for modeling frequency, and it provides researchers with the tools to work not only with the frequency of events, but also with the frequency of *combinations* of events. One can thus estimate the probability of complex events (such as sentences) by combining the probabilities of their subparts.

The presence of frequency effects is not in itself sufficient to warrant adopting a probabilistic view. It is conceivable that at least some of the frequency effects outlined in this book could occur without any kind of probabilistic effect. However, the presence of frequency effects does provide evidence that the basic building blocks of probability theory are stored and exploited. Just as the complete absence of frequency effects

would challenge the foundations of probabilistic linguistics, so their overwhelming presence adds weight to the claim that the language faculty is inherently probabilistic.

1.2.3 Gradience

Frequency effects provide one type of evidence for a probabilistic linguistics. A stronger type of evidence comes from gradience. The chapters in this book are filled with examples of continua and gradience. Here, we outline just a few of these cases—phenomena that at first glance may appear categorical, but upon closer inspection show clear signs of gradience. And probabilities are extremely well suited to capturing the notion of gradience, as they lie in a continuum between 0 (reflecting impossibility) and 1 (reflecting certainty).

1.2.3.1 Category Membership Pierrehumbert argues that phoneme membership is gradient, with phonemes representing continuous probability distributions over phonetic space. Items that are central in such a distribution are good examples of a particular phoneme; more peripheral items are more marginal as members. And distributions may overlap.

Manning suggests that such an approach may also be appropriate for modeling syntactic category membership, which also displays properties of gradience. As a case study, he examines "marginal prepositions" such as *concerning*, *considering*, and *following*. He convincingly demonstrates the gradient behavior of this class, which ranges from fully verbal to fully prepositional, arguing that "it seems that it would be useful to explore modeling words as moving in a continuous space of syntactic category, with dense groupings corresponding to traditional parts of speech" (sec. 8.4).

Categories are central to linguistic theory, but membership in these categories need not be categorical. Probabilistic linguistics conceptualizes categories as distributions. Membership in categories is gradient.

1.2.3.2 Well-Formedness Manning illustrates that, in corpus-based searches, there is no well-defined distinction between sentences generally regarded as "grammatical" in the literature, and those regarded as ungrammatical. Rather, what we see is a cline of well-formedness, wherein some constructions are highly preferred, others are used less frequently, and some are used not at all. The distinction drawn between grammatical and ungrammatical is often somewhere in the middle of the cline, ruling

out those constructions that tend to be less frequent as "ungrammatical." However, nowhere in the cline is there a dramatic drop in frequency; in fact, the cline can often be gradual, so that the decision where to draw the distinction is relatively arbitrary. The difficulty of drawing such lines has led to special notation in formal syntax, to represent questionable grammatical status (the question mark,?). But this middle territory has seldom been the object of theory building, nor has it been incorporated into formal models of syntax. Probabilistic linguistics seeks to account for the full continuum between grammaticality and ungrammaticality.

The gradualness observed in corpus searches is also echoed in grammaticality judgments: speakers do not find it a strange task to rate degrees of acceptability or grammaticality, as we might expect if grammaticality were categorical, rather than gradient.

Similarly, in the realm of phonology, Pierrehumbert summarizes compelling evidence that the judged well-formedness of novel words is incontrovertibly gradient and can be predicted as a function of the probability of the words' subparts.

1.2.3.3 Morphological Productivity It is widely accepted that some affixes are productive and can give rise to new words, whereas others are unproductive—present in extant words, but not available for further word formation. However, as Baayen discusses, not all affixes are equally productive. Some word formation rules give rise to very few words, whereas others are highly productive and spawn many new words. Morphological productivity is a clearly gradient phenomenon. Understanding and accurately modeling it, then, requires a theory of linguistics that can predict degrees of productivity. Drawing a simple categorical distinction between "productive" and "unproductive" is relatively stipulative and captures only a small proportion of the facts.

1.2.3.4 Morphological Decomposition As both Baayen and Pierrehumbert discuss, word formation is not the only morphological process that exhibits symptoms of gradience. Both authors summarize evidence that morpheme boundaries, the very essence of morphology, are gradient—that is, stronger in some complex words than in others. This gradience arises from the role of decomposition in speech perception: complex words that are often decomposed are represented with strong morphological boundaries, those that are seldom decomposed come to be represented with weak ones. Crucially, and as with all other examples

discussed in this section, this gradience is not a simple matter of performance—it has deep linguistic consequences.

1.2.3.5 The Argument/Adjunct Distinction Even syntactic roles may be gradient. Manning argues against a categorical conception of the argument/adjunct distinction, citing documented difficulties with cleanly dividing verbal dependents into freely occurring adjuncts and subcategorized arguments. He suggests that one possibility for modeling the observed gradience is to represent subcategorization information as "a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability" (sec. 8.3.1).

1.2.4 Acquisition

As outlined above, there is a wide range of evidence for gradience and gradient effects in language. Modeling all such factors as artifacts of "performance" would be a massive challenge and would likely constitute serious hoop-jumping. One common reason for wanting to do so stems from skepticism regarding the mind's ability to acquire and store a complex range of generalizations and frequencies. However, the chapters in this book argue that adding probabilities to linguistics in fact makes the acquisition problem easier, not harder.

As Gold (1967) demonstrated, formal languages cannot be learned without negative evidence. Moreover, negative evidence is not readily available to children. Together, these two facts are widely used as evidence that language is special and largely innate, a line of reasoning known as the "argument from the poverty of the stimulus." Manning outlines evidence that challenges this argument—most importantly, evidence (dating from Horning 1969) that, unlike categorical grammars, probabilistic grammars *are* learnable from positive evidence alone.

As outlined by Pierrehumbert, generalizations based on statistical inference become increasingly robust as sample size increases. This holds for both positive and negative generalizations: as the range and quantity of data increase, statistical models are able to acquire negative evidence with increasing certainty. Pierrehumbert also outlines several types of results relating to the acquisition of phonemes and phonological generalizations, which together provide strong evidence that acquisition involves the continual updating of probability distributions.

Many current models of language acquisition rely on probabilistic models, and considerable evidence demonstrates that infants track probabilities in order to tackle such difficult tasks as decomposing a speech stream into words (Goodsitt, Morgan, and Kuhl 1993; Saffran, Newport, and Aslin 1996a,b) and even into phrases (Saffran 2001). It is certainly not the case that the use of probabilities complicates the learning task. On the contrary, if the language faculty is probabilistic, the learning task is considerably more achievable. Variability and continuity both enhance learning.

1.2.5 Universals

Many phenomena or constraints are present in a great many languages, reflecting universal tendencies of the language faculty. They are operative to greater or lesser degrees in different languages and in some cases are highly grammaticalized and categorical. Manning discusses one such case in depth: the interaction of passive, person, and topicality. A categorical formal framework does not enable us to fully capture the different degrees to which constraints are operative in different languages. By contrast, probabilistic linguistics does enable us to formally model such situations, capturing both the ways in which languages are similar (operating under similar constraints) and the ways in which they differ (the probabilities associated with those constraints).

1.3 Probabilities Where?

Clearly, there is a need to integrate probabilities into linguistics—but where? Taken together, the chapters in this book answer, "Everywhere." Probabilities are operative in acquisition (see, e.g., Manning; Pierrehumbert), perception (Zuraw; Baayen; Jurafsky), and production (Pierrehumbert; Baayen; Jurafsky). Moreover, they are not merely a tool for processing: linguistic representations are probabilistic (see, e.g., Pierrehumbert; Baayen; Mendoza-Denton, Hay, and Jannedy), as are linguistic constraints and well-formedness rules (Manning; Bod). Probabilities permeate the linguistic system.

Probabilities are relevant at multiple levels of representation (Pierrehumbert) and can be calculated over arbitrarily complex, abstract representations (Manning; Jurafsky). As Manning discusses, it is a common misconception that probabilities can be recorded only over surface

structure; indeed, there is no barrier to calculating probabilities over hidden structure. Probabilistic linguistics does not abandon all the progress made by linguistics thus far; on the contrary, it integrates this knowledge with a probabilistic perspective.

As Baayen and Pierrehumbert argue, probabilities of both types and tokens play an important role. For example, the number of different words a speaker has encountered containing a particular affix is important (the types), as is the number of times the speaker has encountered each of those words (tokens).

As Pierrehumbert discusses, linguistic constraints consist of statistically robust generalizations. There are many theoretically possible constraints that could be operative in language, but those that are effectively learned, transmitted, and exploited are those that are statistically robust: they can be learned from limited language exposure, and they can be successfully learned by different individuals exposed to language in different ways and to different extents. The robust generalizations are the linguistically important ones.

Here we briefly review some of the many levels of representation that show probabilistic properties.

As noted above, phonemes are probabilistic distributions over a continuous phonetic space (Pierrehumbert). Learning phonemes, and classifying phonetic exemplars as specific phonemes, requires situating them within the appropriate region in this phonetic space. Phoneme membership is probabilistic.

Knowledge of phonotactics involves knowledge of co-occurrence probabilities of phonemes. The well-formedness of a string of phonemes is a function of "the frequency of the subparts and the specific way in which they were combined" (Pierrehumbert, sec. 6.2). Such phonotactic probabilities are exploited in speech perception for segmentation, and they affect well-formedness judgments, influence pronunciation, and affect behavior in linguistic tasks such as creating blends. Phonotactics is probabilistic.

Probabilities are also operative at the morpheme level. Some affixes are much more productive than others; that is, probability of use varies, and forms part of the speaker's linguistic knowledge. Individuals' choice among competing affixes shows a strong bias toward the most probable one, as measured by patterns of occurrence in related words (Baayen). Affix choice is probabilistic.

The processing and representation of words is strongly influenced by lexical frequency: more probable and less probable words behave differently. This is true both of morphologically simple and of morphologically complex words. The many realms in which word frequency manifests itself include ambiguity resolution (Jurafsky), phoneme reduction (Jurafsky; Mendoza-Denton, Hay, and Jannedy; Pierrehumbert), language change (Zuraw), and speed of access (Jurafsky). Word representations are probabilistic.

Relationships between words also exhibit linguistically relevant probabilities. The larger the number of word pairs that instantiate a generalization (or word sets that instantiate a paradigm), the more robust that generalization is. Generalizations that are represented by a great many words pairs tend to be highly salient and productive (Pierrehumbert). Morphophonological relations between words are probabilistic.

Individuals also track co-occurrence probabilities of words (Jurafsky). In comprehension, these influence processing time. In production, high-frequency (or high-probability) word pairs are more phonetically reduced. Low-probability words (given the probability of surrounding words) are more likely to attract a pitch accent. Word combinations are probabilistic.

Verbs take different subcategorization frames with different frequencies. The probability that a specific verb will take various specific subcategorization frames affects ambiguity resolution (Jurafsky). Moreover, there is evidence that subcategorization displays properties of a continuum (Manning). Syntactic subcategorization is probabilistic.

Jurafsky provides evidence that people track the probabilities of syntactic structures. Frequently encountered sentences or sentence fragments are more easily processed than infrequently encountered ones, even controlling for lexical frequency and other relevant factors. And listeners and readers are influenced by the likelihood of a specific structure or word given previously encountered structure. This effect influences processing time and is involved in disambiguation. Bod and Manning discuss methods for the probabilistic combination of syntactic subtrees. Sentence structure is probabilistic.

Cohen discusses cases in which supplementing truth-conditional semantics with probability theory increases the explanatory power of the model. These include the modeling of generics, frequency adverbs, conditionals, and vague terms. He demonstrates clearly that the semantics of

such words are probabilistic. He concludes by discussing prospects for a fully probabilistic semantics, in which judgments of truth conditions are replaced by judgments of probability. In such a semantics, the meaning of a sentence would not be a function from possible worlds to truth values; rather, it would be "a function from sets of possible worlds to probabilities" (sec. 9.8). Such a theory, Cohen argues, would formally capture the idea that "understanding the meaning of a sentence is the ability, given a situation, to assess its probability." Semantics too, then, may be probabilistic.

In short, practically every level of representation provides robust evidence for the involvement of probabilities.

1.4 Conclusion

Language displays all the hallmarks of a probabilistic system. Categories and well-formedness are gradient, and frequency effects are everywhere. We believe all evidence points to a probabilistic language faculty. Knowledge of language should be understood not as a minimal set of categorical rules or constraints, but as a (possibly redundant) set of gradient rules, which may be characterized by a statistical distribution.

Chapter 2

Introduction to Elementary Probability Theory and Formal Stochastic Language Theory

Rens Bod

2.1 Introduction

For a book on probabilistic approaches to a scientific discipline, it may seem unnecessary to start with an introduction to probability theory. The reader interested in probabilistic approaches would usually have a working knowledge of probability theory and would directly read the more specialized papers. However, the situation is somewhat different for linguistics. Since probability theory does not form part of a traditional linguistics curriculum, probabilistic linguistics may not be as accessible as some other areas. This is further reinforced by the disciplinary gap between probabilistic and categorical approaches, the first being dominant in psycholinguistics and natural language processing, the second in generative linguistics. One goal of this book is to show that these two apparently opposing methodologies go very well together: while categorical approaches focus on the endpoints of distributions of linguistic phenomena, probabilistic approaches focus on the gradient middle ground. That linguistic phenomena *are* gradient will not be discussed here, as this is extensively shown in the other chapters. But to make these chapters accessible to the linguistics community at large, there is a need to explain the most important concepts from probability theory first. Any additional concept that may be encountered later can be looked up in the glossary. I will only assume that the reader has some elementary knowledge of set theory (see Partee, ter Meulen, and Wall 1990 for a linguistic introduction).

After a brief introduction to the basics of probability theory, I will show how this working knowledge can be put into practice by developing the concept of *probabilistic grammar*, which lies at the heart of probabilistic linguistics. Since many different probabilistic grammars have been

proposed in the literature, there is a need for a theory that creates some order among them, just as *Formal Language Theory* creates order among nonprobabilistic grammars. While I will only scratch the surface of a *Formal Stochastic Language Theory*, I will show that probabilistic grammars evoke their own stochastic hierarchies.[1]

2.2 What Are Probabilities?

Historically, there have been two interpretations of probabilities: *objectivist* and *subjectivist*. According to the objectivist interpretation, probabilities are real aspects of the world that can be measured by *relative frequencies* of outcomes of experiments. The subjectivist view, on the other hand, interprets probabilities as *degrees of belief* or *uncertainty* of an observer rather than as having any external significance. These two contrasting interpretations are also referred to as *frequentist* versus *Bayesian* (from Thomas Bayes, 1764). Whichever interpretation one prefers, probabilities are numbers between 0 and 1, where 0 indicates impossibility and 1 certainty (percentages between 0% and 100% are also used, though less commonly).

While the subjectivist relies on an observer's judgment of a probability, the objectivist measures a probability through an *experiment* or *trial*—the process by which an observation is made. The collection of *outcomes* or *sample points* for an experiment is usually referred to as the *sample space* Ω. An *event* is defined as any subset of Ω. In other words, an event may be any set of outcomes that result from an experiment. Under the assumption that all outcomes for an experiment are equally likely, the probability P of an event A can be defined as the ratio between the size of A and the size of the sample space Ω. Let $|A|$ be the number of elements in a set A; then

$$P(A) = \frac{|A|}{|\Omega|}. \tag{1}$$

To start with a simple, nonlinguistic example, assume a fair die that is thrown once. What is the chance of obtaining an even number? The sample space of this trial is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

The event of interest is the subset containing all even outcomes. Let us refer to this event as A:

$A = \{2, 4, 6\}$.

Thus, the number of elements in A is 3, and the number of elements in Ω is 6; that is, $|A| = 3$ and $|\Omega| = 6$. Then the probability of A is

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = .5.$$

Let us now turn to a slightly more linguistic example. Assume a small corpus consisting of 50 unambiguous words of which 25 are nouns, 20 are verbs, and 5 are adjectives. Consider the experiment of randomly selecting a word W from this corpus. What is the probability of selecting a verb? The sample space Ω of this trial is the set of all words in the corpus. The event of interest A is the set of verbs, which we may write as $\{W\text{: } W \text{ is a verb}\}$. So,

$$P(A) = \frac{|A|}{|\Omega|} = \frac{|\{W\text{: } W \text{ is a verb}\}|}{|\Omega|} = \frac{20}{50} = .4.$$

For the sake of brevity, we will often write $P(\{\text{verb}\})$ instead of $P(\{W\text{: } W \text{ is a verb}\})$. Thus,

$$P(\{\text{verb}\}) = \frac{|\{\text{verb}\}|}{|\Omega|} = \frac{20}{50} = .4,$$

$$P(\{\text{noun}\}) = \frac{|\{\text{noun}\}|}{|\Omega|} = \frac{25}{50} = .5,$$

$$P(\{\text{adjective}\}) = \frac{|\{\text{adjective}\}|}{|\Omega|} = \frac{5}{50} = .1.$$

Two important observations can now be made. First, note that the probability of selecting either a verb or a noun or an adjective is equal to 1, since in that case the event of interest A is $\{W\text{: } W \text{ is any word}\}$, which is equal to the sample space Ω, and thus $P(A) = |\Omega|/|\Omega| = 1$. This corresponds to the intuition that the probability that something will be sampled in this experiment is equal to 1.

Second, note that the *sum* of the probabilities of each event, {verb}, {noun}, and {adjective}, is also equal to 1; that is, $.4 + .5 + .1 = 1$. If events do not overlap, the probability of sampling either of them is equal to the sum of their probabilities. This is known as the *sum rule*. For example, the probability of selecting either a verb or a noun, usually written as $P(\{\text{verb}\} \cup \{\text{noun}\})$ or $P(\{\text{verb}, \text{noun}\})$, is equal to $45/50 = .9$, which is also equal to the sum $P(\{\text{verb}\}) + P(\{\text{noun}\}) = .4 + .5 = .9$.

It is important to note that the event $\{\text{verb}, \text{noun}\}$ does *not* refer to the event of a word being in the class of words that can be both a noun *and* a verb. As defined above, events are subsets of the sample space, and $\{\text{verb}, \text{noun}\}$ denotes the event of either a noun occurring or a verb occurring.

These two properties just noted are actually the rules a so-called *probability function* should obey (in addition to the fact that it should range over [0, 1]). The first rule says that a trial will always produce an event in the event space. That is, the probability that something in the event space will happen—namely, $P(\Omega)$—is 1:

$$P(\Omega) = 1. \tag{2}$$

The second rule says that if two or more events do not overlap, the probability that either event occurs is equal to the sum of their probabilities. That is, for two disjoint events A and B,

$$P(A \cup B) = P(A) + P(B). \tag{3}$$

As long as these rules hold, P is a probability function, also known as a *probability distribution*. (There are some well-studied probability distributions that appear later in this book, such as the binomial distribution and the normal distribution. See the glossary for definitions.)

Note that rule (3) can be generalized to any number of events. That is, for n disjoint events $A_1, A_2, \ldots, A_n$,

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n). \tag{4}$$

The right-hand side of this sum rule is often conveniently abbreviated by the sum sign Σ:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \Sigma_i P(A_i). \tag{5}$$

Recall that under the frequentist interpretation, the probability of an event is interpreted as its *relative frequency* in a series of experiments. A classical result from statistics shows that the relative frequency of an event converges to its true probability as the number of experiments increases (*Law of Large Numbers*). Thus, if x is an outcome of some experiment (e.g., throwing a die) and $Count(x)$ is the number of times x occurs in N repeated experiments, then the relative frequency $Count(x)/N$ converges to the probability of x if N goes to infinity. The probability of x is also written as $P(X = x)$, where X is called a *random variable* (see also the glossary).

2.3 Joint Probabilities and Conditional Probabilities

Let us now extend our notion of simple probability to that of *joint* probability. Joint probabilities are useful if we are interested in events that contain more than one outcome. For example, in an experiment where we randomly sample *two* words from the corpus described in section 2.1 (rather than just *one* word), what is the probability of an event consisting of a noun and a verb—given that we sample with replacement?[2] We write this probability as $P(\{\text{noun}\} \cap \{\text{verb}\})$, or simply as $P(\{\text{noun}\}, \{\text{verb}\})$. We already computed the probabilities of sampling a noun and a verb separately:

$P(\{\text{noun}\}) = .5,$

$P(\{\text{verb}\}) = .4.$

Intuitively, this amounts to saying that in 50% of the cases we sample a noun, after which in 40% of the cases we sample a verb. This means that we sample them jointly in 40% of 50% of the cases—that is, in 20% of the cases (in our experiment). Thus, intuitively, the joint probability of sampling a noun and a verb is equal to the product of the probabilities of sampling them separately: $P(\{\text{noun}\}, \{\text{verb}\}) = P(\{\text{noun}\}) \times P(\{\text{verb}\}) = .5 \times .4 = .2$.[3] We can do this simple multiplication because we designed our experiment in such a way that sampling a verb is independent of having sampled a noun.[4] We say that the events {noun} and {verb} are *independent*. In general, for two independent events A and B,

$$P(A, B) = P(A) \times P(B) \quad \text{if } A \text{ and B are independent.} \tag{6}$$

It is often the case that two events are not independent, but *dependent*. We could design an experiment where the probability of sampling a verb changes if we know that we previously sampled a noun. This is for instance the case in an experiment where we sample two *consecutive* words. Suppose that in our corpus, 90% of the nouns are followed by verbs. For such an experiment, the probability of sampling a verb given that we first sampled a noun is thus .9 (rather than .4). This probability is written as $P(\{\text{verb}\}|\{\text{noun}\})$ and is called the *conditional probability* of a verb given a noun. But now what is the probability of sampling a noun *and* a verb in this particular experiment? We know that

$P(\{\text{noun}\}) = .5,$

$P(\{\text{verb}\}|\{\text{noun}\}) = .9.$

That is, in 50% of the cases we sample a noun, after which in 90% of the cases we sample a verb (in this experiment). This means that we sample them jointly in 90% of 50% of the cases—that is, in 45% of the cases. Thus, the joint probability $P(\{\text{noun}\}, \{\text{verb}\})$ is equal to the product $P(\{\text{noun}\}) \times P(\{\text{verb}\}|\{\text{noun}\}) = .5 \times .9 = .45$. In general, for two events A and B,

$$P(A, B) = P(A) \times P(B|A), \tag{7}$$

which reads as "The probability of A and B equals the probability of A, times the probability of B given A." Note that this formula generalizes over formula (6): if the events A and B are independent, $P(B|A)$ is equal to $P(B)$, and (7) reduces to (6). Formula (7) is generally known as the *multiplication rule* or *product rule*. The product rule can also be written as a general definition for conditional probability:

$$P(B|A) = \frac{P(A, B)}{P(A)}. \tag{8}$$

It is important to realize that a conditional probability is itself a probability function, and its values sum up to 1 by varying what is on the left-hand side of the bar in (8). Most textbooks on probability theory first define the concept of conditional probability and then, from that, the formula for joint probability. For the current exposition, it seemed more intuitive to do this the other way round.

From (8), *Bayes' rule* can be derived. First, we will rename the variables of (8):

$$P(H|E) = \frac{P(E, H)}{P(E)}, \tag{9}$$

where, in the context of *Bayesian reasoning*, $P(H|E)$ usually reads as "the probability of a hypothesis H given some evidence E." Second, since set intersection is commutative (i.e., $A \cap B = B \cap A$), the joint probability $P(E, H)$ is equal to $P(H, E)$, and we can therefore write the right-hand side of (9) also as $P(H, E)/P(E)$, which according to (7) is equal to $P(H) \times P(E|H)/P(E)$. Thus, (9) can be written as

$$P(H|E) = \frac{P(H) \times P(E|H)}{P(E)}. \tag{10}$$

This formula, known as Bayes' rule, is useful if the conditional probability $P(H|E)$ is more difficult to compute than $P(H)$ and $P(E|H)$. The probability $P(H)$ is usually called the *prior probability*, while $P(E|H)$ is

called the *posterior probability*. We will see later in this book how Bayes' rule can be applied to linguistic phenomena.

Turning back to the concept of joint probability, the product rule (7) for two events can be generalized to multiple events. For example, the joint probability of three events A, B, and C is

$$P(A, B, C) = P(A) \times P(B|A) \times P(C|A, B), \tag{11}$$

which reads as "The probability of A, B, and C equals the probability of A, times the probability of B given A, times the probability of C given A and B." The proof of (11) follows straightforwardly when we combine the associative property of set intersection (i.e., $A \cap B \cap C = A \cap (B \cap C) = (A \cap B) \cap C$) with formula (7): $P(A, B, C) = P(A, (B, C)) = P(A) \times P(B, C|A) = P(A) \times P(B|A) \times P(C|A, B)$. And for n events $A_1, A_2, \ldots, A_n$, the multiplication rule becomes

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \times P(A_2|A_1) \times \cdots \times P(A_n|A_1, A_2, \ldots, A_{n-1}), \tag{12}$$

which is also known as the *chain rule*. Remember that in an experiment where the events $A_1, A_2, \ldots, A_n$ are independent, formula (12) simply reduces to

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \times P(A_2) \times \cdots \times P(A_n). \tag{13}$$

Sometimes, each event depends only on the immediately previous event, in which case formula (12) reduces to

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \times P(A_2|A_1) \times \cdots \times P(A_n|A_{n-1}). \tag{14}$$

Formula (14) stands for what is more commonly known as a *first-order markov model*, where each event depends only on the preceding event; and formula (13) corresponds to a *zero-order markov model*. In general, a *k-th order markov model* assumes that each event depends only on a fixed number of k preceding events, where k is called the *history* of the model. For several decades, markov models were assumed to be inadequate for linguistics because they were applied to word sequences (*n-grams*, such as *bigrams* or *trigrams*) only, without taking into account the grammatical structure of these sequences. Yet we will see in the following section that formulas (12) through (14) can just as well be applied to grammatical structures.

It is useful to introduce the product sign Π, which abbreviates long products (and is analogous to the sum sign Σ, which abbreviates long sums). For example, (12) is often written as

$$P(A_1, A_2, \ldots, A_n) = \Pi_i P(A_i | A_1, A_2, \ldots, A_{i-1}). \quad (15)$$

And (as with (13)), if the events are independent, (15) reduces to

$$P(A_1, A_2, \ldots, A_n) = \Pi_i P(A_i). \quad (16)$$

It is important to understand the difference in use between the sum rule in (4) and the product rule in (6) and (7). The sum rule describes the probability that either event A or event B occurs in some experiment, which is equal to the *sum* of their probabilities (provided that A and B are disjoint[5]). The product rule, on the other hand, describes the probability that both A and B occur as a joint event in an experiment where events can have more than one outcome; and this probability is equal to the *product* of the probabilities of A and B (or in the general case, to the product of the probability of A and the conditional probability of B given A).

2.4 Probabilistic Grammars

With these concepts from probability theory in mind, we can now look at an example of actual linguistic interest: *probabilistic grammars* (also called *stochastic grammars*). As the following chapters will show, probabilistic grammars are used to describe the probabilistic nature of a vast number of linguistic phenomena, such as phonological acceptability, morphological alternations, syntactic well-formedness, semantic interpretation, sentence disambiguation, and sociolinguistic variation.

One of the most widely used probabilistic grammars is the *probabilistic context-free grammar* or *PCFG* (also called *stochastic context-free grammar*). As an introduction to PCFGs, consider a simple example. Suppose we have a very small corpus of phrase structure trees (also called a *treebank*) consisting of only two surface trees for the sentences *Mary hates visiting relatives* and *John likes buzzing bees* (figure 2.1). We will assume that each tree in the treebank corresponds to the structure as it was perceived for that sentence by some hypothetical natural language user. (Some subcategorizations are omitted to keep the example simple.) Note that the only difference between the two structures (apart from the words) is the syntactic label covering the last two words of the sentences, which is *VP* in the first sentence and *NP* in the second. By reading the rules off the trees, we obtain the context-free grammar (CFG) implicit in these structures. Table 2.1 gives these rules together with their frequencies in the

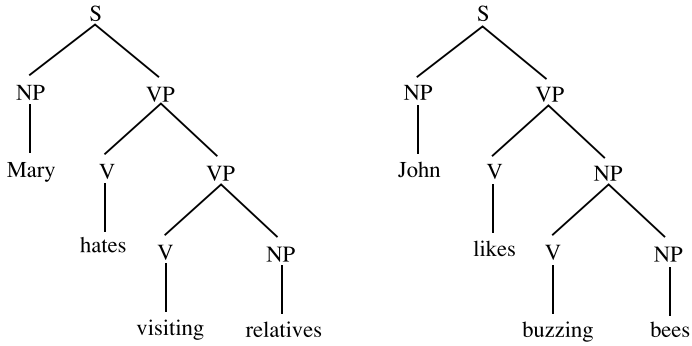

Figure 2.1
A treebank of two trees

Table 2.1
The rules implicit in the treebank of figure 2.1

Rule	Frequency
S → NP VP	2
VP → V NP	2
VP → V VP	1
NP → V NP	1
NP → Mary	1
NP → John	1
NP → relatives	1
NP → bees	1
V → hates	1
V → likes	1
V → visiting	1
V → buzzing	1
Total	14

treebank. This table allows us to derive, for example, the probability of the rule S → NP VP in the treebank—or, more precisely, the probability of randomly selecting S → NP VP from among all rules in the treebank. The rule S → NP VP occurs twice in a sample space of 14 rules; hence, its probability is $2/14 = 1/7$. However, usually we are interested not so much in the probability of a single rule, but in the probability of a combination of rules (i.e., a *derivation*) that generates a particular sentence. The grammar derived from the treebank in table 2.1 generates an infinite number of sentences, including *Mary likes buzzing bees*, *Mary likes visiting buzzing bees*, *Mary likes visiting buzzing visiting bees*. Thus, although these sentences are not in the treebank, they can be generated by productively combining fragments from the treebank trees.[6] For example, *Mary likes buzzing bees* can be generated by combining the rules from table 2.1 that are shown in figure 2.2. This combination of rules, or derivation, produces the tree structure in figure 2.3.[7] Note that the sentence *Mary likes buzzing bees* is ambiguous. That is, it can also be generated by combining the rules from table 2.1 that are shown in figure 2.4, which produce the alternative tree structure in figure 2.5.

One application of probability theory is to provide a ranking of the various tree structures for a sentence by means of their probabilities. How can we determine the probabilities of the two structures in figures 2.3 and 2.5? Using the concepts from sections 2.2 and 2.3, we can view a tree structure as an *event* containing the context-free rules in an *experiment* that parses a particular sentence by a (leftmost) derivation. In this experiment, we thus first select an S-rule from among all possible S-rules. We then select the next rule among the rules that can be combined with the previous rule (i.e., that start with the same category as the leftmost category on the right-hand side of the previous rule), and we repeat this process until only words remain. Note that this experiment is well defined only if each rule can indeed be combined with the previous rule and if the first rule starts with an S. Thus, the probability of the derivation corresponding to the tree in figure 2.3 is the *joint probability* of selecting the rules in table 2.2.

The probability of (1) can be computed by dividing the number of occurrences of rule S → NP VP by the number of occurrences of all rules that start with an S. There are two rules S → NP VP in the treebank, and the total number of S-rules is also two (in fact, they coincide); thus, the probability of (1) is $2/2 = 1$. Note that this probability is actually the conditional probability $P(\text{S} \rightarrow \text{NP VP}|\text{S})$, and thus the sum of the condi-

S → NP VP
NP → Mary
VP → V NP
V → likes
NP → V NP
V → buzzing
NP → bees

Figure 2.2
Treebank rules for deriving *Mary likes buzzing bees*

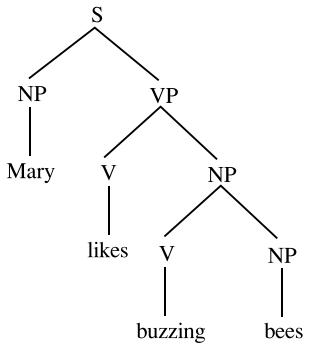

Figure 2.3
Tree structure generated by the rules in figure 2.2

tional probabilities of all rules given a certain nonterminal to be rewritten is 1.

The probability of (2) is equal to 1/5 since the rule NP → *Mary* occurs once among a total of five rules that start with an NP.

The probability of (3) is equal to 2/3 since the rule VP → V NP occurs twice among a total of three rules that start with a VP.

The probabilities of all rules in table 2.2 are given in table 2.3. Having computed these probabilities, how can we now compute their joint probability? That is, are the rules to be taken as dependent or independent? In other words, should we apply formula (12) or (13)? A crucial assumption underlying PCFGs is that the rules in a derivation depend *only* on the nonterminal to be expanded. And this is the assumption we followed in computing the probabilities above by selecting each rule from among the rules that start with the same nonterminal (i.e., we computed the conditional probabilities $P(\text{S} \rightarrow \text{NP VP}|\text{S})$, $P(\text{NP} \rightarrow \text{Mary}|\text{NP})$, etc., rather

S → NP VP
NP → Mary
VP → V VP
V → likes
VP → V NP
V → buzzing
NP → bees

Figure 2.4
Treebank rules for deriving the alternative structure for *Mary likes buzzing bees*

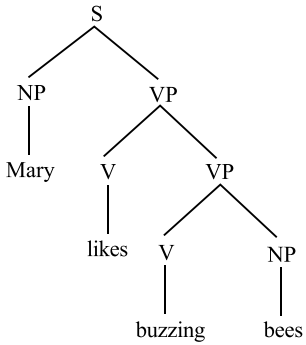

Figure 2.5
Tree structure generated by the rules in figure 2.4

Table 2.2
The probability of a derivation is the joint probability of selecting these rules

Event
(1) selecting the rule S → NP VP from among the rules starting with an S,
(2) selecting the rule NP → *Mary* from among the rules starting with an NP,
(3) selecting the rule VP → V NP from among the rules starting with a VP,
(4) selecting the rule V → *likes* from among the rules starting with a V,
(5) selecting the rule NP → V NP from among the rules starting with an NP,
(6) selecting the rule V → *buzzing* from among the rules starting with a V,
(7) selecting the rule NP → *bees* from among the rules starting with an NP.

Table 2.3
The probabilities of the various rules in table 2.2

Event	Probability
(1) selecting the rule S → NP VP from among the rules starting with an S	1
(2) selecting the rule NP → *Mary* from among the rules starting with an NP	1/5
(3) selecting the rule VP → V NP from among the rules starting with a VP	2/3
(4) selecting the rule V → *likes* from among the rules starting with a V	1/4
(5) selecting the rule NP → V NP from among the rules starting with an NP	1/5
(6) selecting the rule V → *buzzing* from among the rules starting with a V	1/4
(7) selecting the rule NP → *bees* from among the rules starting with an NP	1/5

than the simple probabilities $P(\text{S} \rightarrow \text{NP VP})$ and $P(\text{NP} \rightarrow \text{Mary})$). Thus, for a PCFG, the probability of a rule is independent of the derivation it occurs in and so can be computed off-line. Table 2.4 gives the PCFG probabilities for all rules that can be derived from the treebank in figure 2.1. PCFGs can of course be defined independently of how the rule probabilities are "learned." A PCFG that extracts the probabilities directly from a treebank, as shown above, is known as a *treebank grammar*, a term coined by Charniak (1996).

Let us now turn back to the probability of the derivation for *Mary likes buzzing bees* that generates the tree in figure 2.3. This can be computed as the product of the probabilities in table 2.3, that is, $1 \times 1/5 \times 2/3 \times 1/4 \times 1/5 \times 1/4 \times 1/5 = 2/6,000 = 1/3,000$. This probability is small, reflecting the fact that the grammar produces derivations for infinitely many sentences whose probabilities sum up to 1 only in the limit. But what we are actually interested in is to compare the probability of this derivation with the probability of the other derivation for *Mary likes buzzing bees* (producing the tree in figure 2.5). The latter probability is equal to $1 \times 1/5 \times 1/3 \times 1/4 \times 2/3 \times 1/4 \times 1/5 = 2/3,600 = 1/1,800$. Thus, the probability of the derivation producing the tree in figure 2.5 is higher than the probability of the derivation producing the tree in figure 2.3. Although we must keep in mind that our sample space of two trees is

Table 2.4
Probabilistic context-free grammar (PCFG) probabilities for the rules derived from the treebank in figure 2.1

Rule	PCFG probability
S → NP VP	1
VP → V NP	2/3
VP → V VP	1/3
NP → V NP	1/5
NP → Mary	1/5
NP → John	1/5
NP → relatives	1/5
NP → bees	1/5
V → hates	1/4
V → likes	1/4
V → visiting	1/4
V → buzzing	1/4

unrealistically small (most available treebanks contain 50,000 trees or more), it is somewhat surprising that the tree in figure 2.5 has a higher probability than the one in figure 2.3. We would expect the reverse: since *Mary likes buzzing bees* differs by only one word from the treebank sentence *John likes buzzing bees* but differs much more from the other treebank sentence, *Mary hates visiting relatives*, we might expect a probabilistic grammar to predict that the most probable tree for *Mary likes buzzing bees* would be the same tree associated with *John likes buzzing bees*, rather than the tree associated with *Mary hates visiting relatives*. However, as noted, a crucial assumption underlying PCFGs is that their rules are independent. It is easy to see that this assumption is wrong, even for the subclass of natural language sentences that *are* in fact context free. For example, the words *buzzing* and *bees* in the NP *buzzing bees* are probabilistically dependent: that is, the probability of observing *bees* is not equal to the probability of observing *bees* given that we have first observed *buzzing*. But this dependency is not captured by a PCFG, since it takes the rules V → *buzzing* and NP → *bees* to be independent. Thus, while a CFG may suffice as a grammar formalism for defining the categorical properties for the context-free subset of sentences, its probabilistic counterpart PCFG does not do the same job for the *non*categorical properties of this context-free subset.

Several alternative models have been proposed to redress the shortcomings of PCFGs. These alternative probabilistic extensions of CFGs have resulted in probabilistic grammars that are provably stronger than PCFGs (''stronger'' will be explained more precisely in the next section). One such grammar makes the probabilities of the rules dependent on the previous rules used in a derivation, by effectively applying formula (12) to the rules (Black et al. 1993). However, while such a *history-based grammar* can thereby capture the dependency between *buzzing* and *bees*, it has problems with dependencies between words that are separated by other words, as for example in the sentence *The old man died*, where there is a dependency between *old* and *died* but not between *old* and *man* or between *man* and *died*. It cannot capture this dependency because the rules are made dependent on *directly* preceding rules, and not on any arbitrary previously used rule(s).

Another probabilistic grammar formalism, which has become quite influential in the field of natural language processing, associates each nonterminal of a context-free rule with its lexical head according to the treebank tree (e.g., Collins 1996; Charniak 1997a). However, such a *head-lexicalized probabilistic grammar* neglects dependencies that go beyond simple headword dependencies, such as the one between *nearest* and *to* in the ATIS[8] sentence *Show the nearest airport to Denver*. Since a head-lexicalized probabilistic grammar considers *nearest* to be a non-headword of the NP *the nearest airport*, it incorrectly disambiguates this sentence (it assigns the highest probability to the tree where the PP *to Denver* is attached to *show*, since the dependency between the headwords *show* and *to* is more likely in the ATIS treebank than that between the headwords *airport* and *to*). The shortcomings of head-lexicalized probabilistic grammars are discussed more fully in Bod 2001b.

What we may learn from these different probabilistic formalisms is that the probability of a *whole* (i.e., a tree) can be computed from the combined probabilities of its *parts*, but that it is difficult to decide what the *relevant* parts are. In a PCFG, the relevant parts are assumed to be the simple CFG rules (clearly wrong), while in a head-lexicalized grammar, the parts are assumed to be the rules enriched with their lexical heads (also too limited). Another probabilistic grammar formalism, *probabilistic lexicalized tree-adjoining grammar* (Schabes 1992; Resnik 1992), takes the elementary trees of a tree-adjoining grammar as the relevant parts (see Bod 1998 for a critique of this formalism).

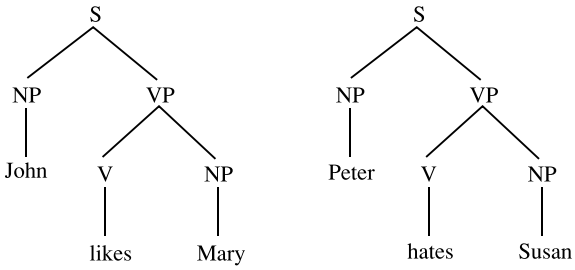

Figure 2.6
A treebank of two trees

Still another formalism generalizes over most other probabilistic grammars. It does so by taking any subtree (of arbitrary size) as a *part*, including the entire trees from a treebank. This formalism, known as a *Data-Oriented Parsing* (DOP) model (Bod 1993, 1998), is formally equivalent to a probabilistic tree substitution grammar. A DOP model captures the previously mentioned problematic dependency between *old* and *died*, or *nearest* and *to*, by a subtree that has the two relevant words as its only lexical items. Moreover, a DOP model can capture arbitrary fixed phrases and idiom chunks, such as *to take advantage of*. Note that a DOP model reduces to a PCFG if the size of the subtrees is limited to the smallest ones.

To see how a DOP model works, consider a simple example. Since the number of subtrees tends to be quite large, we will use the tiny treebank shown in figure 2.6. A total of 34 subtrees, shown in figure 2.7, can be derived from this treebank (at least if we use one specific instantiation of DOP, known as DOP1; see Bod 1998). Notice that some subtrees occur twice: a subtree may be extracted from different trees, and also from a single tree if the same node configuration appears at different positions.

These subtrees form the underlying grammar by which new sentences are generated. Subtrees are combined using a *node substitution operation* similar to the operation that combines context-free rules in a (P)CFG, indicated by the symbol "$\circ$". Given two subtrees T and U, the node substitution operation substitutes U on the leftmost nonterminal leaf node of T, written as $T \circ U$. For example, the sentence *Mary likes Susan* can be generated by combining three subtrees from figure 2.7 as shown in figure 2.8.

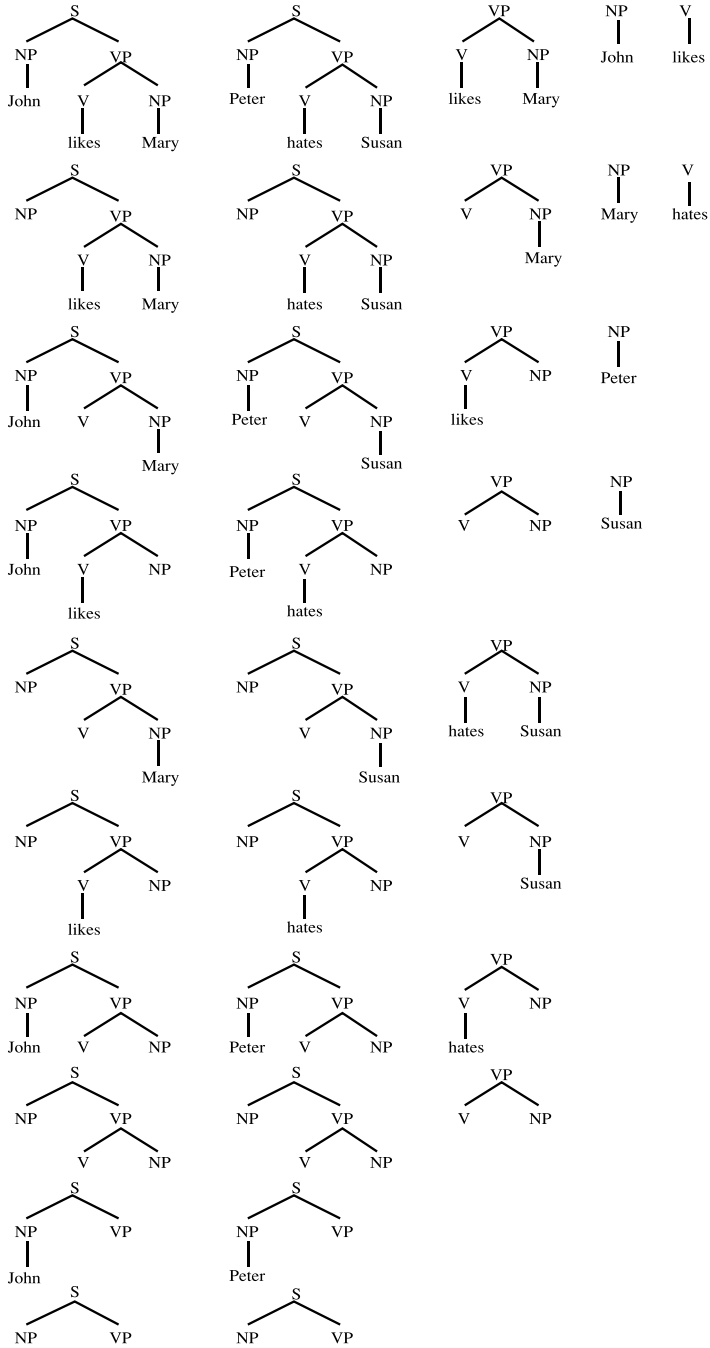

Figure 2.7
The subtrees derived from the trees in figure 2.6

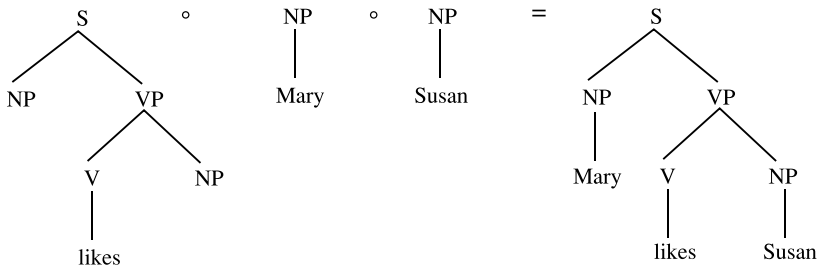

Figure 2.8
Generating *Mary likes Susan* by combining subtrees from figure 2.6

Table 2.5
The probability of a derivation is the joint probability of selecting its subtrees

Event
(1) selecting the subtree $[_S$ NP $[_{VP}[_V$ likes] NP]] from among the subtrees with root label S,
(2) selecting the subtree $[_{NP}$ Mary] from among the subtrees with root label NP,
(3) selecting the subtree $[_{NP}$ Susan] from among the subtrees with root label NP.

The events involved in this derivation are listed in table 2.5. The probability of (1) is computed by dividing the number of occurrences of the subtree $[_S$ NP $[_{VP}[_V$ likes] NP]] in figure 2.7 by the total number of occurrences of subtrees with root label S: 1/20. The probability of (2) is equal to 1/4, and the probability of (3) is also equal to 1/4.

The probability of the whole derivation is the joint probability of the three selections in table 2.5. Since in DOP each subtree selection depends only on the root label and not on the previous selections, the probability of a derivation is, as in PCFG, the product of the probabilities of the subtrees, in this case $1/20 \times 1/4 \times 1/4 = 1/320$. Although it is again assumed that the *parts* of the probabilistic grammar are independent, this assumption is not harmful in DOP, since if the treebank contains any larger subtree that includes two (or more) smaller subtrees, it can directly be used as a unit in a derivation, thereby taking into account the co-occurrence of the smaller subtrees.

This brings us to another feature of DOP: the fact that different derivations can produce the *same* tree. This so-called *spurious ambiguity* may be irrelevant for nonprobabilistic grammars, but for probabilistic gram-

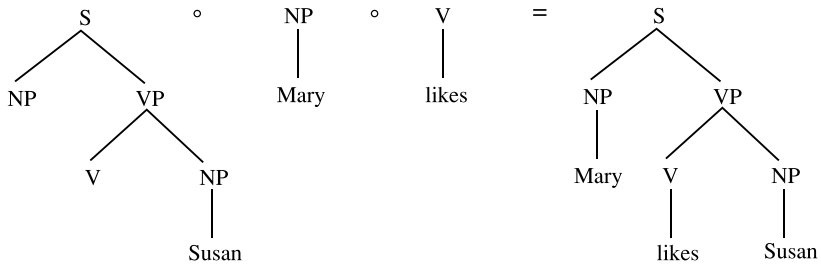

Figure 2.9
A different derivation generated by combining subtrees from figure 2.6, yielding the same parse for *Mary likes Susan*

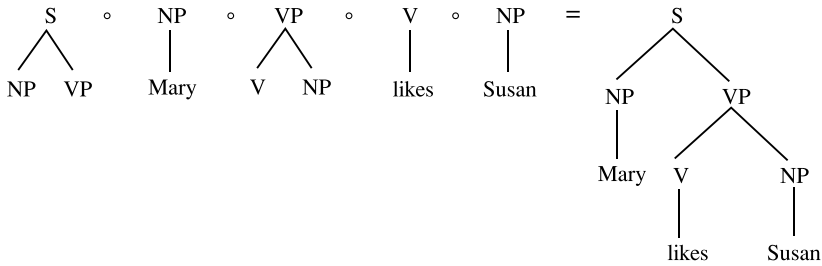

Figure 2.10
Another derivation generated by combining subtrees from figure 2.6, yielding the same parse for *Mary likes Susan*

mars it leads to a different probability model. For example, the tree shown in figure 2.8 for the sentence *Mary likes Susan* can also be derived by combining the subtrees shown in figure 2.9. The probability of this derivation is equal to $1/20 \times 1/4 \times 1/2 = 1/160$, which is different from the probability of the derivation in figure 2.8, even though it produces the same tree. And in fact there are many more derivations that produce this tree, each with its own probability. The one shown in figure 2.10 is analogous to a PCFG derivation for *Mary likes Susan*, in that each subtree exactly corresponds to a context-free rewrite rule. The probability of this derivation is equal to $2/20 \times 1/4 \times 2/8 \times 1/2 \times 1/4 = 1/1,280$, which is again different from the probabilities of the other two derivations generating this tree.

Thus, DOP does not exhibit a one-to-one correspondence between derivation and tree, as PCFG does. Instead, there may be several distinct

derivations for the same tree. The probability that a certain tree occurs is then the probability that *any* of its derivations occurs. According to rule (4), this amounts to saying that the probability of a tree is the *sum* of the probabilities of its derivations (I leave the computation of the tree probability for *Mary likes Susan* to the reader). Intuitively this means that in DOP, evidence for a tree accumulates: the more derivations a tree has, the larger its probability tends to be. This means that if a tree can be constructed (also) from large subtrees found in a treebank, it tends to be ranked higher than a tree than can be constructed only from small subtrees.

Note that if we are interested in the probability of generating a certain sentence, we must sum up the probabilities of all different *trees* that generate that sentence—following the same reasoning as for the probability of a tree. It can be shown that the sum of the probabilities of all sentences generated by a DOP model is equal to 1 (following Chi and Geman 1998).

The DOP model outlined here, DOP1, is just one of many DOP models that have been proposed (see Bod, Scha, and Sima'an 2002 for an overview). The distinctive features of the general DOP approach, when it was proposed in Bod 1992, were that (1) it directly used sentence fragments as a grammar, and (2) it did not impose constraints on the size of the fragments. While (1) is now relatively uncontroversial in probabilistic natural language processing (see Manning and Schütze 1999), (2) has not been generally adopted. Many models still work either with local trees, that is, single-level rules with limited means of information percolation such as headwords (e.g., Collins 1996; Charniak 1997a), or with restricted fragments, as in probabilistic lexicalized tree-adjoining grammar, that do not include nonlexicalized fragments (e.g., Schabes 1992; Chiang 2000). However, the last few years have seen a shift toward using more and larger treebank fragments. While initial extensions of PCFGs limited fragments to the locality of headwords (e.g., Collins 1996; Eisner 1996), later models have shown the importance of including context from higher nodes in the tree (e.g., Johnson 1998b). The importance of including nonheadwords is now widely accepted (e.g., Goodman 1998; Collins 1999; Charniak 2000). And Collins (2000, 176) argues for "keeping track of counts of arbitrary fragments within parse trees," a proposal carried out by Collins and Duffy (2001, 2002), who use exactly the same set of sentence fragments that was proposed in the original DOP model (Bod 1992).

From a linguistic point of view, the more interesting question is whether *language users* store sentence fragments in memory, and if they do, whether they store arbitrarily large fragments as the DOP model proposes. Jurafsky (this volume) reports that people store not only lexical items, but also frequent bigrams (two-word units), frequent phrases, and even whole sentences. For the case of sentences, there is some evidence that language users store not only idioms, but also simple high-frequency sentences such as *I love you* and *I don't know* (Jurafsky, this volume; Bod 2001a). The fact that language users store sentence fragments in memory and that these fragments can range from two-word units to entire sentences suggests that language users need not always generate or parse sentences from scratch using the rules of the grammar, but that they can productively reuse previously heard sentences and sentence fragments. Yet there is no evidence so far that people memorize *all* fragments they hear. Only high-frequency fragments seem to be stored. However, if the language faculty has to *learn* which fragments will be stored, it will initially need to store everything (with the possibility of forgetting some things, of course); otherwise, frequencies can never accumulate. This results in a model that continuously and incrementally updates its fragment memory given new input. We will see that such a model turns out to be important for almost all subfields of (probabilistic) linguistics, ranging from phonology to syntax and from psycholinguistics to sociolinguistics.

Another interesting linguistic question is whether DOP models are too *general*. Since DOP models essentially store all sentences, they perhaps do not provide sufficient constraints for defining the set of possible languages. Since this question is aptly dealt with by Manning (this volume), I will not go into it here. Still another interesting linguistic question is whether DOP models of the type outlined above are actually too *constrained*, since they have the generative power of context-free languages (this follows from the node substitution operation for combining subtrees). Although context-free power may suffice for phonology (Pierrehumbert, this volume) and morphology (Baayen, this volume), there are syntactic phenomena, such as long-distance dependencies and cross-serial dependencies, that are known to be beyond context free. Therefore, a model that is inherently context free is deemed to be linguistically inadequate. In the last few years, various DOP models have been developed whose generative capacity is richer than context free. These models are based on linguistic representations that also allow for syntactic features, functional categories, and semantic forms (see Bod and Kaplan 1998;

Neumann 1998; Hoogweg 2002). Although a detailed description of these models lies beyond the scope of this chapter, it should be noted that for these richer DOP models as well, fragments of arbitrary size are indispensable for predicting correct sentence structure (see Bod 1998; Way 1999; Bod and Kaplan 2002). Manning (this volume) examines some other probabilistic extensions of non-context-free grammars.

2.5 Formal Stochastic Language Theory

We have seen that a DOP model (actually, a DOP1 model) generalizes over a PCFG. But is DOP also probabilistically "richer" than a PCFG? That is, is it impossible to create a PCFG for every DOP model? These two questions lead us to ask how two probabilistic grammars can be compared. First note that in comparing probabilistic grammars, we are not interested in the traditional notion of generative capacity, since for example DOP1, PCFG, history-based grammar, and head-lexicalized grammar are all context free. Instead, we are interested in the probability distributions that these probabilistic grammars define over sentences and their trees.

Two central concepts in traditional Formal Language Theory are *weak equivalence* and *strong equivalence*. Two grammars are said to be weakly equivalent if they generate the same strings, and strongly equivalent if they generate the same strings with the same trees. The set of strings generated by a grammar G is also called the *string language* of G, while the set of trees generated by G is called the *tree language* of G.

Analogously, the two central concepts in *Formal Stochastic Language Theory* are *weak stochastic equivalence* and *strong stochastic equivalence*. Two probabilistic grammars are said to be weakly stochastically equivalent, if and only if they generate the same stochastic string language (where the *stochastic string language* generated by a probabilistic grammar G is the set of pairs $\langle x, P(x)\rangle$ where x is a string from the string language generated by G and $P(x)$ the probability of that string). Two probabilistic grammars are said to be strongly stochastically equivalent if and only if they generate the same stochastic tree language (where the *stochastic tree language* generated by a probabilistic grammar G is the set of pairs $\langle x, P(x)\rangle$ where x is a tree from the tree language generated by G and $P(x)$ the probability of that tree). Note that if two probabilistic grammars are strongly stochastically equivalent, they are also weakly stochastically equivalent.

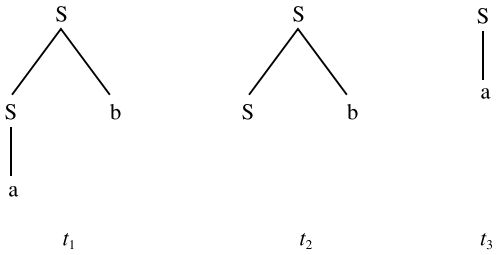

Figure 2.11
A probabilistic tree substitution grammar consisting of three elementary trees

As an illustration of how Formal Stochastic Language Theory can be used to compare different formalisms, we will investigate whether PCFG and DOP are strongly stochastically equivalent (for some other comparisons, see Bod 1998; Carroll and Weir 2000). Since the instantiation of DOP in this chapter is equal to a probabilistic tree substitution grammar (PTSG), we will refer to this DOP model as a PTSG (in accordance with Manning and Schütze 1999, 446–448). Specifically, the question is, Is there a PTSG for which there is a strongly equivalent PCFG but no strongly stochastically equivalent PCFG? As can easily be shown, the answer is yes. Consider the very simple PTSG G in figure 2.11 consisting of three subtrees that are all assigned a probability of 1/3.[9] The string language generated by G is $\{a, ab, abb, abbb, abbbb, \ldots\}$, which can be abbreviated as $\{ab^*\}$. The only PCFG G' that is strongly equivalent to G consists of the following productions:

(1) $S \rightarrow Sb$

(2) $S \rightarrow a$

G' would also be strongly stochastically equivalent to G if it assigned the same probabilities to the parse trees in the tree language as those assigned by G. Let us consider the probabilities of two trees generated by G—specifically, the trees represented by t_1 and t_3.[10] The tree represented by t_3 has exactly one derivation, which consists of the subtree t_3. The probability of generating this tree is hence equal to 1/3. The tree represented by t_1 has two derivations: by selecting subtree t_1, or by combining the subtrees t_2 and t_3. The probability of generating this tree is equal to the sum of the probabilities of its two derivations; that is, $1/3 + (1/3 \times 1/3) = 4/9$.

If G' is strongly stochastically equivalent to G, it should assign the probabilities 4/9 and 1/3 to (at least) the trees represented by t_1 and t_3, respectively. The tree t_3 is exhaustively generated by production (2); thus, the probability of this production should be equal to 1/3: $P(S \rightarrow a) = 1/3$. The tree t_1 is exhaustively generated by applying productions (1) and (2); thus, the product of the probabilities of these productions should be equal to 4/9: $P(S \rightarrow Sb) \times P(S \rightarrow a) = 4/9$. By substitution, we get $P(S \rightarrow Sb) \times 1/3 = 4/9$, from which we derive that $P(S \rightarrow Sb) = 4/3$. This means that the probability of the production $S \rightarrow Sb$ should be larger than 1, which is not allowed. Thus, G' cannot be made strongly stochastically equivalent to G.

This proof shows that *there exists a PTSG for which there is no strongly stochastically equivalent PCFG* (even if it is strongly equivalent). On the other hand, it can easily be shown that *for every PCFG there exists a strongly stochastically equivalent PTSG*: for any rule in any PCFG, one can create a minimal one-level subtree (with the same probability) covering exactly the corresponding rule.

Now, if for every PCFG there is a strongly stochastically equivalent PTSG, but not vice versa, then *the set of stochastic tree languages generated by the class of PCFGs is a proper subset of the set of stochastic tree languages generated by the class of PTSGs*. This is what it means to say that PTSGs are "richer" than PCFGs.

The goal of this section has been to present a framework in which different probabilistic grammars can be compared. The importance of such a comparison should not be underestimated. If we invent a new formalism only to find out that for each grammar in this formalism we can create a strongly stochastically equivalent PCFG, then we haven't made much progress. Thus, rather than being interested in a grammar's place in the Chomsky hierarchy (Chomsky 1959), we are often more interested in its place in the *stochastic* hierarchy within one and the same class of the Chomsky hierarchy.

2.6 Conclusion

Although the background knowledge outlined here (and in the glossary) should suffice for understanding this book, this chapter only scratches the surface of probability theory and probabilistic grammars. Important topics it has not touched on include probabilistic regular grammars

(which are equivalent to markov models), probabilistic attribute-value grammars (which generalize over several richer probabilistic grammars), and consistency requirements for probabilistic grammars (which turn out to be particularly interesting for DOP models—see Bod 2000a; Johnson 2002). If the reader feels cheated and wants to see the full picture, then I have achieved my goal. Excellent textbooks and overview articles on probability theory and formal stochastic language theory are available, some of which are mentioned below.

The reader may wonder whether probability theory is really needed to cope with gradience and frequency effects in language, or whether these effects could just as well be accounted for by other approaches such as Optimality Theory or connectionism. Then it is really time to dive into the following chapters: probabilistic approaches nowadays cover the entire spectrum of linguistics, and other approaches are increasingly turning to probabilistic models, including Optimality Theory and connectionism.

2.7 Further Reading

There are many good introductory textbooks on probability theory and statistics. A very accessible introduction is Moore and McCabe 1989, which focuses on probability distributions. Other textbooks include Ross 2000 and Feller 1970, as well as (at a more advanced level) Breiman 1973 and Shao 1999. For an introduction from a Bayesian standpoint, see DeGroot 1989. Krenn and Samuelsson 1997 offers a tutorial on probability theory from the viewpoint of natural language processing. Oakes 1998 gives an overview of the use of statistics in corpus linguistics. An interesting survey on the emergence of probability in the history of thought is Hacking 1975.

Probabilistic grammars were first studied outside linguistics: they were used for pattern recognition (Grenander 1967), and mathematical properties of PCFGs were explored (Booth 1969). It was shown that PCFGs can be learned from positive data alone (Horning 1969); this result turns out to be quite important for probabilistic linguistics (see Manning, this volume). One of the first papers that argues for PCFGs from a linguistic standpoint is Suppes 1970. Manning and Schütze 1999 gives a good overview of the properties of PCFGs and discusses several enhancements. Jurafsky and Martin 2000 explores the psycholinguistic relevance of PCFGs. Chi and Geman 1998 shows that proper probability distributions

are obtained if the probabilities of the PCFG rules are estimated directly from a treebank (as proposed in Bod 1993 and Charniak 1996).

An overview of probabilistic extensions of CFGs is included in Charniak 1997b, Bod 1998, 2001b, and Manning and Schütze 1999. Probabilistic grammars for languages richer than context free are developed in Abney 1997, Bod and Kaplan 1998, Johnson et al. 1999, and elsewhere. DOP models are discussed in Bod 1998 and Bod, Scha, and Sima'an 2002. Regarding the properties of various probability models for DOP, see Bod 2000a, Bonnema 2002, Goodman 2002, and Johnson 2002.

Initial comparisons of different probabilistic grammars focused on their stochastic string languages (e.g., Fu 1974; Levelt 1974; Wetherell 1980) Bod 1993 distinguishes between weak and strong stochastic equivalence, and Bod 1998 uses these concepts to compare different probabilistic extensions of CFGs, suggesting a hierarchy of probabilistic grammars within the classes of the Chomsky hierarchy. Abney, McAllester, and Pereira 1999 investigates the exact relationship between probabilistic grammars and probabilistic automata. Carroll and Weir 2000 demonstrates the existence of a subsumption lattice of probabilistic grammars with PCFG at the bottom and DOP at the top.

Notes

I wish to thank all coauthors of this book for their helpful feedback on this chapter. I am especially grateful to Jennifer Hay and Chris Manning, whose extensive comments were particularly useful.

1. The word *stochastic* is used as a synonym for *probabilistic*, but is especially used when it refers to results generated by an underlying probability function.

2. In this book and in probabilistic linguistics in general, the word *sampling* always refers to sampling with replacement.

3. Note that the probability of first sampling a verb and then a noun is also .2. This is because set intersection is commutative: $\{\text{noun}\} \cap \{\text{verb}\} = \{\text{verb}\} \cap \{\text{noun}\}$ and therefore $P(\{\text{noun}\} \cap \{\text{verb}\}) = P(\{\text{verb}\} \cap \{\text{noun}\})$. This also means that the probability of sampling a noun and a verb in *any* order is equal to $.2 + .2 = .4$.

4. In this book, multiplications are often written without the multiplication sign. Thus, $P(A) \times P(B)$ is also written as $P(A)P(B)$.

5. If A and B are not disjoint, there is double counting, which means that the counts of the intersection of A and B should be subtracted. Thus, for the general case, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

6. This shows that the "Chomskyan myth" that finite corpora can only generate finite numbers of sentences is fallacious.

7. Without loss of generality, we will assume that a tree or a sentence is produced by a *leftmost* derivation, where at each step the leftmost nonterminal is rewritten.

8. Air Travel Information System (see Marcus, Santorini, and Marcinkiewicz 1993).

9. This PTSG would correspond to a DOP model of which the subtrees are taken from a treebank consisting only of tree t_1.

10. Note that the trees t_1 and t_3 are both elements of the set of subtrees of G and of the tree language generated by G.

This page intentionally left blank

Chapter 3

Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production

Dan Jurafsky

Probability is not really about numbers; it is about the structure of reasoning.
—Glenn Shafer, cited in Pearl 1988

3.1 Introduction

It must certainly be accounted a paradox that probabilistic modeling is simultaneously one of the oldest and one of the newest areas in psycholinguistics. Much research in linguistics and psycholinguistics in the 1950s was statistical and probabilistic. But this research disappeared throughout the '60s, '70s, and '80s. In a highly unscientific survey (conducted by myself) of six modern college textbooks and handbooks in psycholinguistics, not one mentions the word *probability* in the index.

This omission is astonishing when we consider that the input to language comprehension is noisy, ambiguous, and unsegmented. In order to deal with these problems, computational models of speech processing have had to rely on probabilistic models for many decades. Computational techniques for processing of text, an input medium much less noisy than speech, rely just as heavily on probability theory. Just to pick an arbitrary indicator, 77% of the papers presented at the year 2000 conference of the Association for Computational Linguistics relied on probabilistic models of language processing or learning.

Probability theory is certainly the best normative model for solving problems of decision making under uncertainty. Perhaps, though, it is a good normative model but a bad descriptive one. Even though probability theory was originally invented as a cognitive model of human reasoning under uncertainty, perhaps people do not use probabilistic reasoning in cognitive tasks like language production and comprehension. Perhaps human language processing is simply a nonoptimal, nonrational process?

In the last decade or so, a consensus has been emerging that human cognition is in fact rational and relies on probabilistic processing. Seminal work by Anderson (1990) gave Bayesian underpinnings to cognitive models of memory, categorization, and causation. Probabilistic models have cropped up in many areas of cognition, among them categorization (Glymour and Cheng 1998; Rehder 1999; Tenenbaum 2000; Tenenbaum and Griffiths 2001a,b).

Probabilistic models are also now finally being applied in psycholinguistics, drawing from early Bayesian-esque precursors in perception such as Luce's (1959) choice rule. What does it mean to claim that human language processing is probabilistic? This claim has implications for language comprehension, production, and learning.

Probability has been claimed to play three roles in language comprehension. First, consider the task of accessing linguistic structure from the mental lexicon or grammar. Perhaps more probable structures are accessed more quickly, or with less effort. Or perhaps they can merely be accessed with less evidence than less probable structures. Second, consider disambiguation. Ambiguity is ubiquitous in language comprehension: speech input is ambiguously segmented, words are syntactically and semantically ambiguous, sentences are syntactically ambiguous, utterances have ambiguous illocutionary force, and so on. Probability is one of the factors that play a role in disambiguation: the more probable an interpretation, the more likely it is to be chosen. Third, probability may play a key role in explaining processing difficulty. Recent models of what makes certain sentences difficult to process are based, at least in part, on certain interpretations having particularly low probabilities, or on sudden switches of probabilistic preference between alternative interpretations. In this chapter, I will summarize models of all three of these roles for probability in comprehension: access, disambiguation, and processing difficulty.

The claim that human language processing is probabilistic also has implications for production. Probability may play a role in accessing structures from the mental lexicon or grammar. High-probability structures may be accessed faster, or more easily, or simply with more confidence. Disambiguation, itself a phenomenon of comprehension, has a correlate in production: choice. Given multiple possible structures a speaker might say, probability may play a role in choosing among them. I will give an overview of the experimental and modeling literature on probabilistic production, although there is significantly less to summarize here than there is for comprehension.

Probability also plays a role in learning. Many models of how linguistic structure is empirically induced rely on probabilistic and information-theoretic models. I will not focus on learning here; instead, the interested reader should turn to relevant papers such as Brent and Cartwright 1996; Saffran, Aslin, and Newport 1996; Saffran, Newport, and Aslin 1996a,b; Tenenbaum and Xu 2000; and Saffran 2001.

What probabilistic modeling offers psycholinguistics is a model of the structure of evidential reasoning: a principled and well-understood algorithm for weighing and combining evidence to choose interpretations in comprehension, and to choose certain outcomes in production. Throughout the chapter, I will make reference to Bayesian reasoning, and in particular to Bayes' rule (equation (10) in chapter 2). Bayes' rule gives us a way to break down complex probabilities into ones that are easier to operationalize and compute. Suppose we are trying to compute the probability of some interpretation i given some evidence e. Bayes' rule states that this probability can be broken down as follows:

$$P(i|e) = \frac{P(e|i)P(i)}{P(e)}. \tag{1}$$

This says that we can compute the probability of an interpretation i given evidence e by instead asking how likely the interpretation i is a priori, and how likely the evidence e would be to occur if we knew the interpretation was correct. Both of these are often easier to compute than $P(i|e)$.

Probabilistic modeling has been applied to many areas of psycholinguistics: phonological processing, morphological processing, lexical processing, syntactic processing, discourse processing. I will focus in this chapter on lexical and syntactic processing. As for other areas of processing, Baayen (this volume) covers processing with respect to morphology and Pierrehumbert (this volume) touches on processing issues in phonology. Regarding probabilistic work on dialogue and discourse processing, see Jurafsky, in press.

3.2 Summary of Evidence for Probabilistic Knowledge

In this section, I will summarize evidence from psycholinguistic experiments that bears on the use of frequency-based or probabilistic knowledge in human language processing. The focus will be on lexical and syntactic processing, both in comprehension and in production.

What kinds of experiments provide evidence for probabilistic modeling? First, I will summarize many experiments showing that frequencies

of various linguistic structures play a role in processing: frequencies of words, word pairs, lexical categories of words, subcategorization relations, and so on. Why should finding evidence for frequencies support probabilistic models?

One reason is that relative frequency can play the role of prior probability in the computation of conditional probability. Recall from Bayes' rule that the probability $P(i|e)$ of some structure or interpretation i given some evidence e can be computed as follows:

$$P(i|e) = \frac{P(e|i)P(i)}{P(e)}. \tag{2}$$

This means that the conditional probability of an interpretation or structure i is directly related to the prior probability of i. Since the relative frequency of i provides an easy way to estimate the prior probability of i, the Bayesian model predicts that we should find frequency effects for various kinds of structures.

But many complex structures are too rare for their probability to be computed by counting the number of times they have occurred. Language is creative, after all, and many large structures (like sentences) may only occur once. In these cases, we would not expect to see evidence for the frequency of these unique events.

In just these cases, however, probabilistic modeling gives us tools to estimate the prior probability of these structures by making independence assumptions, allowing us to estimate the probability of one large complex object from the counts of many smaller objects. This, for example, is the goal of probabilistic grammar formalisms like Data-Oriented Parsing and stochastic context-free grammars. Since these models compute larger structures and their probabilities by combining smaller structure and their probabilities, probabilistic modeling suggests that the frequency of smaller, more primitive structures should play a role in processing. This once again predicts that we should find effects of various frequencies in processing. I will explore some of these assumptions in more detail in section 3.3.

In addition to evidence for frequencies, I will summarize evidence for various kinds of conditional probabilities. In general, I will define the probabilities of interest as the experiments are introduced.

3.2.1 Lexical Frequency

One of the earliest and most robust effects in psycholinguistics is the word frequency effect. Word frequency plays a role in both the auditory

and visual modalities, and in both comprehension and production. I will summarize results of experiments in this area; but first, it's important to know how lexical frequency is measured.

3.2.1.1 Measuring Lexical Frequency Most studies since 1970 have relied on word frequency statistics calculated from the Brown corpus of American English, a 1-million word collection of samples from 500 written texts from different genres (newspapers, novels, nonfiction, academic prose, etc.), which was assembled at Brown University in 1963–64. Kučera and Francis 1967 reports the frequency for each word-form in the corpus, while Francis and Kučera 1982 uses a lemmatized and part-of-speech tagged version of the Brown corpus to report frequencies for lemmas (e.g., reporting both combined and distinct frequencies for *go*, *goes*, *going*, *went*, and *gone*, and distinct frequencies for, say, *table* the verb and *table* the noun).

From the very beginning of the field, it was clear that deriving frequencies from corpora in this way was not unproblematic. It might seem astonishing that the wide variety of frequency effects reported in the literature are based on using this one corpus. Indeed, the use of such corpora to derive frequencies, either as a control factor or as an explicit part of a probabilistic model, is problematic in three ways. First, consider that a corpus is an instance of language production, but the frequencies derived from corpora are often used to model or control experiments in comprehension. While comprehension and production frequencies are presumably highly correlated, there is no reason to expect them to be identical. Second, the Brown corpus is a genre-stratified corpus. It contains equal amounts of material from newspapers, fiction, academic prose, and so on. But presumably a corpus designed for psychological modeling of frequency would want to model the frequency with which an individual hearer or speaker is exposed to (or uses) linguistic input. This would require a much larger focus on spoken language, on news broadcasts, and on magazines. Third, the Brown corpus dates from 1961; most subjects in psycholinguistics experiments carried out today are college undergraduates and weren't even born in 1961; the frequencies that would be appropriate to model their language capacity may differ widely from Brown corpus frequencies.

I see these problems as introducing a very strong bias *against* finding any effects of corpus frequencies in experimental materials. Nonetheless, as we will see, strong and robust effects of corpus frequencies have been found. One reason for this is that, as studies have long shown, frequencies

from different corpora are very highly correlated (Howes and Solomon 1951). A more important reason is that most studies report only very broad-grained frequencies, often using just three bins: high-frequency, low-frequency, and other. Finally, studies are beginning to use larger and more recent corpora and databases such as the CELEX lexical database (based on a corpus of 18 million words) (Baayen, Piepenbrock, and Gulikers 1995) and the British National Corpus (which has roughly 10 million tokens of tagged spoken English and 90 million tokens of written English). These corpora are large enough to allow for the direct use of unbinned frequencies (see, e.g., Allegre and Gordon 1999; de Jong et al. 2002; Baayen et al. 2002).

3.2.1.2 Lexical Frequency in Comprehension The earliest work studying word frequency effects in comprehension seems to have been by Howes and Solomon (1951), who used a tachistoscope to display a word for longer and longer durations. They showed that the log frequency of a word (as computed from corpora of over 4 million words) correlated highly with the mean time subjects took to recognize the word; more frequent words were recognized with shorter presentations. Later, the naming paradigm, in which subjects read a word out loud, was used to show that high-frequency words are named more rapidly than low-frequency words (Forster and Chambers 1973). The lexical decision paradigm, in which subjects decide if a string of letters presented visually is a word or not, has also been used to show that lexical decisions about high-frequency words are made faster than decisions about low-frequency words (Rubenstein, Garfield, and Millikan 1970; Whaley 1978; Balota and Chumbley 1984). Again, these results are robust and have been widely replicated. Frequency also plays a role in other on-line reading measures such as fixation duration and gaze duration.

Similarly robust results have been found for auditory word recognition. Howes (1957) first found results with speech that were similar to his earlier results with vision: when presented with high- and low-frequency words immersed in noise, subjects were better at identifying high- than low-frequency ones. In an extension of this experiment, Savin (1963) found that when subjects made recognition errors, they responded with words that were higher in frequency than the words that were presented. Grosjean (1980) used the gating paradigm, in which subjects hear more and more of the waveform of a spoken word, to show that high-frequency words are recognized earlier (i.e., given less of the speech waveform) than low-frequency words. Tyler (1984) showed the same result for Dutch.

Table 3.1
Lexically reduced vowels in high-frequency words. (After Fidelholz 1975.)

Reduced vowel [fər]		Full vowel [fɔr]	
Word	Count per million	Word	Count per million
forget	148	forfend	<1
forgive	40	forgo	<1

In conclusion, the evidence shows that in both the visual and auditory domains, high-frequency words are accessed more quickly, more easily, and with less input signal than low-frequency words.

3.2.1.3 Lexical Frequency in Production The effects of lexical frequency on production have been measured via a number of tests, including *latency* (the time to start producing a word), *duration* (the time from word onset to word offset), *phonological reduction* (number of deleted or reduced phonemes), rate of speech errors, and others.

The earliest studies focused on duration; indeed, lexical frequency effects on duration in production have been remarked upon for over a hundred years. Schuchardt (1885) noticed, for example, that more frequent words tend to be shorter. Later, Fidelholz (1975) and Hooper (1976) showed that frequent words such as *forget* are more likely to have lexically reduced vowels (e.g., [fər]) than less frequent words such as *forgo* (e.g., [fɔr]) (table 3.1).

While these early studies showing an effect of frequency on a word's phonological makeup are suggestive, they do not confirm that the effect of frequency on lexical production is on-line and productive. It could be that frequent words have reduced vowels and fewer phonemes because of some diachronic fact statically reflected in the lexicon that is only related to on-line production in a complex and indirect way.

To show that frequency plays an active and on-line role in language production, it is necessary to examine the effect of frequency on some dynamic process. One such process is phonological variation; thus, a number of studies have examined whether frequency dynamically affects variation in production. Bybee (2000) examined word-find /t/ and /d/ in a corpus of spoken Chicano English. After excluding the extremely high frequency words *just*, *went*, and *and*, she classified the remaining 2,000 word tokens into two bins, high-frequency (defined as more than 35 per million in the Brown corpus) and low-frequency (fewer than 35 per

million). She showed that final /t/ and /d/ deletion rates were greater in high-frequency words (54.5%) than in low-frequency words (34.3%). Hay (2000) has shown that for complex words, the ratio of the frequency of the derived word and the frequency of its base is an important predictor of processing time.

Gregory et al. (2000) and Jurafsky et al. (2001) provided further evidence that these frequency effects on reduction are on-line, by controlling for a wide variety of contextual factors, and also by investigating the effect of frequency on a word's duration, in addition to its phonological reduction. They examined the duration of words and the percentage of final-consonant deletion in a 38,000-word phonetically transcribed subcorpus from the Switchboard corpus of American English telephone conversations (Godfrey, Holliman, and McDaniel 1992; Greenberg, Ellis, and Hollenback 1996). They used multiple regression to control for contextual factors like segmental context, rate of speech, number of phones, and word predictability.

They first confirmed Bybee's results by analyzing 2,042 word tokens whose full pronunciation ended in /t/ or /d/. After controlling for contextual factors, they found that these final obstruents are more likely to be deleted in more frequent words. High-frequency words (at the 95th percentile of frequency) were 2.0 times more likely to have deleted final /t/ or /d/ than low-frequency words (at the 5th percentile).

Gregory et al. (2000) and Jurafsky et al. (2001) also investigated the effects of frequency on word duration, using 1,412 monosyllabic word tokens ending in /t/ or /d/. They found a strong effect of word frequency on duration. Overall, high-frequency words (at the 95th percentile of frequency) were 18% shorter than low-frequency words (at the 5th percentile).

Taken together, these results suggest that frequency plays an on-line role in lexical production. Duration studies, however, may not be completely convincing. It is possible, for example, that high-frequency words are stored with multiple phonological lexemes (Jurafsky et al. 2001) or with very detailed phonetic information about the length of each phone in each word (Pierrehumbert 2001b).

The most unambiguous evidence for frequency effects in production, then, must come from latency. Oldfield and Wingfield (1965), for example, showed an on-line effect of word frequency on latency of picture-naming times. Presenting subjects with pictures, they found that pictures with high-frequency names were named faster than pictures with low-

frequency names. Wingfield (1968) showed that this effect must be caused by word frequency rather than the frequency of pictured objects, by showing that the effect was not replicated when subjects were asked to recognize but not verbalize picture names. These results were also replicated for Dutch by Jescheniak and Levelt (1994).

In conclusion, more frequent words are accessed more quickly (shorter latency) and are articulated more quickly (shorter duration).

3.2.2 Frequency of Lexical Semantic Form and Lexical Category

Words are ambiguous in many ways. A word can have multiple senses (*bank* can refer to a location alongside a river or a financial institution), multiple lexical categories (*table* can be a noun or a verb), and multiple morphological categories (*searched* can be a participle or a preterite). These different categories of an ambiguous word vary in frequency; for example, the word *table* is more likely to be a noun than a verb. In this section, I summarize experiments showing that the frequency of these categories plays a role in processing.

A number of experiments have shown that the frequency of a particular sense of an ambiguous word plays a role in comprehension. Simpson and Burgess (1985), for example, studied lexical access in the visual domain. Subjects were first presented with an ambiguous prime word (homograph) that had a more frequent sense and a less frequent sense. Subjects then performed lexical decision on targets that were associated with either the more frequent or the less frequent meaning of the homograph prime. Simpson and Burgess found that the more frequent meaning of the homograph caused faster response latencies to related associates, suggesting that the more frequent meaning is retrieved more quickly. This result is robust and has been replicated with many paradigms, including eye fixation times in reading and cross-modal priming. Evidence for the use of word sense frequency in comprehension has also been reported crosslinguistically—for example, in Chinese (Li and Yip 1996; Ahrens 1998).

The frequency of an ambiguous word's syntactic category plays a role in comprehension as well. One class of studies involves sentence processing and human parsing. Gibson (1991) and Jurafsky (1992, 1996) suggest that lexical category frequencies might play a role in the difficulty of processing some garden path sentences. Fox example, (3) and (4) are known to be difficult to process. Gibson suggests that (3) is difficult to process because *man* is much more likely to be a verb than a noun, while

Jurafsky suggests that (4) is difficult because of the lexical category preferences of *complex* (more likely to be an adjective than a noun) and *house* (more likely to be a noun than a verb):

(3) The old man the boats. (from Milne 1982)

(4) The complex houses married and single students and their families. (from Jurafsky 1992, 1996)

Finally, morphological category frequencies play a role in comprehending ambiguous words. Words such as *searched*, *scratched*, *proposed*, and *selected* are ambiguous between a participle and a preterite (simple past) reading. For some of these words, the participle reading is more frequent. For example, the percentage of participle readings for *selected* (in the Brown corpus) is 89%, while the percentage for the simple past readings is 11%. By contrast, the preferences are reversed for *searched*: 78% for simple past readings, and 22% for participle readings. Burgess and Hollbach (1988) suggested that these lexical category probabilities might play a role in disambiguation.

Trueswell (1996) investigated this hypothesis by embedding these verbs in sentences that have a local ambiguity. Each sentence had an initial word sequence like *the room searched* that was syntactically ambiguous between a relative clause reading (compatible with the participle form) and a main verb reading (compatible with the simple past). Trueswell found that verbs with a frequency-based preference for the simple past form caused readers to prefer the main clause interpretation (as measured by longer reading time for a sentence like (5) that required the other interpretation):

(5) The room searched by the police contained the missing weapon.

This suggests that the frequency with which the different morphological categories of a verb occur plays a role in whether one syntactic parse is preferred or not.

In summary, the frequencies of the semantic, syntactic, or morphological categories associated with an ambiguous word play an important role in comprehension. More frequent categories are accessed more quickly and are preferred in disambiguation.

Rather surprisingly, given this robust effect of the frequency of lexical semantic/syntactic category in comprehension, there may not be any such effect in production. Instead, some studies have suggested that frequency effects in lexical production are confined to the level of the word-form or lexeme, rather than the semantic/syntactically defined lemma level.

Both Dell (1990) and Jescheniak and Levelt (1994), for example, studied whether word frequency effects in production take place at the level of the semantic *lemma* or the phonological *word-form*. Finding an effect of frequency for the semantic/syntactic lemma would be the correlate in lexical production of finding an effect of semantic sense or syntactic category in comprehension. Dell (1990) used experimentally elicited speech errors to study word frequency effects. Previous work had shown that low-frequency words are more susceptible to phonological speech errors than high-frequency words. Dell showed that some low-frequency words are not susceptible to phonological speech errors: specifically, low-frequency words (such as *wee*) with a high-frequency homophone (such as *we*). In other words, a low-frequency word that shares a lexeme with a high-frequency word exhibits some of the frequency properties of the high-frequency word. One way to model this result is to store frequency effects only at the lexeme level; the words *we* and *wee* would then share a single frequency node.

Jescheniak and Levelt (1994) used a novel translation task to study word frequency effects. Like Dell, they looked at homophones, in which two distinct lemmas share one lexeme. If frequency effects are localized at the lemma level, accessing a low-frequency lemma in production should have slower latency than accessing a high-frequency lemma; but if frequency effects are localized at the lexeme level, low-frequency and high-frequency lemmas of the same homophone should have identical latencies. This hypothesis cannot be tested with standard paradigms like picture naming, since it is unlikely that both the high-frequency and low-frequency senses of a word are picturable (e.g., neither *we* nor *wee* is obviously picturable). Jescheniak and Levelt therefore used a novel translation latency task: bilingual Dutch subjects were ask to produce the Dutch translation for a visually presented English word, and the translation latency was recorded. For example, subjects saw the English word *bunch*, whose Dutch translation is *bos*. The Dutch word *bos* has another sense, *forest*. If frequencies are stored at the lexeme level, latencies to low-frequency words like *bunch*/*bos* should match latencies to high-frequency words. This is what Jescheniak and Levelt found. Latency to homophones patterned like latency to high-frequency words, and not like latency to low-frequency words.

One important caveat about Dell's (1990) and Jescheniak and Levelt's (1994) results: they crucially rely on the assumption that lexical production is modular. If lexical production is completely interactionist, frequencies could be stored at the lemma level but activation could spread

from the lemma, down to the lexeme, and back up to both lemmas, allowing a low-frequency lemma to act like a high-frequency one. In fact, this is exactly Dell's proposal, and he built a nonmodular computational model to show that the idea is possible. The evidence for a lack of a lemma effect, then, rests only on the evidence for a modular (noninteractionist) model of lexical production.

To help resolve this dilemma, Jurafsky et al. (2001) proposed a different, corpus-based methodology for studying frequency effects in production. They examined the production of ambiguous words like *to* (which can be an infinitive marker (*We had* ***to*** *do it*) or a preposition (*I would have gone* ***to*** *the store*)) and *that* (which can be (at least) a complementizer, a pronoun, or a determiner). Again using the 38,000-word phonetically transcribed subcorpus from the Switchboard corpus of American English telephone conversations, they measured the duration of the function words. They then used multiple regression to control for known factors affecting duration, including rate of speech, segmental context, contextual predictability, and so on, and to test for an effect of lemma frequency on duration. They found that the different pronunciations and durations of these words could be completely accounted for by other factors such as pitch accent and contextual predictability. They found no evidence that lemma frequency affected lexical production.

Thus, although the frequencies of the semantic, syntactic, or morphological categories associated with an ambiguous word play a role in comprehension, preliminary studies suggest that they may not play a similar role in production.

3.2.3 Neighboring Word-to-Word Probabilities

Having looked at frequency effects for single words, let us now turn to evidence that frequency plays a role in more complex and structured relationships between words and syntactic structure. A number of studies show that the probabilistic relationships between neighboring words play a role in both comprehension and production.

Some of these studies looked at raw frequency, while others looked at various probabilistic measures. Researchers have investigated both the conditional probability of a word given the previous word $P(w_i|w_{i-1})$ and the joint probability of two words together $P(w_{i-1}w_i)$. The joint probability of two words is generally estimated from the relative frequency of the two words together in a corpus, normalized by the total number N of word-pair tokens in the corpus (which is one more than the total number

of words in the corpus):

$$P(w_{i-1}w_i) = \frac{Count(w_{i-1}w_i)}{N}. \tag{6}$$

Some experiments use this normalized joint probability, while others simply use the raw joint frequency.

Another common metric is the first-order markov relation: the conditional probability of a word given the previous word (sometimes called the transitional probability (Saffran, Aslin, and Newport 1996; Bush 1999). The conditional probability of a particular target word w_i given a previous word w_{i-1} can be estimated from the number of times the two words occur together $Count(w_{i-1}w_i)$, divided by $Count(w_{i-1})$, the total number of times that the first word occurs:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}. \tag{7}$$

MacDonald (1993) studied the effect of word-pair (joint) frequencies on comprehension. Investigating the processing of a noun followed by a word that is ambiguous between a noun and verb, such as the pair *miracle cures*, she hypothesized that if the noun-noun pair was frequent (like *miracle cures*), its interpretation would be biased toward the noun reading of the second word. She predicted no such bias for infrequent noun-noun pairs (like *shrine cures*). She confirmed this hypothesis by looking at reading time just after the ambiguous word in sentences that were otherwise biased toward a verb reading. For example, subjects spent more time reading the word *people* in (9) than in (8), since the frequent noun-noun phrase in (9) biases the reader toward the noun reading of *cures*, whereas the word *people* is compatible only with the verb reading:

(8) The doctor refused to believe that the *shrine cures* people of many fatal diseases ...

(9) The doctor refused to believe that the *miracle cures* people of many fatal diseases ...

Extending this study, McDonald, Shillcock, and Brew (2001) showed using eye-tracking that bigram probability $P(w_i|w_{i-1})$ is a good predictor of gaze duration on word w_i. Bod (2000b, 2001a) showed in a recognition task that this extends to structures larger than bigrams: frequent three-word (subject-verb-object) sentences proved to be recognized more easily and faster than infrequent three-word sentences.

Table 3.2
Examples of word boundary coronals that are more or less likely to palatalize

High $P(w_i\|w_{i-1})$ More palatalized	Low $P(w_i\|w_{i-1})$ Less palatalized
did you	at you
told you	but you
would you	good you

The psycholinguistic role of word-to-word frequencies or probabilities has also been extensively studied in production. The production studies have generally investigated the effect that frequency or probability given neighboring words has on the phonetic form of a word. The main result is that words in high-frequency word-pairs or high-probability word-pairs are phonetically reduced in some way.

Krug (1998), for example, showed that cliticization is more common in more frequent word-pairs. Bybee and Scheibman (1999) and Bush (1999) found that word boundary coronals are more likely to be palatalized between word sequences with high conditional probabilities, as shown in table 3.2.

Gregory et al. (2000), Jurafsky et al. (2001), and Bell et al. (2001) studied the effect of different kinds of probability on reduction. As mentioned above, they used a phonetically transcribed portion of the Switchboard telephone corpus of American English. They used multiple regression to control for other factors affecting reduction, and looked at more measures of predictability and more measures of reduction. In particular, they looked at both the joint probability and conditional probability of target words with both previous and following words. Confirming earlier studies, they found that words that have a higher probability given neighboring words are reduced. In particular, Gregory et al. (2000) and Jurafsky et al. (2001) found that high-probability content words are shorter in duration and more likely to have final /t/ or /d/ deleted. Jurafsky et al. (2001) and Bell et al. (2001) found that high-probability function words are shorter and undergo more vowel reduction and more coda deletion. All of the studies found more reduction in high-probability words no matter how probability was defined (conditional probability given the previous word or given the next word; joint probability with the previous word or with the next word).

Pan and Hirschberg (2000) have shown that conditional bigram probability correlates highly with location of pitch accent; specifically, pitch accent is more likely to occur on low-probability words. Gregory (2001) has extended this result by showing that conditional probability given previous and following words is a significant predictor of pitch accent even after controlling for other contextual factors such as position in the intonation phrase, part of speech, and number of syllables.

In summary, the probability of a word given the previous or following word plays a role in comprehension and production. Words with a high joint or conditional probability given preceding or following words have shorter durations in production. In comprehension, any ambiguous words in a high-frequency word-pair are likely to be disambiguated consistently with the category of the word-pair itself.

3.2.4 Syntactic Subcategorization Frequencies

Quite a bit of attention has been paid to the next type of probability we will look at: the frequency of the different subcategorization frames of a verb. For example, the verbs *remember* and *suspect* are both subcategorized for either a direct object noun phrase or a sentential complement, as in (10)–(13):

(10) The doctor remembered [$_{NP}$ the idea].

(11) The doctor remembered [$_{S}$ that the idea had already been proposed].

(12) The doctor suspected [$_{NP}$ the idea].

(13) The doctor suspected [$_{NP}$ that the idea would turn out not to work].

While both verbs allow both subcategorization frames, they do so with different frequencies. *Remembered* is more frequently used with a noun phrase complement, while *suspected* is more frequently used with a sentential complement. These frequencies can be computed either from a parsed or a transitivity-coded corpus (Merlo 1994; Roland and Jurafsky 1998) or by asking subjects to write sentences using the verbs (Connine et al. 1984; Garnsey et al. 1997).

Since these frequencies are contingent on the verb, they are the maximum likelihood estimate of the conditional probability of the subcategorization frame given the verb $P(\text{frame}|\text{verb})$. These conditional probabilities have been shown to play a role in disambiguation. For

example, the noun phrase *the idea* is ambiguous in the following sentence prefix:

(14) The doctor suspected the idea ...

The idea could function as a direct object noun phrase complement of *suspected*, or it could be the syntactic subject of an embedded sentential complement. A wide variety of experiments shows that the verb's "bias" (its probabilistic preference for a subcategorization frame) influences which of these two continuations subjects expect.

This idea of subcategorization bias was first suggested in a slightly different context by Fodor (1978), who predicted that a verb's preference for being transitive or intransitive could affect whether the human parser hypothesizes a gap following the verb. Ford, Bresnan, and Kaplan (1982) proposed a generalization of Fodor's hypothesis: that each verb has strengths for different subcategorization frames, that these strengths are based on some combination of frequency and contextual factors, and that these strength-based expectations are used throughout parsing. When they tested this idea by asking subjects in an off-line experiment to perform a forced choice between two interpretations of an ambiguous utterance, they found that some set of subcategorization strengths could be used to predict the interpretation selected by the subjects.

Ford, Bresnan, and Kaplan (1982) did not actually test whether these transitivity preferences were related to frequency in any way. Although Jurafsky (1996) later confirmed that some of the preferences corresponded to Brown corpus frequencies, this was not clear at the time Ford, Bresnan, and Kaplan conducted their study; furthermore, the fact that their experiment was off-line left open the possibility that semantic plausibility or some other factor rather than subcategorization frequency was playing the causal role. Clifton, Frazier, and Connine (1984) tested the model more directly by using the frequency norms collected by Connine et al. (1984) to show that a frequency-based interpretation of transitivity preference predicted quicker understanding in filler-gap sentences, and Tanenhaus, Stowe, and Carlson (1985) showed that anomalous fronted direct objects required extra reading time at transitive-bias verbs, but not intransitive-bias verbs.

Trueswell, Tanenhaus, and Kello (1993) extended these results to show that these frequency-based subcategorization preferences play an on-line role in the disambiguation of various syntactic ambiguities. One experiment was based on cross-modal naming (so called because the stimulus

is auditory while the target is orthographic). Subjects heard a sentence prefix ending in either an S-bias verb (*The old man suspected* ...) or an NP-bias verb (*The old man remembered* ...). They then had to read out loud ("name") the word *him*. Previous research had shown that naming latencies are longer when the word being read is an ungrammatical or unexpected continuation. In Trueswell, Tanenhaus, and Kello's study, naming latency to *him* was longer after S-bias verbs (*The old man suspected ... him*) than after NP-bias verbs (*The old man remembered ... him*). This suggests that subjects preferred the more frequent frame of the verb and were surprised when this preference was overturned, causing longer naming latencies. Trueswell, Tanenhaus, and Kello also confirmed these results with an eye-tracking study that focused on the difference in reading times between sentences with and without the complementizer *that*. Controlled first-pass reading times at the disambiguating verb phrase were longer for NP-bias verbs but not for S-bias verbs, indicating that subjects attached the postverbal noun phrase as a direct object for NP-bias verbs but not for S-bias verbs.

MacDonald (1994) showed that the effect of subcategorization frame frequency also plays a role in resolving a different kind of ambiguity: main clause/relative clause (MC/RR) ambiguities. These ambiguities have been the object of much study since Bever (1970) first pointed out the difficulty of the garden path sentence in (15):

(15) The horse raced past the barn fell.

Until the word *fell*, this sentence is ambiguous between a reading in which *raced* is a main verb and one in which it is a part of a reduced relative clause modifying *the horse*. The difficulty of the sentence is caused by the fact that readers incorrectly select the main verb sense and then are confused when they reach *fell*.

MacDonald (1994) suggested that the subcategorization frequencies proposed by earlier researchers could play a role in explaining processing difficulties in main verb/reduced relative ambiguities. Her test materials used transitive-bias verbs like *push* and intransitive-bias verbs like *move*, in sentences like these:

(16) The rancher could see that the nervous cattle *pushed* into the crowded pen were afraid of the cowboys.

(17) The rancher could see that the nervous cattle *moved* into the crowded pen were afraid of the cowboys.

MacDonald found that corrected reading times in the disambiguation region *were afraid* were longer for intransitive-bias verbs like *move* than transitive-bias verbs like *push*.

Jennings, Randall, and Tyler (1997) extended Trueswell, Tanenhaus, and Kello's (1993) study on the effect of verb subcategorization bias on disambiguation, using a similar cross-modal naming paradigm. One goal of this study was to clear up some potential problems with Trueswell, Tanenhaus, and Kello's materials. But perhaps its most significant result addressed an important issue that no previous research on the role of frequency in syntactic disambiguation had addressed. Previous studies had generally clustered their verb-bias frequency into two bins: high transitive-bias versus low transitive-bias, or high S-bias versus low S-bias. All previous results on syntactic disambiguation, then, were compatible with a model in which subcategorization preferences were represented as a ranked or ordered list, with no link to an actual frequency or probability. Jennings, Randall, and Tyler showed a correlation between the strength of a verb's bias and reading time at the target word. The stronger the verb's bias for one subcategorization frame over the other, the larger the advantage they found in naming latency for the preferred over the nonpreferred continuation.

Despite the many studies of subcategorization frequencies in comprehension, there are no equivalent studies in production. Of course, the frequencies used to model the comprehension studies are derived from production data. But there have been no production tests showing clearly that verb-argument probability plays an active on-line role here, as opposed, say, to merely being correlated with semantics or world knowledge. There is at least one suggestive study, by Stallings, MacDonald, and O'Seaghdha (1998), who note a similarity between sentential complement-taking verbs and heavy-NP shift: in both cases, the verb and the complement can be separated by other material (e.g., sentential complements can be separated from the verb by adverbial expressions: *She said the other day that* ...). By contrast, direct objects cannot be separated in this way from their verbs. In a production experiment based on this contrast, Stallings, MacDonald, and O'Seaghdha showed that verbs that can take either sentential complements or noun phrase direct objects are more likely to undergo heavy-NP shift than verbs that take only noun phrase direct objects. They also showed that verbs that frequently undergo heavy-NP shift elicit slower responses when placed in a nonshifted context. They suggest that each verb is stored with a "shifting disposition"—

Table 3.3
Brown corpus part-of-speech percentages for *that*. (From Juliano and Tanenhaus 1993.)

	Determiner	Complementizer
At start of sentence	35%	11%
After verb	6%	93%

a frequency-based preference for appearing contiguous with its arguments or not.

In summary, the conditional probability of a subcategorization frame given a verb plays a role in disambiguation in comprehension. The higher the conditional probability of the frame, the more it will be preferred in disambiguation. In production, the evidence is less conclusive and awaits further study.

3.2.5 Conditional and Lexicalized Syntactic Frequencies

The subcategorization bias of a verb is a kind of conditional probability: the probability of seeing a noun phrase or a sentence given the verb. A number of experiments have found evidence that sentence comprehension makes use of another kind of conditional probability: the probability of a word or the probability of a lexical category conditioned on previous context or on particular syntactic structure.

Juliano and Tanenhaus (1993) studied the role of the frequency of the different lexical categories of *that*, which can be a determiner, a complementizer, a pronoun, or an intensifier. Overall, the pronoun reading of *that* is more frequent than any of the other readings. But Juliano and Tanenhaus noticed that the frequencies of these different categories depend on the syntactic context, as shown in table 3.3. Accordingly, they conducted a self-paced reading study using sentences like those in (18)–(21). In (19) and (21), *that* must be a complementizer, while in (18) and (20), *that* must be a determiner. The word *diplomat/s* provides the disambiguating information (the plural is compatible only with the complementizer reading).

(18) The lawyer insisted *that* experienced **diplomat would** be very helpful.

(19) The lawyer insisted *that* experienced **diplomats would** be very helpful.

(20) *That* experienced **diplomat would** be very helpful to the lawyer.

(21) *That* experienced **diplomats would** be very helpful made the lawyer confident.

If subjects make use of the conditional probability of the part of speech given the context, they should treat sentence-initial *that* as a determiner and postverbal *that* as a complementizer. This would predict increased reading time for the sentence-initial complementizer reading (21) and for the postverbal determiner reading (18). Juliano and Tanenhaus found just such an interaction: reading times for *would* were longer in (21) and (18) and shorter in (19) and (20). Notice that the simple unconditioned use of the different lexical-category frequencies for *that* would not predict this interaction.

A second piece of evidence that probabilities conditioned on previous structure are used in disambiguation comes from Trueswell, Tanenhaus, and Kello's (1993) experiment discussed above. Recall that these researchers showed that cross-modal naming latency to *him* was longer after hearing S-bias verbs (*The old man suspected ... him*) than after hearing NP-bias verbs (*The old man remembered ... him*), a result that supports the use of verb subcategorization frequencies in comprehension. But in a separate analysis, Trueswell, Tanenhaus, and Kello also showed that the longer naming latency after S-bias verbs is not uniform for all S-bias verbs. It has often been noted that the complementizer *that* is optional after some S-bias verbs. Trueswell, Tanenhaus, and Kello measured the frequency with which each S-bias verb occurred with an explicit *that*, to compute the "*that*-preference" for each verb. They found that this *that*-preference correlated with the increased reading time: the more an S-bias verb expected to be followed by *that*, the longer the latency on naming *him*. Once again, this suggests that subjects are computing the probability of hearing the *that* complementizer given previous structure (in this case, the verb).

A third study supporting the idea of probabilities conditioned on previous structure was carried out by MacDonald (1993). Recall that MacDonald was looking at the processing of a noun followed by a word that can be either a noun or a verb, such as the pair *miracle cures*. She examined the frequency with which the first word occurs as the head of a noun phrase versus the frequency with which it occurs as the modifier of a noun phrase. For example, the noun *warehouse* appears in the Brown corpus more often as a modifier, while *corporation* occurs more often as a head. If subjects make use of this probability, they should treat nouns that are

more frequently heads as complete noun phrases, parsing the following word as a verb; nouns that are more likely to be modifiers should cause the following word to be treated as a noun. MacDonald found that the frequency with which a word occurs as a head versus modifier in the Brown corpus did predict reading time difficulty on the word following these bigrams.

In summary, the conditional probability of a word (like *that*) or a lexical category (like determiner or complementizer) given previous words or structure plays a role in disambiguation. Words or categories with higher conditional probabilities are preferred.

3.2.6 Constructional Frequencies

The frequency effects described so far are all lexical in some way. Indeed, the vast majority of frequency effects that have been studied involve lexical structure. A small number of studies have looked for frequency effects for larger (supralexical) structures, but the results are relatively inconclusive.

For example, studying the effect of idiom frequency on word-by-word reading of Dutch idioms, some of which had syntactic errors (such as agreement errors) inserted in them, d'Arcais (1993) found that subjects were able to locate the errors more quickly in frequently than in less frequently occurring idioms.

A number of researchers have suggested that one factor contributing to the difficulty of the main verb/reduced relative ambiguity is the relative rarity of reduced relative clauses. In a norming study for an experiment using 32 verbs, Tabossi et al. (1994) checked 772 sentences from the Brown corpus containing *-ed* forms of the verbs. They found that the verb occurred as part of a simple main clause in 37% of the sentences, a relative clause in 9%, and a reduced relative clause in 8%.

Jurafsky (1996), McRae, Spivey-Knowlton, and Tanenhaus (1998), and Narayanan and Jurafsky (1998), among others, showed that (various) models that include the corpus-based frequency of the main clause versus reduced relative construction are able to model certain reading time effects in main clause/reduced relative sentences such as (15), repeated here as (22):

(22) The horse raced past the barn fell.

Jurafsky, for example, showed that the stochastic context-free grammar (SCFG) probability for the main clause parse was significantly lower than

the SCFG probability for the reduced relative parse because of two factors: first, the reduced relative construction includes one more SCFG rule, and second, this SCFG rule introducing the reduced relative structure has a very low probability.

A series of studies by Mitchell and colleagues (Mitchell 1994; Cuetos, Mitchell, and Corley 1996) has focused on a model of disambiguation called *tuning*. Tuning models claim that people tabulate every ambiguity they encounter, together with the disambiguation decision. Future disambiguation decisions are based on choosing the most likely previously chosen disambiguation for the ambiguity. Tuning models thus claim that syntactically ambiguous sentences are resolved to whichever choice has been made more often in the past. As a simplifying assumption, Mitchell and colleagues assume that the frequency of this choice is isomorphic to the total frequency of the structures in the language.

I discuss tuning models in this section because although such models could hypothetically apply to any kind of disambiguation, all research so far has focused on the frequency of two specific complex syntactic constructions. In particular, Cuetos, Mitchell, and Corley (1996) looked at ambiguities like those in (23), where a relative clause *who was on the balcony* can attach to either the first of two noun phrases (*the servant*) or the second (*the actress*) (not counting the subject *someone*):

(23) Someone shot [$_{NP_1}$ the **servant**] of [$_{NP_2}$ the **actress**] *who was on the balcony*.

Figure 3.1 shows a simplified schematic of the two parse trees whose frequency is being computed.

Cuetos, Mitchell, and Corley (1996) found crosslinguistic differences in disambiguation preference between English and many other languages,

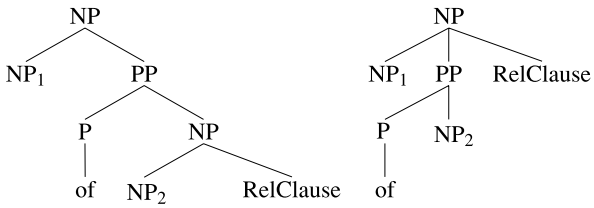

Figure 3.1
The tuning hypothesis predicts that frequencies are stored for these parses for *Someone shot the servant of the actress who was on the balcony*, with attachment of the relative clause to NP_2 (*left*) and NP_1 (*right*)

including Spanish and Dutch. Supporting the tuning hypothesis, English speakers preferred to attach relative clauses to NP_2, and the NP_2 attachment construction shown on the left of figure 3.1 was more common in a corpus. Also supporting the hypothesis, Spanish speakers preferred to attach relative clauses to NP_1, and the NP_1 attachment shown on the right of figure 3.1 was more frequent in a Spanish corpus. But more recent studies have cast doubt on the link between the frequency of the two constructions and the disambiguation decisions. Studies on three-site relative clause ambiguities in English have not found a link between corpus frequency of these (very) complex constructions and disambiguation preference (Gibson, Schütze, and Salomon 1996). Another study found that Dutch speakers preferred to attach relative clauses to NP_1, but that the NP_2 attachment construction shown on the left of figure 3.1 was more common in a corpus (Mitchell and Brysbaert 1998).

Still other studies, however, have shown that human preferences do match corpus preferences when the animacy of the NP_1 is held constant (e.g., Desmet, Brysbaert, and Baecke, in press). Since corpora are used to estimate frequencies in most probabilistic models, this is an important result; I will return to this issue in section 3.4.3. But since this control factor concerned the semantics of the noun phrases, it suggests that a purely structure-frequency account of the tuning hypothesis cannot be maintained.

It has also been shown (Bod 2000b, 2001a) that frequent three-word (subject-verb-object) sentences (e.g., *I like it*) are recognized more easily and quickly than infrequent three-word sentences (e.g., *I keep it*), even after controlling for plausibility, word frequency, word complexity, and syntactic structure. These results suggest that frequent sentences or at least some structural aspects of these frequent sentences are stored in memory.

In addition to studies of the complex structures posited by the tuning hypothesis, which do not show strong evidence for frequency effects in comprehension, there have been some studies on frequency effects for simpler syntactic structure in production. Bates and Devescovi (1989) performed a crosslinguistic series of production studies that attempted to control for semantic and pragmatic factors. They found that relative clauses, which are generally more frequent in Italian than in English, occurred more frequently in Italian in their production study even after these controls. They suggest that the frequency of the relative clause construction in Italian may play a role in its being selected in production.

In conclusion, while some studies seem to suggest that the frequency of larger nonlexical syntactic structures plays a role in disambiguation, the evidence is still preliminary and not very robust. None of the studies that found an effect of nonlexical syntactic or idiom structure did so after carefully controlling for lexical frequencies and two-word or three-word frequencies. Although Bod's (2001a) results clearly point to storage of three-word chunks, it is not necessarily higher-level structure that is playing a causal role. But of course complex constructions are much less frequent than words, and so we expect frequency effects from larger constructions to be harder to find. This remains an important area for future research.

3.2.7 Summary of Psycholinguistic Results on Frequency and Probability

Frequency plays a key role in both comprehension and production, but solid evidence exists only for frequency related in some way to lexical items, or to the relationship between lexical items and syntactic structure.

High-frequency words are recognized more quickly, with less sensory input, and with less interference by neighbors than low-frequency words. High-frequency words are produced with shorter latencies and shorter durations than low-frequency words. Low-frequency words are more subject to phonological speech errors.

The frequencies of various lexical categories a word belongs to play a role in language processing. For words that are morphologically, syntactically, or semantically ambiguous, the more frequent morphological category, part of speech, or sense is accessed more quickly and is preferred in disambiguation. But this effect of lexical semantic/syntactic category does not seem to extend to production.

The frequency of multiple-word structures plays a role in both comprehension and production. Frequent word-pairs or idioms are more quickly accessed and/or preferred in disambiguation. Frequent word-pairs or words that have a high markov bigram probability given neighboring words are shorter in duration and phonologically more reduced.

Various kinds of conditional probabilities play a role in comprehension and production. For verbs that have more than one possible syntactic subcategorization, the more frequent subcategorization frame is preferred in disambiguation. The probability that a verb appears separated from its complement plays a role in production. For words that can belong to more than one part of speech, the part of speech with higher conditional probability given the preceding part of the sentence is preferred.

Finally, a frequency effect for other, larger syntactic structures, while not disproved, remains to be shown.

Although I have focused only on knowledge for which a frequency effect has been found, many other kinds of knowledge of course play a role in probabilistic evidence-combination. One of these is the relationship between lexical and thematic knowledge. For example, animate nouns are more likely to be agents, while inanimate nouns are more likely to be patients; the word *cop* is more likely to be the agent of the verb *arrested* than is the noun *crook*. Many studies have shown that this kind of thematic role information plays a role in comprehension (Trueswell, Tanenhaus, and Garnsey 1994; Garnsey et al. 1997; McRae, Spivey-Knowlton, and Tanenhaus 1998).

3.3 Probabilistic Architectures and Models

Having shown that frequencies of linguistic structure, especially linguistic structure related to lexical items, play a role in language processing, in this section I turn to probabilistic architectures for modeling these frequency effects. Practically all of these models address the process of comprehension, most of them focusing on syntactic comprehension. I will discuss a few preliminary directions toward probabilistic models of production.

3.3.1 Constraint-Based Models

A large class of experimental and modeling work in sentence comprehension belongs to the *constraint-based* (sometimes *constraint-based lexicalist*) framework (Spivey-Knowlton, Trueswell, and Tanenhaus 1993; MacDonald, Pearlmutter, and Seidenberg 1994; Trueswell and Tanenhaus 1994; Trueswell, Tanenhaus, and Garnsey 1994; Spivey-Knowlton and Sedivy 1995; McRae, Spivey-Knowlton, and Tanenhaus 1998; Seidenberg and MacDonald 1999; Kim, Srinivas, and Trueswell 2002). Specific models differ in various ways, but constraint-based models as a class focus on the interactions of a large number of probabilistic constraints to compute parallel competing interpretations.

Much work in the constraint-based framework has focused on experiments showing that certain frequency-based constraints play a role in sentence processing, via either regression or full factorial analysis of reading time data. Above, I summarized the results of a number of these experiments on the roles of verb bias, collocation frequencies, and so on, in

sentence comprehension. The constraint-based framework includes some computational models in addition to experimental results. In general, these are neural network models that take as input various frequency-based and contextual features, which they combine via activation to settle on a particular interpretation (Burgess and Lund 1994; Pearlmutter et al. 1994; Spivey-Knowlton 1996; Tabor, Juliano, and Tanenhaus 1997; Kim, Srinivas, and Trueswell 2002).

I have chosen one of these models to describe, the competition-integration model of Spivey-Knowlton (1996), because it has been most completely implemented; because it, more than other such models, is clearly intended to be probabilistic; and because it has been widely tested against experimental results from a number of reading time studies (McRae, Spivey-Knowlton, and Tanenhaus 1998; Spivey and Tanenhaus 1998; Tanenhaus, Spivey-Knowlton, and Hanna 2000). The input to this model is a set of probabilistic features like the bias for main clauses versus reduced relatives, the verb's preference for participle versus preterite, the contextual support for a particular interpretation, and so on. Some input features are derived from frequencies; others come from rating studies. All features are then normalized to estimate a probabilistic input feature varying between 0 and 1. The model uses a neural network, shown in figure 3.2, to combine these constraints to support alternative interpretations in parallel. Each syntactic alternative is represented by a prebuilt localist node in a network; thus, the network models only the disambiguation process itself rather than the generation or construction of syntactic alternatives. The alternatives compete until one passes an activation threshold.

Each interpretation receives activation from the constraints, which is then fed back to the constraint nodes within each cycle of competition. The algorithm first normalizes each pair of constraints. Let $C_{i,a}$ be the activation of the ith constraint node connected to the ath interpretation node. $C'_{i,a}$ will be the normalized activation; the activation of each constraint thus ranges from 0 to 1:

$$C'_{i,a} = \frac{C_{i,a}}{\sum_a C_{i,a}}. \tag{24}$$

The activation I_a from the constraints to interpretation a is a weighted sum of the activations of the constraints, where w_i is the weight on constraint i:

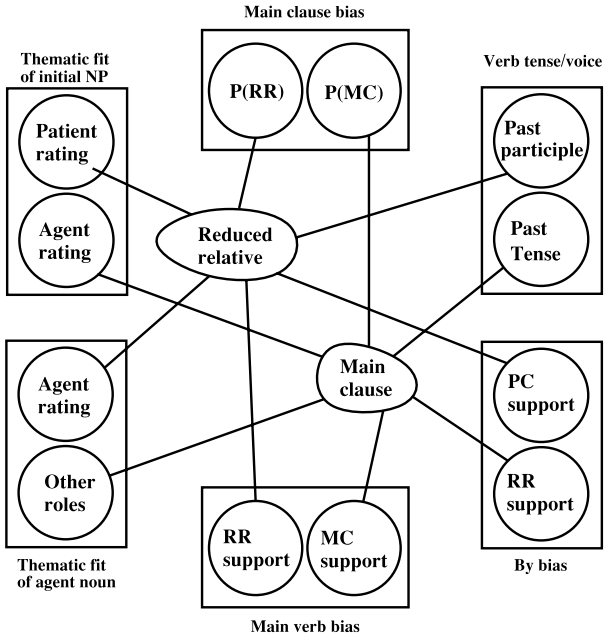

Figure 3.2
A schematic of the competition-integration model. (From McRae, Spivey-Knowlton, and Tanenhaus 1998.)

$$I_a = \sum_i w_i \times C'_{i,a}. \tag{25}$$

Finally, the interpretations send positive feedback to the constraints:

$$C_{i,a} = C'_{i,a} + I_a \times w_i \times C'_{i,a}. \tag{26}$$

These three steps are iterated until one interpretation reaches criterion. Reading time is modeled as a linear function of the number of cycles it takes an interpretation to reach criterion.

This model accounts for reading time data in a number of experiments on disambiguation of main verb/reduced relative ambiguities (McRae, Spivey-Knowlton, and Tanenhaus 1998; Spivey and Tanenhaus 1998; Tanenhaus, Spivey-Knowlton, and Hanna 2000). Let us look at McRae, Spivey-Knowlton, and Tanenhaus's (1998) study, which included two experiments. The first was a sentence completion experiment. For each verb in their study, McRae, Spivey-Knowlton, and Tanenhaus had subjects complete four sentence fragments like these:

The crook arrested

The crook arrested by

The crook arrested by the

The crook arrested by the detective

For each fragment, they measured the proportion of reduced relative clause completions. They then showed that combining a number of probabilistic factors via the competition-integration model correctly predicted the completion preferences for main clauses versus reduced relatives.

McRae, Spivey-Knowlton, and Tanenhaus (1998) also showed that the thematic fit of a subject with the verb plays a role in reading time. Consider the difference between good agents for *arrested* (e.g., *cop: The cop arrested* ...) and good patients for *arrested* (e.g., *crook*). Figure 3.3 shows that controlled human reading time for good agents like *cop* gets longer after reading the *by* phrase (requiring *cop* to be a patient), while controlled reading time for good patients like *crook* gets shorter.[1] McRae, Spivey-Knowlton, and Tanenhaus again showed that the competition-integration model predicts this reading time difference.

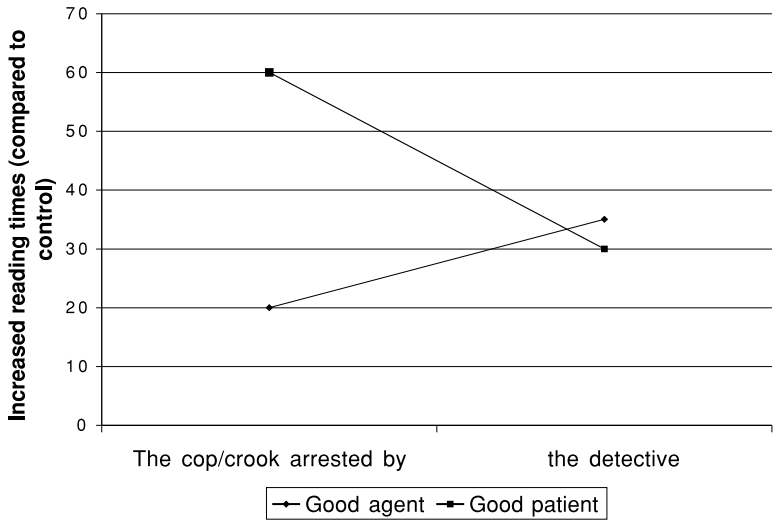

Figure 3.3
Corrected self-paced reading times for words in these regions. (From figure 6 of McRae, Spivey-Knowlton, and Tanenhaus 1998.)

3.3.2 Rational and Utility-Based Probabilistic Models

3.3.2.1 The Competition Model There are remarkable similarities between some of the earliest and some very recent probabilistic models of sentence processing. They all attempt to combine the ideas of probability with utility, cost, or other ideas of rationality in processing.

The *competition* model (MacWhinney, Bates, and Kliegl 1984; MacWhinney and Bates 1989) may have been the first probabilistic model of sentence processing. Its goal is to map from the "formal" level (surface forms, syntactic constructions, prosodic forms, etc.) to the "functional" level (meanings, intentions). Since input is ambiguous and noisy, the model assumes that the sentence processor relies in a probabilistic manner on various surface cues for building the correct functional structures. The model focuses on how these cues probabilistically combine to suggest different interpretations, and on how these probabilities differ from language to language. Consider the problem of assigning agent and patient roles to noun phrases in an input sentence. An English-speaking comprehender relies heavily on word order cues in making this mapping, while a German speaker relies more heavily on morphological (case) cues.

The competition model formalizes the notion of cues via *cue validity*, which is generally defined in this model as a combination of *cue availability* and *cue reliability*. Consider the task of identifying an interpretation i given a cue c. Cue availability is defined by Bates and MacWhinney (1989) as the ratio of the cases in which the cue is available to the total number of cases in a domain. Probabilistically, we can think of this as an estimate of the prior probability of a cue. Cue reliability is defined by Bates and MacWhinney as the ratio of cases in which a cue leads to the correct conclusion to the number of cases in which it is available. Probabilistically, this relative frequency is the maximum likelihood estimate of $P(i|c)$. If cue availability and cue reliability are combined by multiplication, as suggested by McDonald (1986), then cue validity $v(c,i)$ of cue c for interpretation i works out to be the joint probability of cue and interpretation:

$$v(c,i) = \text{availability}(c) \times \text{reliability}(c) = P(c) \times P(i|c) = P(c,i). \qquad (27)$$

The competition framework views cue validity as an objectively correct value for the usefulness of a cue in a language, derived from corpora and studies of multiple speakers. *Cue strength* is the subjective property of a

single human language user, the probability that the language user attaches to a given piece of information relative to some goal or meaning. This use of joint probability as a measure of cue strength seems to be equivalent to the cue strength used in memory models like SAM (Search of Associative Memory) (Raaijmakers and Shiffrin 1981; Gillund and Shiffrin 1984).

How do cues combine to support an interpretation? McDonald and MacWhinney (1989) formalize cue combination by assuming that the contribution of each cue toward an interpretation is independent and that cue strengths vary between 0 and 1. Given these assumptions, they propose the following equation for cue combination, where A and B are interpretations and C is the set of all cues $c_1 \ldots c_n$.

$$P(A|C) = \frac{\prod_i P(A|c_i)}{\prod_i P(A|c_i) + \prod_i P(B|c_i)}. \tag{28}$$

The numerator in (28) factors the probability $P(A|C)$ into separate terms $P(A|c_i)$ for each of the individual cues c_i, while the denominator acts as a normalizing term. Multiplying the factors of the individual cues c_i implies that they are independent (recall equation (6) in chapter 2). Assuming that the contribution of each cue is independent is a simplifying assumption that resembles the "naive Bayes" independence assumption often used in the categorization literature.

The competition model also considers another kind of validity in addition to cue validity: *conflict validity*. Conflict validity is based on how useful a cue is in a competition situation. It is defined (Bates and MacWhinney 1989) as the number of competition situations in which a cue leads to a correct interpretation divided by the number of competition situations in which that cue participates. Thus, the absolute frequency or validity of a cue is not as important as how useful the cue is in disambiguation situations. Conflict validity is thus related to the idea of discriminative training in machine learning, and to the tuning hypothesis that ambiguities are resolved to whichever interpretation has been chosen more frequently in the past (Mitchell 1994; Cuetos, Mitchell, and Corley 1996).

Finally, the competition model also considers factors related to the cost of a cue. For example, certain cues may be difficult to perceive ("perceivability cost"), or holding them until they are integrated may use up short-term memory ("assignability costs").

3.3.2.2 Rational Models Anderson (1990) proposed a rational model for human cognitive processing. The rational framework claims that human cognitive processing makes optimal use of limited resources to solve cognitive problems. The optimal solution to many problems of decision given noisy data and limited resources is known to be probabilistic. Anderson thus applies a probabilistic formulation of his rational model to the task of modeling human memory and categorization. In the course of doing this, he shows how his model explains some results in lexical access.

Anderson assumes that a rational system for retrieving memory would retrieve memory structures serially ordered by their probabilities of being needed p, and would consider the gain G associated with retrieving the correct target and the cost C of retrieving the item. Such a memory should stop retrieving items when

$$pG < C. \tag{29}$$

Anderson shows that by making certain strong independence assumptions, it is possible to produce a relatively straightforward equation for $P(A|H_A\&Q)$, the probability that memory trace A is needed, conditional on some history H_A of its being relevant in the past, and context Q. Let i range over elements that make up the context Q. Anderson gives the following equation:

$$P(A|H_A\&Q) = P(A|H_A) * \prod_{i \in Q} \frac{P(i|A)}{P(i)}. \tag{30}$$

Equation (30) says that we can estimate the posterior probability that A is needed from two terms: a term representing A's past history (how frequent it was and how often it was needed) and a term representing the ratio of the conditional probabilities of the cues given that the structure is relevant or not relevant. Anderson proposes that an increase in need probability $P(A|H_A\&Q)$ maps monotonically into higher recall probability and faster latency (reaction time). He shows that his model predicts a number of basic results in recall rates and latencies, including some results in lexical processing. For example, his model predicts the result that low-frequency words are better recognized than high-frequency words (Kintsch 1970). This sketch of equation (30) and the model has been necessarily but unfortunately brief; the interested reader should turn to Anderson 1990.

Chater, Crocker, and Pickering (1998) apply Anderson's rational model to sentence processing. They first propose that the goal of the human parser is to maximize the probability of obtaining the globally correct parse. Extending Anderson's serial model of memory, they assume that as each word is input, the parser considers all possible parses in series. But they suggest that ordering these parses just by their probabilities may not be optimal. A parse that seems (locally) to be optimal may turn out to be the wrong parse. A serial parser would garden-path at this point, then have to backtrack and reanalyze a sentence. They therefore suggest that an optimal parser would need to include the cost of this backtracking in its algorithm for choosing a parse to follow at points of ambiguity. In particular, they suggest that it is important to balance the probability of a hypothesis, how long it would take to *settle* on the hypothesis (i.e., follow it and build the parse tree), and how long it would take to *escape* from the hypothesis (test and reject it). Given this assumption, they show that a serial parser should first consider the hypothesis H_i with the highest value of the following function f of H_i:

$$f(H_i) = P(H_i) \times P(\text{settle } H_i) \times \frac{1}{1 - P(\text{escape } H_i)}. \tag{31}$$

This proposal—that the function f, rather than unaugmented probability p, is the utility function maximized by the human parser—is an intriguing claim about sentence processing that remains to be tested.

3.3.3 Markov Models of Lexical Category Preference

The previous sections motivated the use of probabilities as a tool for ranking interpretations or actions taken by the human comprehension mechanism. Let us turn now to the details of some probabilistic models. This section and the next two describe increasingly sophisticated probabilistic models of lexical category and syntactic disambiguation: hidden markov models, stochastic context-free grammars, and Bayesian belief networks. All of these are instances of what are often called *graphical models* (Jordan 1999).

Corley and Crocker (1996, 2000) focus on the problem of lexical category disambiguation as part of human sentence processing. They propose that human lexical category disambiguation can be modeled by a hidden markov model (HMM) part-of-speech tagging algorithm (or a variant of the HMM algorithm known as the Church tagger—Church 1988).

Table 3.4
Hidden markov model tag probabilities for *race*. (From Jurafsky and Martin 2000.)

Words	$P(t_i\|t_{i-1})P(w_i\|t_i)$	P
to/INF race/VERB	P(Verb|InfTo) × P(race|Verb)	.00001
to/INF race/NOUN	P(Noun|InfTo) × P(race|Noun)	.000007

HMM taggers are used to compute the probability of a sequence of part-of-speech tags given a sequence of words. For example, given the sequence of words in (32), a tagger would produce the series of tags in (33):

(32) the miracle cures

(33) Det Noun Noun

HMM taggers rely on a very simple intuition: given a word, choose its most likely tag in context. They operationalize "most likely" by using only two probabilistic sources of knowledge: the probability of a word given a lexical category tag $P(w_i|t_i)$ and the probability of one lexical category tag following another $P(t_i|t_{i-1})$.

For example, the word *race* can be a verb or a noun. The noun is vastly more frequent; but verbs are more common after the infinitive marker *to*. Table 3.4 (taken from Jurafsky and Martin 2000, with probabilities from the combined Brown and Switchboard corpora) shows that an HMM tagger correctly chooses the part of speech verb in the context *to race*.

HMM taggers actually produce the most likely sequence of tags $\hat{t}_1^n$ for an entire sentence or sequence of words of length n rather than just for a single word $\hat{t}_i$. We can use the function $\text{argmax}_x\, f(x)$, which returns the x that maximizes $f(x)$, to write the equation for what the tagger is maximizing:

$$\hat{t}_1^n = \underset{t_1^n}{\text{argmax}}\ P(t_1^n|w_1^n). \tag{34}$$

This equation can be rewritten by Bayes' rule (chapter 2, equation (10)), as

$$\hat{t}_1^n = \underset{t_1^n}{\text{argmax}}\ \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)}. \tag{35}$$

Since an HMM tagger is choosing the most likely tag sequence for a fixed

Table 3.5
Corley and Crocker's (1996) probability computation for hidden markov model tagger on data from Juliano and Tanenhaus 1993

Context	Part of speech	$P(t_i\|t_{i-1})P(w_i\|t_i)$	P
Sentence-initial	Comp	P(Comp|#)P(that|Comp)	.0003
	Det	P(Det|#)P(that|Det)	**.0011**
Following verb	**Comp**	P(Comp|Verb)P(that|Comp)	**.023**
	Det	P(Det|Verb)P(that|Det)	.00051

set of words w_1^n, we can drop the denominator term, producing

$$\hat{t}_1^n = \underset{t_1^n}{\text{argmax}}\ P(w_1^n|t_1^n)P(t_1^n). \tag{36}$$

The HMM tagger makes two large simplifying assumptions: first, that the probability of a word depends only on its own tag, and not any neighboring tags; and second, that the words are independent of each other. This results in the following equation by which a bigram tagger estimates the most probable tag sequence:

$$\hat{t}_1^n = \underset{t_1^n}{\text{argmax}}\ P(t_1^n|w_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}). \tag{37}$$

Corley and Crocker (1996, 2000) show that this probabilistic model accounts for a number of the psycholinguistic results discussed above. For example, they model Juliano and Tanenhaus's (1993) finding that subjects seem to treat sentence-initial *that* as a determiner, but postverbal *that* as a complementizer. Table 3.5 shows that the HMM probabilities predict a determiner reading sentence-initially, but a complementizer reading after a verb. Corley and Crocker also show that their tagging model accounts for three other results on lexical category disambiguation.

3.3.4 Stochastic Context-Free Grammars

Jurafsky (1996) proposed a probabilistic model for syntactic disambiguation. His probabilistic parser kept multiple interpretations of an ambiguous sentence, ranking each interpretation by its probability. The probability of an interpretation was computed by multiplying two probabilities: the stochastic context-free grammar (SCFG) "prefix" probability of the currently seen portion of the sentence, and the "valence" (syntactic/semantic subcategorization) probability for each verb.

A stochastic context-free grammar, first proposed by Booth (1969), associates each rule in a context-free grammar with the conditional probability that the left-hand side expands to the right-hand side. For example, the probability of two of the expansions of the nonterminal NP, computed from the Brown corpus, is

[.42] NP → Det N,

[.16] NP → Det Adj N.

Jurafsky's model was on-line, using the left-corner probability algorithm of Jelinek and Lafferty (1991) and Stolcke (1995) to compute the SCFG probability for any initial substring (or "prefix") of a sentence.

Subcategorization probabilities in the model were also computed from the Brown corpus. For example, the verb *keep* has a probability of .81 of being bivalent (*keep something in the fridge*) and a probability of .19 of being monovalent (*keep something*).

While the model kept multiple interpretations, it was not fully parallel. Low-probability parses were pruned via beam search. Beam search is an algorithm for searching for a solution in a problem space that only looks at the best few candidates at a time. The name derives from the metaphor of searching with a flashlight; only things that lie within the beam of the light are retained. The use of beam search in the algorithm, rather than full parallel search, means that the model predicts extra reading time (the strong garden path effect) when the correct parse has been pruned away and the rest of the sentence is no longer interpretable without reanalysis.

Jurafsky (1996) showed that this model could account for a number of psycholinguistic results regarding parse preferences and garden path sentences. For example, the corpus-based subcategorization and SCFG probabilities for *keep* and other verbs like *discuss* correctly modeled the preferences for these verbs in the off-line forced-choice experiment carried out by Ford, Bresnan, and Kaplan (1982). The SCFG grammar also correctly modeled the misanalysis of garden path sentences like (38), by claiming that the correct parse (in which *houses* is a verb) gets pruned:

(38) The complex houses married and single students and their families.

Finally, the combination of SCFG probability and subcategorization probability modeled the garden path effect for preferentially transitive verbs like *race* and the weaker garden path effect for preferential intransitive verbs like *find*:

(39) The horse raced past the barn fell.

(40) The bird found in the room died.

In summary, Jurafsky's (1996) parser has the advantages of a clean, well-defined probabilistic model, the ability to model the changes in probability word by word, a parallel processing architecture that can model both lexical and syntactic processing, accurate modeling of parse preference, and a probabilistic beam search architecture that explains difficult garden path sentences. The model has many disadvantages, however. First, it only makes very coarse-grained reading time predictions; it predicts extra reading time at difficult garden path sentences, because the correct parse falls out of the parser's beam width. It does not make fine-grained reading time predictions of any kind. In addition, although the description of the model claims that the interpreter can combine probabilistic information of any sort, the model as described specifies only SCFG and subcategorization probabilities. Finally, the model has been tested on only a handful of examples.

Crocker and Brants (2000) propose a probabilistic model of sentence processing that is similar to Jurafsky's (1996) but that, unlike Jurafsky's, is designed to have wide coverage and efficient scalability. Their *incremental cascaded markov model* (ICMM) is based on the broad-coverage statistical parsing techniques of Brants (1999). ICMM is a maximum likelihood model, which combines stochastic context-free grammars with HMMs, generalizing the HMM/SCFG hybrids of Moore et al. (1995). The original nonincremental version of the model constructs a parse tree layer by layer, first at the preterminal (lexical category) nodes of the parse tree, then at the next higher layer in the tree, and so on. In the incremental version of the model, information is propagated up the layers of the model after reading each word. Each markov model layer consists of a series of nodes corresponding to phrasal (syntactic) categories like NP or AdvP, with transitions corresponding to trigram probabilities of these categories. The output probabilities of each layer are structures whose probabilities are assigned by a stochastic context-free grammar. Figure 3.4 shows a part of the first markov model layer for one sentence. Each markov model layer acts as a probabilistic filter, in that only the highest-probability nonterminal sequences are passed up from each layer to the next higher layer. The trigram transition probabilities and SCFG output probabilities are trained on a treebank.

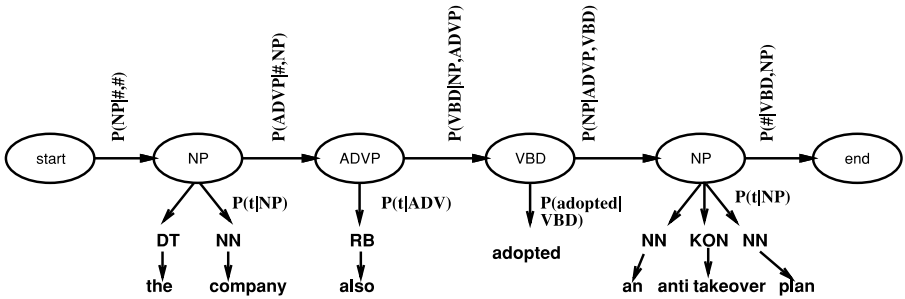

Figure 3.4
Part of the first markov model layer for one sentence. (From Crocker and Brants 2000.) The letter t indicates the subtrees generated by the stochastic context-free grammar. For example, $P(t|\text{NP})$ is the conditional probability of the subtree NN $\rightarrow$ *company* given the NP.

Crocker and Brants's (2000) model accounts for a number of experimental results on human parsing. For example, because the ICMM is a generalization of Corley and Crocker's (1996) model, it handles the same lexical category effects described in the previous section, including Juliano and Tanenhaus's (1993) conditional probability effect of *that*.

The ICMM also models a finding by Pickering, Traxler, and Crocker (2000), who were looking at disambiguation of the role of noun phrases like *his goals* in NP/S ambiguities like the following:

(41) The athlete realized $[_{\text{NP}}$ his goals] at the Olympics.

(42) The athlete realized $[_{\text{S}}[_{\text{NP}}$ his goals] were out of reach].

Realize is an S-bias verb. Nonetheless, Pickering, Traxler, and Crocker showed that readers must be considering the NP interpretation of *his goals*. They did this by creating pairs of sentences with sentential complements. In one sentence, (43), the noun phrase *potential* was a plausible direct object for *realize*. In the other sentence, (44), the noun phrase *her exercises* was not a plausible direct object:

(43) The young athlete realized her potential one day might make her a world-class sprinter.

(44) The young athlete realized her exercises one day might make her a world-class sprinter.

Pickering, Traxler, and Crocker showed that reading time was delayed on the phrase *might make her* after the implausible direct object *her exercises*

but not after the plausible direct object *her potential*. In order to be influenced by the plausibility of the direct object, the human parser must be building the direct object interpretation, despite the S-bias of the verb *realized*.

Crocker and Brants (2000) use the structure of the SCFG to model this result. Sentential complements involve one more SCFG rule than direct objects (the rule VP $\rightarrow$ S). The probability of the sentential complement will thus be lower than it would be otherwise; since probabilities are less than 1, multiplying by an additional rule lowers the probability of a parse. Thus, Crocker and Brants's model predicts that the probability of the direct object reading of (42) is actually higher than the probability of the sentential complement reading.

Like Jurafsky (1996) and Crocker and Brants (2000), Hale (2001) proposes to model human sentence processing via a probabilistic parser based on SCFG probabilities. But Hale's model offers an important new contribution: much more fine-grained predictions about parsing difficulty and hence reading time. Hale proposes that the cognitive effort needed to integrate the next word into a parse is related to how surprising or unexpected that word is. The *surprisal* of a word is an alternate term in information theory for the word's information value (Attneave 1959), which can be computed by the negative log of its probability:

$$h(w_i) = -\log P(w_i). \tag{45}$$

Thus, Hale's proposal is that reading times at a word are a function of the amount of information in the word. A word that is surprising or informative (has a large negative log probability and hence a large positive information content) will cause extended reading times and hence a garden path sentence.

How should the probability $P(w_i)$ be estimated? This is of course a cognitive modeling question; the appropriate probability is whatever people can be shown to use. Hale proposes to use a simple syntax-based probability metric: the conditional SCFG probability of the word given the parse tree of the preceding prefix.

The conditional probability of a word given the previous structure can be computed from an SCFG by using the *prefix* probability. Recall that the prefix probability is the probability of an initial substring of a sentence given the grammar. Unlike computing the probability of an entire sentence, computing the probability of a prefix is somewhat complex, since it involves summing over the probability of all possible recursive

structures before the parser knows exactly how many recursions will be seen. Jelinek and Lafferty (1991) show how this prefix probability can be computed, and Stolcke (1995) shows how this computation can be integrated into a probabilistic Earley parser. If α_i represents the prefix probability of words $w_0 w_1 \ldots w_i$, then the conditional probability of a new word given all the previous words is

$$P(w_i | w_1, w_2 \ldots w_{i-1}) = \frac{P(w_1 \ldots w_i)}{P(w_1 \ldots w_{i-1})} = \frac{\alpha_i}{\alpha_{i-1}}. \tag{46}$$

Hale's proposal is that reading times at a word will be proportional to the information value assigned by this probability, or

$$h(w_i) = -\log \frac{\alpha_i}{\alpha_{i-1}}. \tag{47}$$

Hale actually gives a different but equally valid way of thinking about this equation. He proposes that the cognitive effort needed for parsing any sequence of words is proportional to the total probability of all the structural analyses that are incompatible with that sequence. That is, cognitive effort, and particularly the garden path effect, occurs wherever the parser disconfirms potential parses that together comprise a large probability mass. The simplest way to measure the amount of probability mass that is disconfirmed is to look at the amount of probability mass in the prefix leading up to the previous word that is no longer in the prefix leading up to the current word, which is the difference between α_i and α_{i-1}.

Hale shows that his model predicts the large increases in reading time corresponding to two well-known cases of processing difficulty: reduced relative clauses and subject-object asymmetries. First, he shows that the surprise at the word *fell* is very high in the reduced relative garden path sentence (48) by hand-building a mini context-free grammar for the rules in the sentence and setting the probabilities from a sample of the Penn Treebank. Figure 3.5 shows the prediction of extra surprise, hence extra reading time at *fell*.

(48) The horse raced past the barn fell.

Hale's model predicts a large increase in reading time at *fell* because the probability of *fell* is extremely low. In this sense, Jurafsky's (1996) pruning-based model is just a special case of Hale's. Jurafsky's model predicts extra reading time because the probability of *fell* is zero; the

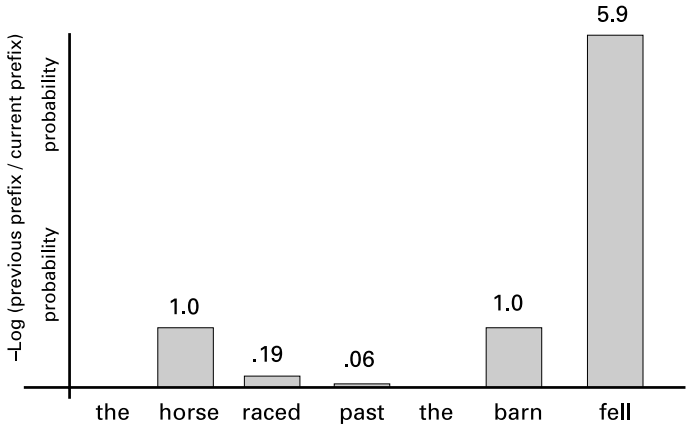

Figure 3.5
Hale's (2001) prediction of reading time based on surprise values computed from a simple stochastic context-free grammar

potential parse that could have incorporated *fell* was pruned away. Hale's model is thus able to make more fine-grained reading time predictions than Jurafsky's.

These more fine-grained predictions can be seen in Hale's probabilistic explanation for a second source of processing difficulty: subject-object relative asymmetry. Many researchers had noticed that object relative clauses (49) are more difficult to parse than subject relative clauses (50); see Gibson 1998 for a summary of previous research (and a nonprobabilistic model):

(49) The man who you saw saw me.

(50) The man who saw you saw me.

Figure 3.6 shows the reading time predictions of Hale's model; note that the object relative has a much higher maximum (and mean) surprisal than the subject relative.

In summary, probabilistic models of human parsing based on markov models and stochastic context-free grammars use the SCFG or HMM probability to predict which parse of an ambiguous sentence a human will prefer. These models also make some predictions about timecourse. Jurafsky (1996) and Crocker and Brants (2000) use the beam search paradigm to prune low-probability interpretations, hence predicting longer reading time when the next word is compatible only with a parse that has

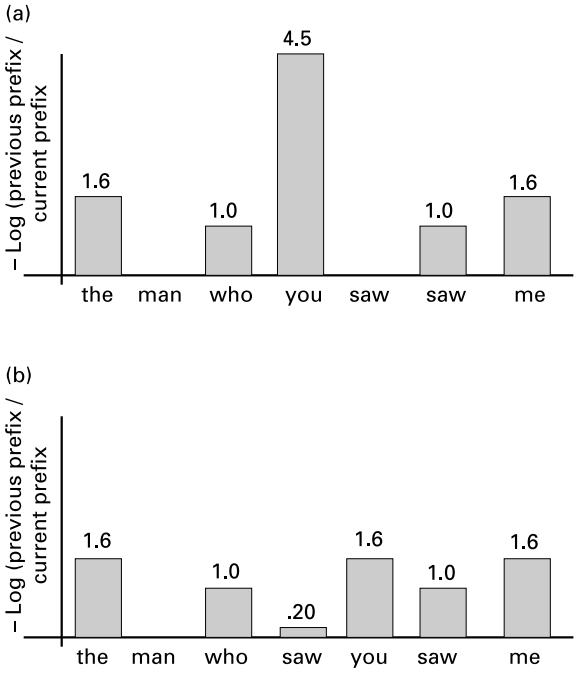

Figure 3.6
Hale's (2001) prediction of reading time based on surprise values computed from a simple stochastic context-free grammar for (*a*) object relatives and (*b*) subject relatives

already been pruned. Hale (2001) offers more fine-grained predictions about reading time, predicting an increase in reading time for surprising or unexpected words, as measured by parse probability.

3.3.5 Bayesian Belief Networks

As noted at the outset, the goal of probabilistic modeling in language processing is to solve the problem of choosing among possible alternatives in comprehension and production, given only incomplete and noisy evidence. The models summarized so far go a long way toward this goal. The competition and constraint satisfaction models both focus on the use of multiple probabilistic or frequency-based cues. The markov and SCFG models just described extend this use of probabilistic cues to show how some of these cues can be combined in a probabilistically correct manner; in particular, they focus on how independence assumptions,

like the assumptions of SCFGs, can be used to combine syntactic probabilities in a structured and motivated way.

In this section, I introduce a more general framework for combining probabilistic knowledge: the Bayesian belief network. Bayesian belief networks are data structures that represent probability distributions over a collection of random variables. A network consists of a directed acyclic graph, in which nodes represent random variables (unknown quantities) and the edges between nodes represent causal influences between the variables. The strengths of these influences are quantified by conditional probabilities; thus, for each variable node A that can take values $a_1 \ldots a_n$, with parents $B_1 \ldots B_n$, there is an attached conditional probability table $p(A = a_1|B_1 = b_x, \ldots, B_n = b_z)$, $p(A = a_2|B_1 = b_x, \ldots, B_n = b_z)$, and so on. The table expresses the probabilities with which the variable A can take on its different values, given the values of the parent variables. The structure of the network reflects conditional independence relations between variables, which allow the joint distribution to be decomposed into a product of conditional distributions. The Bayesian network thus allows us to break down the computation of the joint probability of all the evidence into many simpler computations.

Recall that the advantage of a Bayesian approach to language processing is that it gives a model of what probability to assign to a particular belief, and of how beliefs should be updated in the light of new evidence. Bayesian belief networks are thus on-line models; for example, if we are estimating the probabilities of multiple possible interpretations of an ambiguous utterance, the network will allow us to compute the posterior probability of each interpretation as each piece of evidence arrives. In addition, the use of a Bayesian belief network as a probabilistic estimator allows us to incorporate any kind of evidence: syntactic, semantic, discourse. This in turn allows us to capture the syntactic probabilities captured by graphical models like HMMs and SCFGs, while augmenting them with other probabilities, all in an on-line manner.

Jurafsky (1996) suggested that access and disambiguation of linguistic knowledge follow an evidential Bayesian model, though he gave only the briefest sketch of what the model should look like. Narayanan and Jurafsky (1998, 2002) followed up on this proposal by implementing a Bayesian model of syntactic parsing and disambiguation.

In this model, each interpretation of an ambiguous input is assigned a probability by combining multiple probabilistic sources of evidence, such as SCFG probabilities, syntactic and thematic subcategorization

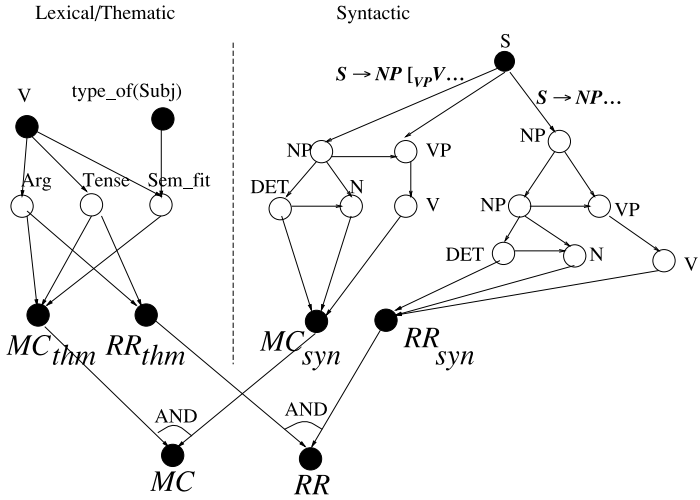

Figure 3.7
A belief network combining stochastic context-free grammar probabilities (*syn*) with subcategorization, thematic (*thm*), and other lexical probabilities to represent support for the main clause (MC) and reduced relative (RR) interpretations of a sample input. (From Narayanan and Jurafsky 1998.)

probabilities, and other contextual probabilities using a Bayesian belief network.

For example, after seeing the first few words of a main clause/reduced relative ambiguous sentence (*The horse raced*), the Bayesian model assigns probabilities to both the main clause (MC) and reduced relative (RR) interpretations using the belief network sketched in figure 3.7. This particular belief network combines multiple sources of probabilistic evidence, such as the subcategorization probability of the verb *raced*, the probability that *horse* is the semantic theme of a racing event, and the syntactic probability that a noun phrase will include a reduced relative clause, computed using SCFG probabilities.

This network is actually composed of two subnetworks, one computing the SCFG probabilities and one computing the lexical and thematic probabilities. The SCFG probabilities can be directly computed by the first subnetwork; the conditional independence assumptions in a stochastic context-free parse of a sentence can be translated into the conditional independence statements entailed by a Bayesian network. Figure 3.8 illustrates the belief network representations that correspond to the SCFG

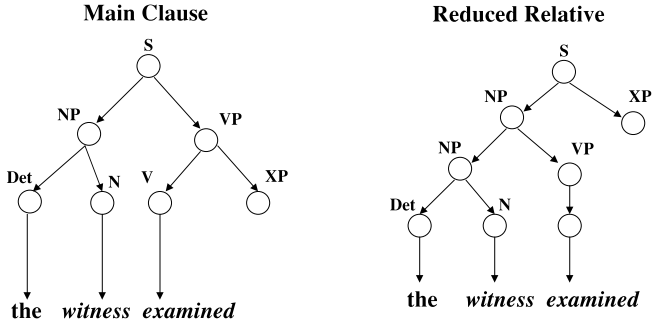

Figure 3.8
Pieces of belief networks corresponding to two stochastic context-free grammar parses for the prefix *The witness examined* . . .

parses for the main clause and reduced relative interpretations of an ambiguous prefix like *The witness examined*.

Figure 3.9 gives the structure of the Bayesian network that computes lexical and thematic support for the two interpretations. The model requires conditional probability distributions specifying each verb's preference for different argument structures, as well as its preference for different tenses. Narayanan and Jurafsky (2002) also compute probabilities from the semantic fit between head nouns (like *crook* or *cop*) and semantic roles (agent or patient) for a given predicate (like *arrested*) by normalizing the preferences given by McRae, Spivey-Knowlton, and Tanenhaus (1998). Thus, the probabilities include

$$P(\text{agent} \mid \text{subject} = \text{crook}, \text{verb} = \text{arrested}),$$
$$P(\text{patient} \mid \text{subject} = \text{crook}, \text{verb} = \text{arrested}),$$
$$P(\text{transitive} \mid \text{verb} = \text{arrested}),$$
$$P(\text{preterite} \mid \text{verb} = \text{arrested}),$$

and so on. As shown in figure 3.9, the MC and RR interpretations require the conjunction of specific values corresponding to tense, semantic fit, and argument structure features. Note that only the RR interpretation requires the transitive argument structure.

In some cases, as with the SCFG, we have relatively complete models of the independence assumptions between probabilities. In other cases, as with thematic and syntactic probabilities, we do not yet have a good idea what the exact causal relationship is between probabilities. The simplest

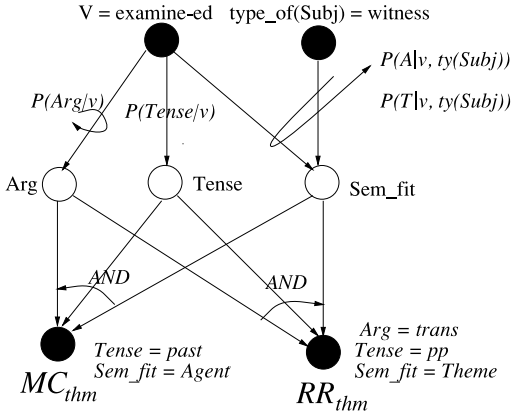

Figure 3.9
The belief network that represents lexical and thematic support for the two interpretations shown in figure 3.8 ($A = agent$, $T = theme$)

thing to do in such cases is to assume the probabilities are independent and to multiply them. Figure 3.7 shows that Narayanan and Jurafsky (1998) make a somewhat weaker assumption by using the noisy-and model (Pearl 1988) in computing the conjunctive impact of the lexical/thematic and syntactic support to compute the probabilities for the MC and RR interpretations. A noisy-and model assumes that whatever *inhibits* a specific source (syntactic) from indicating support for an interpretation is independent of mechanisms that inhibit other sources (lexical) from indicating support for the same interpretation. This assumption, called the assumption of exception independence, is used widely with respect to both disjunctive (noisy-or) and conjunctive sources. In the case of the RR and MC interpretations, as each piece of new evidence is introduced by reading new words, the posterior support for the different interpretations is computed using the following equation:

$$\begin{aligned} P(MC) &= 1 - P(\neg MC) = 1 - P(\neg MC|Syn, Lex, Thm) \\ &= 1 - (P(\neg MC|Syn) \times P(\neg MC|Lex, Thm) \\ P(RR) &= 1 - P(\neg RR) = 1 - P(\neg RR|Syn, Lex, Thm) \\ &= 1 - (P(\neg RR|Syn) \times P(\neg RR|Lex, Thm). \end{aligned} \quad (51)$$

Let's walk through the model word by word as it assigns probabilities to the different parses of the initial prefix of three sentences:

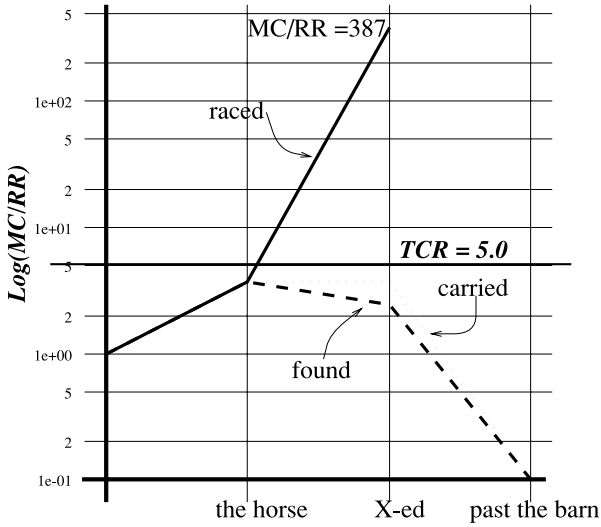

Figure 3.10
The main clause/reduced relative (MC/RR) posterior probability ratio for *raced* falls above the threshold and the reduced relative interpretation is pruned. For *found* and *carried*, both interpretations are active in the disambiguating region. (From Narayanan and Jurafsky 1998.)

(52) The horse raced past ...

(53) The horse carried past ...

(54) The horse found in ...

Previous research has found that (52) causes a severe garden path effect, while (53) and (54) do not (Pritchett 1988; Gibson 1991). Narayanan and Jurafsky's (1998) approach models this garden path effect via the beam search assumption of Jurafsky (1996); interpretations whose probability falls outside the beam width of the best interpretation are pruned. Figure 3.10 shows the relevant posterior probabilities for the example *The horse raced past the barn fell* and the replacement of *raced* by *carried* or *found* at different stages of the input, expressed in terms of the probability of the MC interpretation to the RR interpretation, or the MC/RR ratio.

At the first point in the graph, the network expresses the probability ratio after seeing the phrase *the horse*. The network is thus computing the following probabilities:

P(MC, S → NP . . . , NP → Det N, Det → the, N → horse | the, horse),

P(RR, S → NP . . . , NP → NP . . . , NP → Det N, Det → the,
N → horse | the, horse).

Next, the word *raced* appears, and the network computes the new posterior probability ratio given this new information:

P(MC, S → NP VP, NP → Det N, Det → the, N → horse, VP → V . . . ,
V → raced, Vform, Agent | Vform = preterite, subject = "horse",
verb = race);

P(RR, S → NP VP, NP → NP VP, NP → Det N, Det → the,
N → horse, VP → V . . . , V → raced, Vform, Agent | Vform = participle,
subject = "horse", verb = race).

As shown in figure 3.10, Narayanan and Jurafsky's (1998) model predicts that the MC/RR ratio exceeds the threshold immediately after the verb *raced* is accessed (MC/RR $\approx 387 \gg 5$), leading to the pruning of the RR interpretation. In the other cases, while the MC/RR ratio is temporarily rising, it never overshoots the threshold, allowing both the MC and RR interpretations to be active throughout the ambiguous region.

Narayanan and Jurafsky (2002) tested their (1998) data further, by modeling both sentence completion probabilities and reading time data on 24 sentences from McRae, Spivey-Knowlton, and Tanenhaus 1998. They also included in the model new probabilities taken from McRae, Spivey-Knowlton, and Tanenhaus's study that allow conditioning on the identity of the preposition. Finally, they extended the model's reading time predictions by predicting an increase in reading time whenever an input word causes the best interpretation to drop in probability enough to switch in rank with another interpretation.

The first experiment modeled by Narayanan and Jurafsky (2002) was the sentence completion experiment conducted by McRae, Spivey-Knowlton, and Tanenhaus (1998), summarized above. Narayanan and Jurafsky showed that the same factors integrated by McRae, Spivey-Knowlton, and Tanenhaus using the competition-integration model can instead be integrated by the Bayesian network shown in figures 3.8 and 3.9.

Figure 3.11 shows the human fragment completion preferences and the probabilities the model assigned to the RR and MC completions. The Bayesian model's results correspond closely to the human judgments about whether a specific ambiguous verb was used in the MC or RR

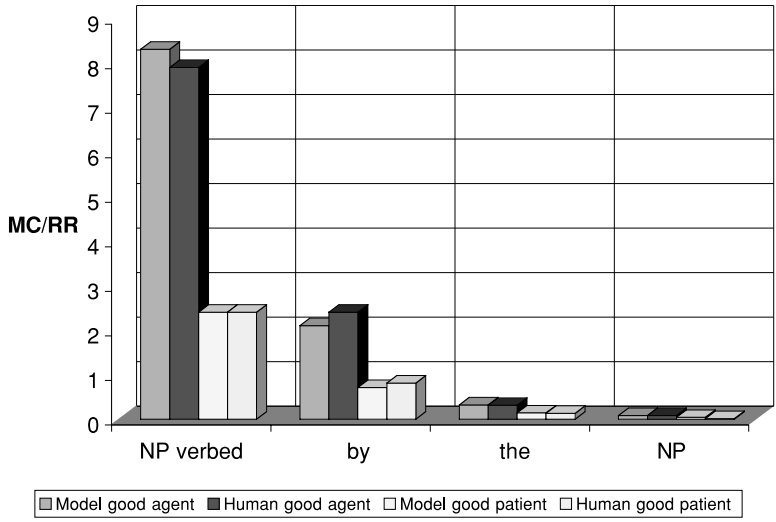

Figure 3.11
The main clause/reduced relative (MC/RR) posterior probability ratio for sentence completion after each word, from the Bayesian model (Narayanan and Jurafsky, in press) and human completion data (McRae, Spivey-Knowlton, and Tanenhaus 1998)

construction. As in McRae, Spivey-Knowlton, and Tanenhaus 1998, the data show that thematic fit clearly influenced the sentence completion task. The probabilistic account further captured the fact that at the *by* phrase, the posterior probability of producing an RR interpretation increased sharply; thematic fit and other factors influenced both the sharpness and the magnitude of the increase.

Narayanan and Jurafsky (2002) also modeled aspects of on-line reading experiments from McRae, Spivey-Knowlton, and Tanenhaus 1998 discussed above. Recall that the latter authors showed that controlled human reading time for good agents for *arrested* (e.g., *cop*) gets longer after reading the *by* phrase (requiring *cop* to be a patient), while controlled reading time for good patients (e.g., *crook*) gets shorter. Narayanan and Jurafsky's (1998) model predicts this larger effect from the fact that the most probable interpretation for the good agent case *flips* from the MC to the RR interpretation in this region. No such flip occurs for the good patient case.

Figure 3.12(a) shows that the good patient results already have an MC/RR ratio of less than one (the RR interpretation is superior), while a flip

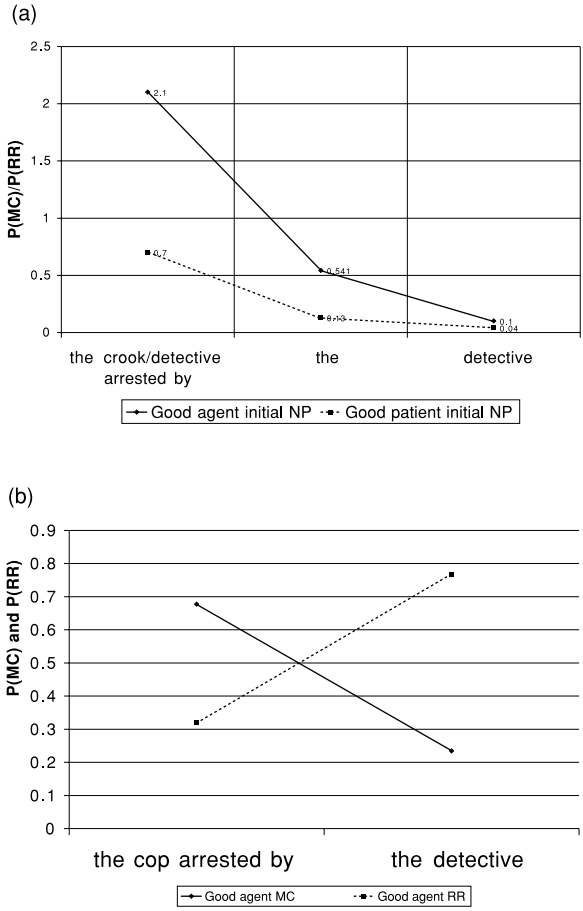

Figure 3.12
(*a*) Main clause/reduced relative ratio for the ambiguous region showing a flip for the good agent case at the word *the*, as the ratio goes below 1, but no such flip for the good patient case. (*b*) $P(\text{MC})$ and $P(\text{RR})$ for the good agent cases alone.

occurs for the good agent sentences (from the initial state where MC/RR > 1 to the final state where MC/RR < 1). Whereas figure 3.12(a) shows MC/RR ratios for different initial NPs, figure 3.12(b) focuses just on the good agent case and breaks the MC and RR probabilities into two separate lines, showing the crossing point where the flip occurs.

3.3.6 Probabilistic Modeling of Production

We have now looked at the architectures of many probabilistic models of comprehension. In production, by contrast, there seem to be no worked-out probabilistic models. Perhaps the main cause of this, as Harald Baayen (personal communication) points out, is that obtaining probabilities for production studies is difficult, because it is so difficult to control the circumstances that will prompt a subject to coin a particular sentence. In any case, modern models of lexical production such as those developed by Levelt, Roelofs, and Meyer (1999), Dell (1986), and Dell et al. (1997)—indeed most models, back to Morton's (1969) seminal *logogen* model—are based on some idea of activation that is tied in some way to frequency. In a logogen-style model, a high-frequency word has a lower activation threshold and hence is quicker to access. In other models, frequency instead plays its role via the weights on links that pass activation into a word node. In any case, Bates and Devescovi (1989) proposed that this sort of frequency-based activation model also be used to model syntactic frequency effects. In their model, given a semantic input, the production system allows various possible syntactic constructions and lexical realizations to compete for access. Frequency and function both play a role in ranking these competing realizations. Evidence for this role of syntactic frequency comes from the study by Bates and Devescovi (1989) discussed in section 3.2.6, which showed that in a controlled production study, relative clauses occurred far more often in Italian production than English production. They suggest that the frequency of relative clauses in Italian may play a direct role in their being chosen in production.

In addition, Stallings, MacDonald, and O'Seaghdha (1998) and Roland and Jurafsky (2001) suggest that various conditional probabilities relating to a verb's subcategorization frame play a role in production. The experiment by Stallings, MacDonald, and O'Seaghdha summarized in section 3.2.4 suggests that each verb is stored with a "shifting disposition"—a frequency-based preference for whether it expects to appear contiguous with its arguments or not. Stallings, MacDonald, and O'Seaghdha suggest that this preference plays an on-line role in choosing

an interpretation in production. Roland and Jurafsky suggest that a verb subcategorization probability is stored with the verb lemma and used in production to select among alternative subcategorizations.

3.3.7 Conclusion

I have sketched, at a surface level, a number of probabilistic models of comprehension. Most models focus on ambiguity, showing that human preference for one interpretation of an ambiguous input can be predicted by the probability of that interpretation. The rational models extend this idea to investigate the role of cost and utility in disambiguation preferences. Finally, the most recent work has begun to explore a more fine-grained relationship between probability and processing time.

3.4 Potential Challenges to and Confusions about Probabilistic Models

In this section, I discuss some commonly encountered challenges to probabilistic models and introduce some frequently asked questions.

3.4.1 Surely You Don't Believe That People Have Little Symbolic Bayesian Equations in Their Heads?

No, probabilistic modeling of human language processing does not imply that little Bayesian equations are somehow symbolically dancing around in the head of a speaker. This misguided conception gives rise to a common objection to probabilistic models of language processing: that it seems hard to believe that people are "doing complex math in their heads."

Rather, many probabilistic modelers assume that probability theory is a good model of language processing at what Marr (1982) called the "computational level": it characterizes the input-output properties of the computations that the mind must somehow be doing. How this model is realized at lower levels (of "implementation" or "algorithm") is an interesting question that unfortunately cannot be addressed here (though see some discussion in Baayen, this volume). The most common assumption, however, is that probability is realized either as an activation level of some mental structure or as a distributed pattern of activation. Stored frequencies or probabilities can thus be encoded either as resting activation levels or as weights on connections. It is well known that the link between neural network or connectionist models and probabilistic ones is close (see, e.g., McClelland 1998). Other realizations of probabilistic

models are possible, however, such as the exemplar models of phonology developed by Pierrehumbert (2001b) and others.

It is important to mention an alternative possible relation between probabilistic models and neural networks or other activation-based models, suggested by Ariel Cohen (personal communication)—namely, that probabilistic models are simply wrong, and that neural network or connectionist models are a better and more explanatory model of human language processing. Unfortunately, very little research has focused on discriminating between probabilistic models and connectionist models. Deciding whether probabilistic models are merely a higher-level description of connectionist models, or whether the two are mutually exclusive alternatives, remains a key problem for future research.

3.4.2 Are Probabilistic Models Always Nonmodular?

Frequency-based, constraint-based, and probabilistic models have often been opposed to models that exhibit Fodorian modularity (Fodor 1983), or models based on rich linguistic structure. While any individual model may make any particular confluence of claims, there is no *necessary* link between probability and antimodularity. The probabilistic models discussed in this chapter are generally models of the probability *of* something—generally the probability of a certain linguistic structure, as computed by humans in the course of linguistic processing. This fact should make it clear that these probabilistic models are not meant to argue for "using numbers instead of linguistic structure" or "random number generators in people's heads" (to cite two more standard but misguided objections to probabilistic models). Some of the probabilistic models sketched in this chapter are modular (e.g., the one proposed in Crocker and Brants 2000); others are nonmodular. Some involve "emergent" structure; others involve explicit structure. Furthermore, the use of probability says nothing about whether many and varied probabilistic constraints are used immediately, as most probabilistic researchers believe, or after a delay of a few milliseconds, as some psycholinguists have argued in offering the garden path and construal models.

3.4.3 But Corpus Frequencies Don't Match Norming Study Frequencies

The probabilities in the models described here are generally estimated from corpus frequencies. A number of researchers have noticed that these corpus frequencies do not always match the frequencies derived from various psychological experiments. This mismatch might suggest that

frequency-based models are not psychologically plausible. Recall, for example, the work of Gibson, Schütze, and Salomon (1996) on English and Mitchell and Brysbaert (1998) on Dutch, which suggested that attachment preferences differ between corpora and production experiments. Desmet, Brysbaert, and Baecke (in press) showed that, at least in Dutch, this difference disappears when the animacy of the NPs is controlled.

Other such mismatches have been reported, however. For example, Merlo (1994) compared verb subcategorization frequencies computed from corpora with frequencies computed from psychological norming studies. In a kind of psychological norming study called a "sentence production" study, subjects are asked to write a sentence using a particular verb. Transitivity biases are then computed from a collection of such sentences. Merlo found that transitivity preferences in a corpus of *Wall Street Journal* and DARPA Air Travel Information System sentences differed from transitivity preferences in norming studies such as Connine et al.'s (1984).

Roland and Jurafsky (2001) followed up on Merlo's research by looking at the causes of subcategorization differences between corpora such as the Brown corpus and subcategorization norming studies such as Connine et al.'s (1984). Their analysis suggests that most of the differences between these verb subcategorization frequencies came from two factors. The first factor is word sense: different corpora tend to use different senses, and different senses tend to have different subcategorization biases. The second factor is discourse and genre effects: for example, the single-sentence production tasks were much less likely to display passives, zero anaphora, and other discourse-related phenomena than natural corpus sentences. Roland et al. (2000) extended this study, examining the subcategorization probabilities for 69 verbs. They found that after controlling for verb sense, the binned subcategorization probabilities (high, medium, and low transitive bias) for each verb were relatively stable across corpora.

What are the implications for probabilistic models? Roland and Jurafsky (2001) and Roland (2001) proposed that the locus of verb subcategorization probabilities is the semantic lemma rather than the lexeme, and suggested that the frequency of a particular verb subcategorization in a corpus is a product of multiple factors. In particular, in lexical production, lexical subcategorization probabilities, which are stored at the semantic lemma level, might be combined with other probabilistic influ-

ences from discourse and genre to produce the subcategorization patterns observed in corpora.

Thus, the mismatch between corpus frequencies and psychological norming studies is to be expected. These are essentially two different kinds of production studies, with different constraints on the production process. A probabilistic model of production that is correctly conditioned on sense, genre, and other factors would correctly model the different observed frequencies for these two kinds of corpora.

3.4.4 Maybe Frequency Is Just an Epiphenomenon

Another common objection to probabilistic and other frequency-based models is that frequency is only an epiphenomenon of other structural factors. One such claim relates to the processing of unaccusative and unergative verbs in English. Unaccusative verbs (*bloom*, *melt*, *blush*, etc.) and unergative verbs (*race*, *slide*, *sail*, etc.) are both typically intransitive, but have been modeled as differing in underlying lexical-syntactic form:

Unergatives	NP [$_{VP}$ V] (external argument, no internal argument)
Unaccusatives	____ [$_{VP}$ V NP/CP] (internal argument, no external argument)

Both unaccusative and unergative verbs generally alternate with a causative transitive form (e.g., unaccusative *melt* can be both intransitive and transitive). Kegl (1995) has claimed that unaccusatives are particularly hard for agrammatic aphasics to process. Her argument is based on a study of an agrammatic aphasic subject, whose productions showed a significant absence of unaccusatives when compared with those of a matched control. Kegl's explanation is that unaccusatives are like passives in involving traces, which are claimed to be generally difficult for agrammatic aphasics to process (Grodzinsky 2000).

An alternative explanation might be based on frequency: as Gahl et al. (in press) have suggested, the comprehension difficulty of a verb might vary with its frequency-based subcategorization bias, and unaccusative verbs could occur more frequently in their intransitive than in their transitivized form. Gahl et al. tested this hypothesis with eight aphasic subjects using a plausibility-in-comprehension task, with both transitive and intransitive sentences. A given sentence thus either matched or didn't match the transitivity bias of the verb. Gahl et al. predicted that a sentence should be easier to understand if its structure matches the tran-

sitivity bias of its verb, and they predicted that there was no reason to expect unaccusatives to act like passives. They found that unaccusatives as a whole were much easier for their aphasic subjects to understand than passives, that unaccusatives as a whole were not harder than unergatives, and that in general sentences were easier when their syntactic structures matched the subcategorization frequency bias of the verb. Thus, processing of unaccusatives was influenced by frequency bias, rather than by structural problems with traces.

Another claim that structure rather than frequency causes processing difficulty comes from Stevenson and Merlo (1997), who noticed that the causativized form of unergative verbs (see (52a)) is more difficult to process than the causativized form of unaccusative verbs (see (52b)).

(52) a. *Causativized unergatives*
The students *advanced* to the next grade had to study very hard.
The clipper *sailed* to Portugal carried a crew of eight.
The ship *glided* past the harbor guards was laden with treasure.
b. *Causativized unaccusatives*
The witch *melted* in *The Wizard of Oz* was played by a famous actress.
The oil *poured* across the road made driving treacherous.

Stevenson and Merlo were extending a proposal of Hale and Keyser (1993), in which verbs project their phrasal syntax in the lexicon. Stevenson and Merlo proposed that causativized (transitive) forms of unergatives are more complex than causativized (transitive) forms of unaccusatives, in terms of number of nodes and number of binding relations. This complexity, together with limitations on creating and binding empty nodes, caused Stevenson's (1994) parser to be unable to activate the structure needed to parse transitivized unergatives, hence explaining the garden path effect.

But an alternative explanation for the garden path effect relies on the subcategorization frequency biases discussed earlier. As Stevenson and Merlo (1997) and Gahl (1999) show, unergative verbs like *race* have a huge intransitive bias, and unaccusatives have a slight causative/transitive bias, if anything (see table 3.6, and see Gahl 1999 for further details of the comparison).

A frequency explanation for the difficulty of these garden path sentences also has the advantage of explaining gradient effects. Filip et al.

Table 3.6
Transitivity counts for unergative versus unaccusative verbs (From Gahl 1999.)

	Transitive		Intransitive	
Unergative	2,869	13%	19,194	**87%**
Unaccusative	17,352	**54%**	14,817	46%

(2002), for example, have shown that some unergatives are easier to understand in reduced relative clauses than others, which would be difficult to explain with a purely structural model.

The fact that frequency might be the psychological actor rather than structural factors in this instance does not mean that structural, semantic, or functional factors might not often be the causal force that is grammaticalized via frequency. Thus, it is crucial to continue to investigate semantic or functional factors like those proposed by Stevenson and Merlo (1997).

3.5 Conclusion

What is the state of knowledge about probabilistic modeling in 2002? We know that the frequency of many kinds of linguistic structure plays a role in processing. The strongest evidence for this role, however, exists only for frequency related in some way to lexical items or to the relationship between lexical items and syntactic structure. The role of probabilities in nonlexical syntactic structure, while assumed in many probabilistic models, rests on very little psychological evidence. This is perhaps unsurprising, since the psychological evidence for constituency itself is so weak. Nonetheless, understanding the role of frequency of larger structures is an important unsolved problem.

As for models, it is clear that probabilistic models of linguistic processing are still in their infancy. Most models include only a very small number of probabilistic factors and make wildly unjustified assumptions about conditional independence. Furthermore, there is a dearth of work exploring the crucial relationship between the neural network models, which focus on emergence, distributional evidence, and the details of input features, and Bayesian models, which focus on the mathematics of evidence combination and independence assumptions. Nonetheless, some conclusions are already possible. Probabilistic models do a good job of selecting the preferred interpretation of ambiguous input and are starting

to make headway in predicting the timecourse of this disambiguation process.

Many unsolved problems remain. How exactly should prior probabilities be estimated from corpora? What exactly is the relationship between probability and reading or production time? We know that this relationship is logarithmic, but little about how or why.

The constraints of space and time have made this survey of probabilistic work in psycholinguistics unfortunately brief. I have given short shrift to the role of frequency in recall, to the role of phonological and orthographic neighborhood frequencies in processing, and, most distressing, to the vast connectionist literature that is so closely related to probabilistic modeling. Alas, those areas will have to await another survey.

Notes

Many thanks to the editors, Susanne Gahl, Alan Bell, and Ariel Cohen, and special thanks to Harald Baayen for help and detailed comments above and beyond the call of duty. An early version of this chapter was given as a talk at the AMLAP 2001 conference in Saarbrücken; particular thanks to Matt Crocker, Don Mitchell, and Brian McElree for helpful comments. Of course, all remaining errors are my own.

1. Controlled reading times are computed by subtracting reading times for reduced relative clauses from those for unreduced relative clauses.

This page intentionally left blank

Chapter 4

Probabilistic Sociolinguistics: Beyond Variable Rules

Norma Mendoza-Denton, Jennifer Hay, and Stefanie Jannedy

4.1 Overview

In this chapter we outline issues facing quantitative approaches in contemporary variationist sociolinguistic theory, surveying trends that led scholars (1) to reject intuitive and categorical descriptions of language data and (2) to use frequency-based and probabilistic approaches to the modeling of language variation. We discuss the importance of linguistic and social contexts in the description of variable linguistic behavior by analyzing our data on the monophthongization of the diphthong /ay/ in the speech of the popular African-American talk show host Oprah Winfrey. We compare VARBRUL (variable rule-based logit analysis) results with CART (classification and regression trees) results to highlight the strengths and weaknesses of different tools in the modeling of probabilistic phenomena. Implications for the theory of sociolinguistic variation and for models of cognition are emphasized throughout. We advocate a usage-based account for both linguistic processes and social identity construction. Such an account allows for the continuous and incremental updating of mental representations on the basis of new input, and it synergistically captures advances in probabilistic linguistics and in social identity construction theory. Social identities are transmitted simultaneously with linguistic structures, and as such they represent dynamic processes, continuously negotiated in interaction.

4.2 Background and History

4.2.1 Introduction

The study of sociolinguistic variation has faced different problems from other fields of linguistics. While other branches of quantitative linguistics

have competed with schools of intuitive and categorical thinking (Bod, this volume), sociolinguists have always started from empirical premises. The very first statistically sophisticated studies that were conducted in a modern sociolinguistic framework laid the foundation for debates on statistical modeling within this field. Past debates within sociolinguistics have included the search for a unified statistical model and tools (Bickerton 1971; Sankoff and Rousseau 1974); the interpretation of correlational statistics linking social structure to linguistic forms, especially in the field of language and gender (Eckert 1989; Labov 1990; Cameron 1990); and the positing of alternative models for the diffusion of change through a population, such as the implicational scale versus quantitative model debate (Bickerton 1973; Romaine 1985; Rousseau 1989; see summary in Rickford 2001). Several of these debates have accorded privileged status to questions of how to model the mathematics of sociolinguistics, while paying short shrift to cognitive issues of the mental representation of linguistic categories and of social processes. Recent work by Mendoza-Denton (2001) and Eckert (1999) has pointed out that advances within social theory and the evolution of understanding of sociolinguistic processes challenge researchers to move beyond viewing social categories as static, relegating them to simple decisions made by the analyst prior to data analysis. Primary questions now surfacing are: How do social categories emerge from the distribution of data? How do abstractions such as ethnicity and gender emerge from the many different ways that speakers have of fashioning themselves as classed, gendered, or ethnic social agents? Although some of the current methods (such as VARBRUL and CART) constrain researchers in selecting discrete variables within sociodemographic categories (coding tokens for age, ethnicity), we propose utilizing a variety of techniques (including discourse and conversation analysis) to more closely examine specific instances of variables and the contexts of their use to determine how social meaning is constructed.

Exemplar theory, a frequency-based model emerging in areas such as phonology and morphology (Pierrehumbert, this volume), can lead the way to unification with social-theoretic understandings of the role of innovative social actors in communities of practice. In exemplar theory, categories are not preexisting, but are established as dynamic (continuously and incrementally updated) generalizations of linguistic data over increasingly abstract domains. The robustness of the categories depends on frequency of the input that can be classified under that category, and on the recency of the stimulus.

There is a groundswell of evidence that much social information is carried in moment-to-moment performances by key individuals—icons—in local communities (Eckert 1999; Labov 2001; Mendoza-Denton 2001; Schilling-Estes 2001). Performances by these social brokers in the linguistic marketplace are subject to the same cognitive constraints of robustness and frequency that underlie other areas of symbolic manipulation.

After reviewing some of the early sociolinguistic literature on variation and on the variable rules framework, we present an extended example analyzing a socially iconic speaker—Oprah Winfrey—with two statistical modeling techniques, supplemented with discourse analysis, showing how her use of specific variants contributes to the construction of her linguistic style.

4.2.2 Against Intuition

Sociolinguistics explores the social correlations of patterns of human linguistic behavior at all levels of grammar, ranging from phonology and syntax to semantics and discourse. The quantification of performance data to explore and explain speakers' linguistic competence in social situations has been a staple of the sociolinguistic paradigm. Unlike the methods used in some other areas of linguistics, those deployed by sociolinguists are empirical in nature and require the modeling of quantitative patterns to draw conclusions about speaker competence. It is not assumed that linguistic innovation, nuances in speech patterns, or variants of lexical choice are in free variation. Rather, they are manifestations of the subtle patterning and interaction of linguistic and social competence.

A speaker has choices to make when selecting which words to use in crafting a sentence, whether to release a word-final stop, or whether to raise a high vowel to display more extreme formant values. These choices carry social meaning at the moment of utterance, and the gradual cumulative steps of innovators may lead to category shifts with the power to rearrange entire linguistic systems. Through the analysis of historical records we gain insight into the succession of linguistic changes, such as those precipitated by the English Great Vowel Shift. Historical evidence and contemporary recordings can be used to show the gradualness of these changes, the lexical diffusion of their carrier items through the population, and their continuing consequences in current structural reorganizations, such as the Northern Cities Chain Shift in the United States (Eckert 1989; Labov 2001).

Sociolinguistics is concerned with capturing not only patterns of change, but also variation across speakers of different speech communities, among speakers in a single speech community, and in the speech of individuals. Variability follows the twin constraints of (1) being conditioned by language-internal factors and (2) participating in processes of social semiosis—a dual meaning-making system par excellence. Because there is little room in generative linguistic frameworks to explore and explain either noncategorical changes or stable variation, much work in that vein has been devoted to describing the endpoints of changes, variability being dismissed as randomness or noise. Categorical descriptions of language data ignore the triggers and mechanisms of variability, their social motivation, and the productivity of such linguistic patterns.

As far back as 1937, Bronislaw Malinowski outlined a view of the essential dilemma facing linguistics:

> ... whether the science of language will become primarily an empirical study, carried out on living human beings within the context of their practical activities, or whether it will remain largely confined to deductive arguments ... (1937, 63)

This chapter will argue that current quantitative models of language behavior may still benefit from further investigation precisely of the form that Malinowski advocated: carried out on living individuals in the course of practical activity, shedding light on both linguistic form and questions of social structure.

Hymes exhorted his linguistic contemporaries to take up research in a nascent field called sociolinguistics, the goal of which was to "identify rules, patterns, purposes, and consequences of language use, and to account for their interrelations" (1974, 71). The definitional core of this field was and remains a theoretical concern for the interrelationship and the codependence between components of linguistic structure and of social structure. Why is this inherently a probabilistic problem? Sociolinguists commonly understand the *linguistic variable* as "a construct that unites a class of fluctuating variants within a language set" (Wolfram 1991, 23), reflecting a decision point at which a speaker chooses between alternative ways of saying the same thing.

The central probabilistic sociolinguistic questions then become: What factors affect a speaker's decision to use one variant over another? How can we best model the simultaneous influence of these linguistic and social factors at that particular decision point? How does the use of a particular linguistic variant reflect social membership? And what can the distribu-

tion of alternative forms in the social landscape reveal about the internal synchronic and diachronic workings of linguistic structure?

We take the multiple determination of variables as a given: it is not only the internal organization of linguistic structure (i.e., phonological context) that shapes variation, but also social settings and characteristics of speakers, all operating in concert and reflected in language (cf. Bayley's (2001) principle of multiple causation).

Labov (1966, 1972) showed that through the frequencies of the various phonetic manifestations of underlying phonological /r/, New Yorkers displayed finely tuned linguistic performance reflecting social classes, ethnic groups, and even such subjective factors as the level of formality in the speech situation. Rhoticity, the presence or absence of a pronounced syllable-coda /r/, varied in predictable and replicably measurable ways. However disparate their informal production, New Yorkers demonstrated their orientation to the current rhotic standard by exhibiting variation wherein formal speech was always more rhotic than informal speech, across all social classes and across all "styles" (word list, reading passage, formal interview, unstructured interview). Figure 4.1 illustrates class stratification in New York City as reflected in a linguistic variable. The vertical axis represents a phonological index for (r), where 100 would reflect a completely *r*-ful dialect and 0 would reflect a completely *r*-less one. The interview contexts that appear on the horizontal axis are designed to elicit increasingly careful, standardized speech. This figure shows that as the formality of the context increases, from casual speech through minimal pairs, so does the production of rhotic speech across all social groups. No group is categorically *r*-ful or *r*-less, and all groups exhibit a finely grained pattern of linguistic behavior that indicates consciousness of the *r*-ful form as the prestigeful target. On the basis of their collective orientation toward the same prestigeful targets across different variables—Labov studied (r), (th), and (-ing)—these randomly sampled New Yorkers could be classified as a single speech community. In the production arena, social differences are shown in patterns of variation. In the perceptual arena, social inferences are drawn from the architecture of variation.

Labov's (1966, 1972) study precipitated a scientific paradigm shift in the study of language and society. Since then, much sociolinguistic work has been carried out using the methodology of the *sociolinguistic interview*, a structured oral interview protocol that was originally designed to be administered to a large, randomly sampled, stratified urban population

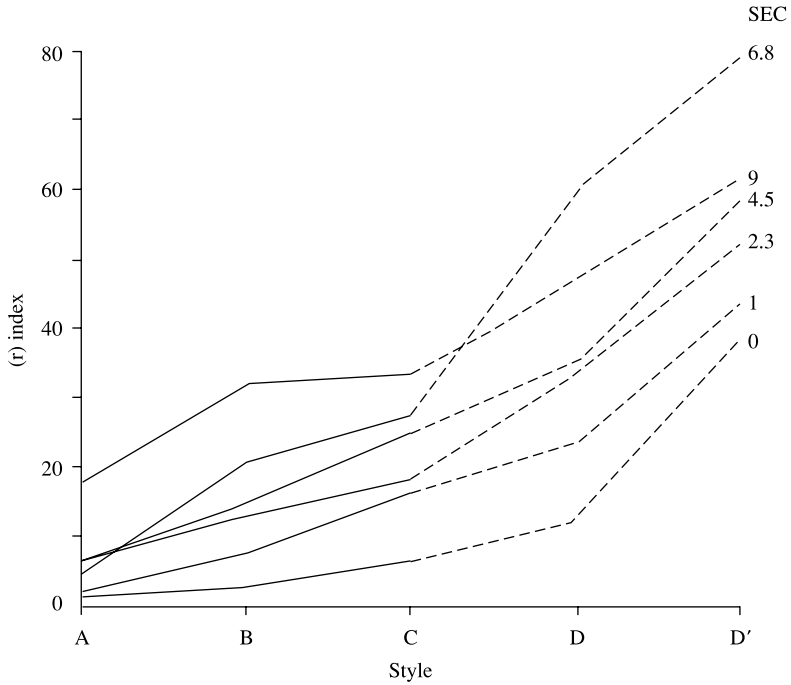

Figure 4.1
Class stratification of a linguistic variable in the process of change: (r) in *guard*, *car*, *beer*, *beard*, *board*, and so on. SEC (socioeconomic class) scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists; D′, minimal pairs. (From Labov 1972, 114.)

of the sort studied by sociologists. Indeed, Labov's innovative interview method was first undertaken as part of a sociological survey of New York City. Soon thereafter, in the 1960s and 1970s, large-scale, quantitative studies began in other urban areas. Such studies aimed to model different strata of speech communities by including large numbers of speakers, varying with respect to age, ethnicity, socioeconomic status, and gender. Modern sociolinguistics is firmly grounded in the belief that language change is propelled by social variation, where innovative speakers push the envelope of preexisting changes, simultaneously abstracting from and constrained by structural linguistic factors. Linguistic facts that appear synchronically categorical—the lack of grammatical gender agreement in

English, for instance—appear from a diachronic perspective as the endpoint of a change that has been carried through to completion. Much of the motivation for presenting data on age-graded, style-based, or gendered stratification is to support claims of changes in progress. Indeed, the old notion of "free variation" has been entirely replaced in sociolinguistics by the notion of a change in progress, where the variable in question is assumed to be part of a system in flux, and the task of the sociolinguist is to identify, from a naturalistic sample, which is the conservative usage, which is (are) the innovative usage(s), who the innovators are, and what structural constraints they face.

To date, hundreds of urban studies in the United States and around the world have applied some version of the sociolinguistic interview methodology (though it is not without its problems—see Wolfson 1976; Briggs 1986), eliciting informal speech by asking interviewees what their childhood games were like, whether they have ever come close to death, and what kinds of changes they have experienced in their lifetimes (for a detailed explanation of the design of a sociolinguistic interview, see Feagin 2001). This method has proved remarkably productive and has served in creating finely stratified models of the speech of urban populations. This particular area of inquiry has come to be called *urban dialectology*, and here we cite but a few recent examples: Silva-Corvalán 1989 for Santiago, Chile; Thibault and Daveluy 1989, Thibault and Sankoff 1993 for Montreal, Canada; Tagliamonte 1999 for York, U.K.; Kontra and Váradi 1997 for Budapest, Hungary; Lennig 1978 for Paris, France; Trudgill 1974, 1988 for Norwich, U.K.; Horvath 1985 for Sydney, Australia; Rickford 1986 for Guyana; Haeri 1998 for Cairo, Egypt; and Labov, Ash, and Boberg, in press, for 145 cities in the United States alone, where sociolinguistic interview methodology and minimal pair elicitation have been combined to produce *The Atlas of North American English*.

4.2.3 Beginning with Frequency

Some of the first studies of language variation were done on sociolinguistic interview corpora, using frequency-based information to locate patterning in the production of linguistic variables. For instance, Wolfram (1974, 202) investigated linguistic contact among African-American and Puerto Rican speakers in New York City by examining their rates of monophthongization of /ay/. Monophthongization of /ay/ is understood

Table 4.1
Percentages of monophthongized /ay/ tokens in the speech of African-American speakers (AA), Puerto Rican speakers with extensive African-American contacts (PR/AA), and Puerto Rican speakers with limited African-American contacts (PR). (Adapted from Wolfram 1974.)

	AA	PR/AA	PR
No./Total	190/247	104/148	261/657
% monophthongized	76.9	70.3	39.7

to be an African-American English feature typically not present in Euro-American dialects in the northern United States such as that of New York City. Wolfram hypothesized that linguistic influence from African-Americans was the source of greater frequencies of monophthongization among Puerto Rican speakers with extensive contacts in the African-American community, as compared to those with limited contacts (see table 4.1). Although this brief example does not fully portray the complexity of Wolfram's findings, we will borrow it to help illustrate two extended points, one sociolinguistic-methodological and one mathematical.

First, in appealing to social explanations for the patterning of linguistic data, and to ensure their validity and replicability, students of variation begin by thoroughly investigating the social categories extant in a given community. Often this takes the form of prolonged ethnographic, participant observation fieldwork within the community in question. This particular feature of investigative inquiry minimizes the observer's paradox and creates a number of close connections between sociolinguistics and qualitative social sciences such as anthropology. In this case, Wolfram based his categorization on participant observation in addition to a follow-up interview designed to probe aspects of social contact between the African-American and Puerto Rican communities. Note that his categories go beyond census-based "ethnic" categories, instead reflecting associative groups in the community.

Second, the reasons behind quantitative variationists' shift in the direction of probabilistic approaches are also apparent in this example. Looking at the distribution of the variants in table 4.1 is not enough, for instance, to determine the comparability of the distribution of linguistic contextual factors in the interviews of different associative groups, or whether the contributions by subvariants within the variables are compa-

rable (Wolfram 1991, 25). Consider as examples of possible disparities two imaginary conditions: (1) that the distribution in the above case could be the result of a particularly frequent discourse marker that carries the monophthongized realization of the variable in question (such as *like* [la:k]); and (2) that such a marker is unevenly distributed in the speech community, with one of the groups using it much more frequently than the others. In such a case, we would have an irrecoverable distributional anomaly in the data, and the comparison of marginals (raw percentages) would be misleading. Providing frequency counts for each particular phonological context runs into a similar problem, since there are different numbers of total tokens in each group, and contexts before /k/ would be overrepresented in one group versus the other, causing a similar skew in those data.

And yet the following questions remain: Is the skew resulting from unevenly distributed linguistic contexts an artifact of the data collection method? Why do sociolinguistic data require collection methods different from those used in collecting other linguistic data? Couldn't all the distributional anomalies be easily avoided if the researcher controlled contexts and used laboratory elicitation? Part of the challenge of sociolinguistics is to take up the Malinowskian question introduced at the beginning of this chapter: shall we study language as a static entity, as it may occur word by word in isolation, or shall we study it as it unfolds in vivo, minimizing the effects of the laboratory and of the interviewer as much as possible?

The construction of a sample of naturally occurring speech is a different enterprise from the construction of a random sample in a demographic study, or of an experimental paradigm that can control exact numbers of presentations of stimuli, repetitions, ordering of contexts, and so on. Sociolinguistic data differ from census or experimental psychology data in that it is usually impossible to predict how often the relevant phenomenon will occur in the flow of naturally occurring conversation. Contributions to numerical skew and unreliability of pure proportional information and frequency counts may include the following:

1. *Unevenly populated speaker categories.* These may emerge because of distributional facts about the subject population, including rates of response in a door-to-door interview situation, or nature and number of participants in a naturalistic speech activity. Investigating a talk show situation such as *The Oprah Winfrey Show*, with a female talk show host

and a preponderance of female guests, easily illustrates such difficulties. These demographic difficulties as well as the time-intensiveness of transcription lead researchers to rely on a small sample size for speakers and to concentrate on collecting relatively long speech samples from each speaker, the standard length of a sociolinguistic interview being at least an hour.

2. *Widely disparate frequency of forms.* Certain variants of the variable in question may be possible but rare in naturalistic discourse. For example, Matsuda's (1993) study of analogical leveling found that some of the target variants of the potential forms of vowel-stem verbs seldom occurred in Tokyo Japanese discourse, with a frequency of four or five tokens per 90-minute interview. By its very design, the sociolinguistic interview is structured but not controlled, and additional methods may have to be devised (Matsuda's solution was to deploy ingeniously worded questions designed to elicit the elusive constructions (1993, 7)).

3. *High proportion of empty cells.* This is an extension of point 2, but often a mathematically fatal condition for certain kinds of statistical models (i.e., analysis of variance, chi-square) that demand controlled data. For example, phrases that appear to be possible in the combinatorics of generative syntax may be pragmatically restricted or may simply be unattested in the data set.

These factors contribute to the poor fit of sociolinguistic data to summary statistics such as percentages, and to analyses such as sum-of-squares approximations, setting the stage for multivariate probabilistic methods.

4.3 Incorporating Probability into Sociolinguistics

4.3.1 What Is/Was a Variable Rule?

Shortly following his first sociolinguistic studies of New York City, Labov (1969) proposed the *variable rule*. Working within the rule-based framework used in Chomsky and Halle's (1968) *The Sound Pattern of English*, Labov introduced the variable rule by distinguishing it from the categorical rule and

> associat[ing] with each variable rule a specific quantity ϕ which denotes the proportion of cases in which the rule applies as part of the structure of the rule itself. This proportion is the ratio of cases in which the rule actually does apply to the total population of utterances in which the rule can possibly apply, as defined by the specified environment. The quantity ϕ [in a variable rule] thus ranges between 0 and 1; for all categorical rules ... it follows that $\phi = 1$. (1969, 738)

This quantitative extension of the categorical rule framework was followed by a mathematical model developed by Cedergren and Sankoff (1974) and Sankoff and Labov (1979).

A new family of notational conventions accompanied the positing of this new theoretical possibility. One of the best-studied variables in sociolinguistics is word-final *-t/-d* deletion (e.g., [wes] for *west*), a common process that displays morphophonological, class-stratified variability in all English dialects. Variable rules soon ceased to be written with specific frequencies, because depending on a speaker's level of formality or social class the researcher would get differing frequency information, though the ordering and strength of constraints was similar for speakers in the same speech community (Fasold 1991). Thus, the constraints were assigned Greek alphabet letters in order of strength (α being the strongest). The following variable rule describes *-t/-d* deletion in Wolfram's data (Fasold 1991, 4; based on Wolfram 1974):

(1) $[\mathrm{d}] \rightarrow \langle \varnothing \rangle / \langle [-\gamma \mathrm{stress}] \rangle \langle -\beta \# \rangle \_\_\_\_ \# \# \langle -\alpha \mathrm{V} \rangle$

This rule states that word-final [d] optionally deletes when it is (1) in an unstressed syllable, (2) not a suffix, or (3) not followed by a vowel. Deletion is most likely when condition (3), the strongest constraint, is met.

Ordering the constraints is helpful, but it cannot fully describe which choice in the set will be used by a speaker as a member of a particular group. A probabilistic model can be derived to evaluate the contributing influences of each variable constraint. The VARBRUL family of statistical programs was originally developed by Rousseau and Sankoff (1978a) specifically to deal with the quantitative modeling of sociolinguistic data displaying the complexities described above. It is important to keep in mind the distinction between the variable rule theoretical framework for understanding sociolinguistic variation and the VARBRUL family of statistical programs, which is still used despite relative agnosticism by practitioners about what it actually models (Fasold 1991).

4.3.2 A Variable Rule Is Not a Generative Rule

The theoretical proposal of variable rules was immediately viewed with skepticism by generative grammarians and castigated in a series of articles, notable among which is Kay and McDaniel 1979. Here we examine the nature of this debate and its implications for underlying cognitive structures.

Although Labov, Cedergren, and Sankoff did not see the introduction of variable rules as a major departure from the concept of rules in linguistic theory, Kay and McDaniel argued that the variable rule was in fact such a radical departure that "it leads to a conceptual muddle in so far as its proponents think they are working within the generative framework" (1979, 152). To illustrate, Kay and McDaniel borrowed Chomsky's hypothetical context-sensitive rules for a simple natural language. Here rule (2b) is optional:

(2) a. $S \rightarrow ab$
b. $ab \rightarrow aSb$

These rules generate the set of all strings in a language where n instances of a are followed by n instances of b, as in $\{ab, aabb, aaabbb, \ldots\}$. Within this framework, there are different kinds of rules: obligatory rules like (2a), and optional rules like (2b) that specify more than one possibility in the derivation and allow for the generation of infinite sets of sentences with fixed rules. In terms of the hypothetical language above, a third optional context-sensitive rule might be posited, yielding strings such as $\{acbb, aacbbb, \ldots\}$:

(3) $a \rightarrow c \,/\, \_\_\_\, b$

This rule is already extremely close to being a variable rule in the sense introduced by Labov (1969). The only difference is that in addition to having contextual information, a variable rule has frequency information, and where (3) can be stated as "Realize a as c in the context before b sometimes," a variable rule might be stated as "Realize a as c in the context before b 69% of the time, when conditioned by the following variables ..." Kay and McDaniel argued that the leap from "sometimes" to a specific frequency is unwarranted, since "[t]he frequency with which a sentence is produced as an utterance (token) is completely irrelevant. Hence a 'rule' which is concerned with predicting token frequencies is not a rule of (generative) grammar" (1979, 153). Kay and McDaniel noted with alarm the break and incompatibility between the categorical nature of rules in closed, discrete, deductive-inferential systems and the gradient quality of the new variable rules, based on open-ended, continuous, and inductive-inferential systems (Romaine 1985; Givón 1979). But what are the different cognitive implications in these two representational statements?

Sankoff argued that "the formal models of grammatical theory have discrete structures of an algebraic, algorithmic and/or logical nature"

(1985, 75), allowing speakers to make a choice between two or more equivalencies (e.g., allophones) that might carry the same denotation. He continued, "By allowing a degree of randomness into the choice between such alternates, the grammatical formalisms are converted into probabilistic models of linguistic performance." Here, Romaine argued, is precisely where the chasm lies: generative grammars "do not generate true sentences or actual utterances, which are then checked against some corpus; they generate *correct* sentences.... In the most general terms, this type of grammar is a set of devices which check derivations for well-formedness" (1985, 59). Much like the laws of abstract algebra or subatomic physics, which cannot be tested against a corpus, so the aim of linguistic grammars is not to compare their output to naturalistic speech. Romaine further argued that if one were to truly extend the generative framework, a central characteristic of a sociolinguistic grammar would have to be sociolinguistic well-formedness. This sensitivity to social context is already about utterances in the world, and by its very violation of the principles of abstract derivation described above, it fatally fails to conform to the notion of what is meant as the object of description of a generative grammar.

Sankoff did not see variable rules as claiming a particular type of ontological status for the surface output they describe (Sankoff 1988), and yet Labov stated, "We can say that the kinds of solutions offered to problems such as consonant cluster simplification, copula deletion, and negative concord represent abstract relations of linguistic elements that are deeply embedded in the data. It is reasonable to suppose that they are more than the constructions of the analyst, they are the properties of language itself" (1972, 259). This does not necessarily imply that Labov believed in exact isomorphism between models and the phenomena described by the models, as Romaine suggested (1985, 65), but it does point to the possibility of understanding variable rules in two different ways: as a building block in a progressively more exact description of how humans cognitively organize language (Labov), or simply as a statistical "display tool" (Fasold 1991), which sociolinguists may use to discern the various influences in their data.

While during the 1970s much of the debate over variable rules revolved around challenges from generative theoreticians and increasing refinements in the mathematical model, urban dialectology scholarship from the 1980s onward split in two directions: one that adopted variable rules as a modus operandi and applied them in different sociolinguistic contexts

and to larger linguistic domains such as syntax (Weiner and Labov 1983; Rickford et al. 1995) and discourse (Vincent 1991); and one that challenged the use of variable rules altogether because of the perceived lack of a coherent stance on the nature of representation (Gazdar 1976; Sterelny 1983), or over the issue of whether percentages can be part of a speaker's knowledge of the language (Bickerton 1971; Butters 1971, 1972, later reversed in Butters 1990). Other challenges have arisen with the charge that because of their reliance on aggregate data, variable rules obscure information about individual performance (Itkonen 1983; Bickerton 1971; for a refutation see Guy 1980). Especially as generative linguists have moved away from rule-based frameworks and toward constraint-based frameworks like Optimality Theory and the Minimalist Program, most sociolinguists have been less inclined to make statements about the psychological reality of variable rules (Fasold 1991).

Fasold (1991, 10) observes that variable rules are designed to make objectivist predictions about the frequencies with which certain rules would apply under certain contextual conditions. However, we must also consider possible subjectivist probabilistic interpretations—choice models—of variable rules such as that espoused by van Hout (1984).

4.3.3 The VARBRUL Program

As a family of computer programs developed specifically to deal with the data of sociolinguistic variation, the VARBRUL programs are similar to logistic regression models. Practitioners working within the VARBRUL framework use the criterion of maximum likelihood estimation for determining how well a model with a given set of factors fits the data. The full details of the mathematical development of VARBRUL and its relationship to the variable rule framework appear in Cedergren and Sankoff 1974; Rousseau and Sankoff 1978a,b; Sankoff 1985, 1988; Sankoff and Labov 1979 (a reply to Kay and McDaniel 1979); and Rousseau 1989. Detailed instructions for employing the software are available in Young and Bayley 1996.

Binary logistic regression is also available in most modern statistics packages. It either goes by a name such as "logistic regression" (e.g., LOGISTIC in SAS, or Binary Logistic in SPSS) or can be implemented within a generalized linear model (e.g., GENMOD in SAS, or glm in S-Plus), by selecting a link function of "logit" and/or distribution of "binomial." One difference between VARBRUL and commercially available alternatives is the form of reporting of the coefficients, or

"weights," assigned to the selected independent variables. VARBRUL reports weights as probabilities, whereas other programs report them in logit form (i.e., as natural log of an odds). VARBRUL probabilities range between 0 and 1, with values below .5 indicating a disfavoring effect and values above .5 indicating a favoring effect. Corresponding logit values range between negative infinity and positive infinity, and when p is .5, the logit is 0. While no upper or lower bound exists for the logit, it is undefined when p equals exactly 1 or 0 (see discussion in Knoke and Bohrnstedt 1994, 334). Probability weights can be transformed into logit values by taking the log odds; that is, logit $= \log_e(p^i/(1-p^i))$. For further discussion of the logit function, see Manning, this volume, and Zuraw, this volume.

The formulas for the logistic or generalized linear model of VARBRUL in use today are as follows. Formula (1) is the generalized linear model:

$$\log\left(\frac{p}{1-p}\right) = w_0 + w_1 + w_2 + \cdots + w_n, \tag{1}$$

where w_0 is an input weight and $w_1 \ldots w_n$ are contextual factor weights. $\mathrm{Log}(p/(1-p))$ is the logit function, while *log* stands for the natural logarithm (with base e).

For each n, w_n is equivalent to $\log(p_n/(1-p_n))$. Thus, (1) is equivalent to

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{p_0}{1-p_0}\right) + \log\left(\frac{p_1}{1-p_1}\right) + \log\left(\frac{p_2}{1-p_2}\right) + \cdots + \log\left(\frac{p_n}{1-p_n}\right), \tag{2}$$

where p_0 is an input probability and $p_1 \ldots p_n$ are contextual probabilities.

And since $\log xy = \log x + \log y$, (2) is also equivalent to (3), one of the most currently used multiplicative equivalents of (1):

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{p_0}{1-p_0} * \frac{p_1}{1-p_1} * \frac{p_2}{1-p_2} * \cdots * \frac{p_n}{1-p_n}\right). \tag{3}$$

VARBRUL estimates the contextual factor probabilities by combining the input probability (p_0, the likelihood that this variable "rule" may apply in the overall data set, regardless of any contextual influences) with the specific factor weights for all the factors included in the model.

Using a technique based on the dynamic clustering of Diday and colleagues (Sankoff 1985; Bochi et al. 1980), Rousseau (1978, 1989) further developed the log likelihood test, a procedure that tests whether a constraint has an effect that is significantly different from another in its constraint family. This test partitions the data into subsets and compares the difference between the log likelihoods of the subsets, comparing them to an analysis of the data without any partitions. From this test, it is possible to arrive at the optimal likelihood analysis as well as the optimal number of factors within each factor group.

4.4 Stylin' Oprah: A Case Study Exercise in Probabilistic Sociolinguistics

This section will illustrate the use of the VARBRUL program with an extended example drawn from our work on the speech of the American daytime TV talk show host Oprah Winfrey. We will begin with a description of the larger project and then discuss the application of the VARBRUL program to our data.

4.4.1 Data and Analysis

Our work attempts to describe variation in the speech of Oprah Winfrey. Style shifting (intraspeaker variation) in Winfrey's speech has been observed by other analysts and has been characterized as a device to appeal to a cross-section of viewers; most analyses in the literature have centered on topic and lexical choice (Lippi-Green 1997; Peck 1994).

Winfrey herself is originally from Kosciusko, Mississippi. She spent all of her language acquisition years in the U.S. South, attending high school and college (and beginning her broadcasting career at the age of 19) in Nashville, Tennessee. She later moved to Baltimore and then to Chicago where she currently hosts her show. We may then expect that in her speech she would draw from two overlapping repertoires: regional African-American English phonological features of the U.S. South and the supraregional speech variety that is normative in commercial broadcasting.

We suspected context-dependent style shifting at the sociophonetic level in Winfrey's speech and have thus far analyzed some early results on monophthongization of /ay/ (Hay, Jannedy, and Mendoza-Denton 1999; Hay, Mendoza-Denton, and Jannedy 2000). We call the phenomenon *monophthongization* simply for the sake of convenience. It is not our intent here to investigate which variant is underlying, the monophthongal

or the diphthongal one, but merely to posit that Winfrey's speech does vary, and that it does so in a patterned way. We do not assume an abstract representation for the phoneme /ay/; rather, we assume a distribution that encompasses the range from fully monophthongized to fully diphthongized forms.

Style shifting has been shown to be sensitive to many elements of the speech situation, including addressees, topics, referees, and even overhearers (Mendoza-Denton 1999). Rickford and McNair-Knox (1994) found that syntactic and phonological features of African-American Vernacular English covaried in the speech of an African-American teenage girl (Foxy) along two axes: Foxy's speech changed depending on whether her interlocutor was African-American or European-American, and depending on whether the topic was school-related or non-school-related (friendships and recreation). A similar result suggesting a strong unity of "styles" correlating with topics was found by the California Style Collective (1993), who looked at sociophonetic, prosodic, and discourse-marking features and their co-occurrence patterns in the speech of a Euro-American California teenager, nicknamed Trendy. Trendy's index of innovative features, like Foxy's, correlated with school topics, and even with subtopics, such as descriptions of individual groups of people within the social landscape of her school.

In our study, we have isolated samples of *The Oprah Winfrey Show* where Winfrey is talking into the camera or to a television studio audience, without a specific interlocutor. The lack of a specific addressee is crucial: this is the closest we can come in this naturally occurring situation to controlling for the effects of specific interlocutors. Concentrating on the absent persons to whom Winfrey refers in the various segments (who happen to be both topics and referees in this case) allows us to code the segments "about" a referee under a single code and to include the characteristics of these referees as our independent variables.

Most of the segments we coded were short passages describing a particular guest or brief announcements of upcoming shows. For instance, the following transcribed segment, all coded as keyed to the referee "Tina Turner," includes five examples of /ay/ (*my*, *wildest*, *Friday*, *night*, *trying*):

> But let me tell you about tomorrow's show. Tina Turner, we're following Tina around the country, Tina Turner made one of *my wildest* dreams come true, and you're gonna get to see it tomorrow, that's *Friday*. Actually last *night*, we were onstage dancing with Tina Turner. There's a brief look at our rehearsal: that's me, *trying* to keep in step with Miss Tina, you'll see that on tomorrow's show, it's great fun. (*The Oprah Winfrey Show*, May 2, 1997)

It is important here to note that our codings for individual referees are not strictly codings of referents but codings of global referee-as-topic. Thus, in this instance, the coding of the vowel in the first person pronoun *my* is keyed to the referee "Tina Turner," on the basis of prior findings about the importance of topics in the organization of variation (Rickford and McNair-Knox 1994).

A probabilistic model of sociophonetic-level behavior seeks to understand each instance of dependent variable /ay/ as a decision point for the speaker. Following the analogy of Preston (1991), the speaker must decide how to flip the variable coin: whether to pronounce the phonological diphthong /ay/ with a diphthongal phonetic realization [ay], a monophthongal one [a:], or something in between. For the purposes of reporting these results, we will look at the monophthongal realization as the surface variant we are trying to model. We begin with an input weight of .32 for the data set from this speaker (the likelihood that the monophthongal variant will occur across all contexts in her speech), since the monophthongal variant occurs about 32% of the time. Various independent variables such as situational characteristics, variables in the linguistic context, or characteristics of the referee will weight each particular "coin toss" in one direction or another. We attempt to account for factors that may modify this input weight and affect monophthongal realization either by promoting it or inhibiting it. In investigating whether Winfrey will choose to monophthongize /ay/ (if indeed this process can be characterized as residing solely in the speaker's choice space), the question we mean to ask through the use of probabilistic methodology is: What possible social or linguistic factors, or their combination, influence this choice? Possible factors might be sociodemographic characteristics of the referee (in this case the African-American singer Tina Turner), the phonological and prosodic environments of the segment, or the frequency of the carrier lexical item. To test these questions, we coded the data with factors that include both the linguistic or "internal" and referee-sociological or "external" factors.

We coded 229 words containing /ay/ taken from discontinuous selections of approximately six hours of *The Oprah Winfrey Show*, from segments that aired in the 1996–97 season. We examined tokens by means of both auditory and acoustic criteria. Two phonetically trained listeners performed an auditory analysis of the data: a token was coded as monophthongized if and only if the listeners agreed on the classification. To provide acoustic verification of the auditory analysis, the vowel quality

was coded on the basis of spectrographic displays: each token in the data set was labeled either as a monophthong or as a diphthong from wide-band spectrograms. Although monophthongization of /ay/ is a continuous phonetic phenomenon, for the purpose of data entry into the VARBRUL program it must be treated as discrete: preferably as a binary variable, ternary variables being possible but necessitating collapse into the most predictive binary set. This limitation is one of the disadvantages of using VARBRUL analysis when treating continuous variables. Its implications are considerable and will be discussed at length in the next sections, where we compare VARBRUL analysis with other possible analyses.

We were able to distinguish three auditory possibilities for the realization of /ay/: fully diphthongized, fully monophthongized, and somewhere in between. Statistical analyses were carried out for two possible groupings of the tokens in the data set: one that considered only the fully monophthongal tokens as monophthongs, and one that considered both the slightly monophthongal and the fully monophthongal tokens in one category. According to these analyses, the most predictive and consistent results emerged with the latter grouping. Of the 229 tokens of /ay/ in our sample, 32% (74/229) were monophthongized according to the more inclusive definition, and 68% (155/229) were diphthongs. Since the diphthongal realization of /ay/ is normative in the standard language of the media, it is noteworthy that one-third of the tokens were monophthongal.

All the factor groups initially tested in this analysis are listed in table 4.2; statistically significant results, with raw frequencies and probability weights, are reported in table 4.3.

4.4.2 Explanation of Factor Groups and Results

The data were analyzed using Goldvarb Version 2.0 (Rand and Sankoff 1990), a variable rule program for the Macintosh computer. Both the application and its documentation are available online at ⟨http://www.CRM.UMontreal.CA/~sankoff/GoldVarb_Eng.html⟩.

Widely accepted by sociolinguists, the VARBRUL family of programs of which Goldvarb is a member utilizes the maximum likelihood estimate (Sankoff 1988) discussed above. Goldvarb computes probability weights that are expressed as likelihoods, with a probability weight of .5 neither favoring nor disfavoring application of the process in question. Probability weights between .5 and 1 favor the process more strongly the closer they are to the asymptotic 1, while probability weights between .5 and 0

Table 4.2
Variables, factor groups, and factors tested in study of monophthongization in the speech of Oprah Winfrey

Variable status	Factor groups	Factors
Dependent variable	monophthongal vs. diphthongal /ay/	diphthongized slight monophthongization full monophthongization
Independent variables (linguistic/ internal)	preceding phonetic context	voiced obstruents voiceless obstruents nasals liquids vowels/glides
	following phonetic context	voiced obstruents voiceless obstruents nasals liquids vowels/glides
	word class	open closed
	frequency in corpus	infrequent = occurring < 5 times in corpus frequent = occurring > 5 times in corpus
	log-converted CELEX frequency	unattested < log 2 between log 2 and log 4 between log 4 and log 6 between log 6 and log 8 between log 8 and log 10 between log 10 and log 12 < log 12
Independent variables (social/external)	referee gender	male female indeterminate or inanimate
	referee ethnicity	African-American zero referee non-African-American
	individual referee	18 individual referees (see appendix) were given separate codes; "other" category was also used
Variable interactions (social/linguistic)	ethnicity and frequency	African-American infrequent (< log 10) African-American frequent non-African-American infrequent non-African-American frequent zero infrequent zero frequent

disfavor the application of the process more strongly the closer they are to asymptotic 0.

VARBRUL analysis makes the mathematical assumption of an ideal data set with crosscutting factor effects but without significant interactions, where all the factors are independent of one another (Sankoff 1988, 4–19). However, certain factors in this data set are extremely likely to display collinearity. In practice, then, many factors (like word class and raw frequency), being highly correlated, could not appropriately be run together. As a result, only factors that could be assumed to be fairly independent of each other were run together. It is widely believed that social factors may show a high degree of correlation (Bayley 2001), but researchers think that it is relatively rare to find correlations across internal and external variables. Labov (2001, 84) states:

> A full assessment of the effects of intersecting social parameters, and a complete account of sociolinguistic structure, is only possible with multivariate analysis. A multivariate approach was first introduced into sociolinguistic studies in the form of the variable rule program (Rand and Sankoff 1990). It was motivated not by the need to analyze external, social factors, but rather to deal with the language-internal configuration of internal, linguistic constraints on variation (Cedergren and Sankoff 1974). The basic fact about internal factors that the variable rule program continually displays is that they operate independently of each other (Sankoff and Labov 1979). However it was realized from the outset that social factors are typically not independent. Though it is convenient and useful to incorporate external and internal factors in the same analysis, a considerable amount of information can be lost in the typical VARBRUL analysis of speech communities.

Including both internal and external factors is crucial to our data, however, since we found an interaction between lexical frequency calculated on the CELEX database (Baayen et al. 1995), presumably a purely linguistic variable, and ethnicity of referee, a social variable. The results presented in table 4.3 are explained by factor group below.

4.4.2.1 Preceding Phonetic Context and Following Phonetic Context

We coded the immediately surrounding phonetic context of each /ay/ token within and across syllables and words, utilizing categories that have been shown in the literature to affect monophthongization in both African-American and Euro-American U.S. South populations.

Coding monophthongization according to a sonority hierarchy (Selkirk 1984) follows widely accepted methodology outlined by Hazen (2001). We included two other categories as well: vowel/glide and pause. Several studies (Thomas 1995; Schilling-Estes 1996; Wolfram, Hazen, and

Table 4.3
Raw frequencies and VARBRUL probability weights for significant categories in the analysis of monophthongal /ay/. Application = monophthongal /ay/; nonapplication = diphthongal /ay/. First run, with factor groups 1, 2, 3: input 0.323, log likelihood = −110.371, $p < .005$. Second run, with factor groups 1 and 4 only: input 0.263, log likelihood = −106.939, $p < .000$.

Factor group	Factors	Apps.	Non-apps.	Total *N*	% of total *N*	VARBRUL probability weight
1 Following segment	vowel/glide	9	35	44	19	0.799
	%	20	80			
	liquid	22	17	39	17	0.804
	%	56	44			
	nasal	13	35	48	21	0.436
	%	27	73			
	voiced obstruent	9	35	44	19	0.384
	%	20	80			
	voiceless obstruent	13	62	75	33	0.320
	%	17	83			
	pause	3	1	4	2	unreported owing to low token count
	%	75	25			
2 Ethnicity of referee	zero referee	19	15	34	15	0.700
	%	56	44			
	African-American referee	39	49	88	38	0.622
	%	44	56			
	non-African-American referee	16	91	107	47	0.336
	%	15	85			

3 CELEX frequencies						
3a Log value, 5-way split	< log 6	1	20	21	9	0.063
	%	5	95			
	log 6–log 8	17	33	50	22	0.478
	%	34	66			
	log 8–log 10	10	39	49	21	0.418
	%	20	80			
	log 10–log 12	17	36	53	23	0.596
	%	32	68			
	> log 12	29	27	56	24	0.734
	%	52	48			
3b Log value, binary split	infreq = < log 10	28	92	120	52	0.370
	%	23	77			
	freq = > log 10	46	63	109	48	0.642
	%	42	58			
4 Interactions	African-American infrequent	12	31	43	19	0.437
	%	28	72			
	African-American frequent	27	20	47	21	0.781
	%	57	43			
	non-African-American infrequent	6	58	64	28	0.177
	%	9	91			
	non-African-American frequent	10	31	41	18	0.576
	%	24	76			
	zero infrequent	9	4	13	6	0.783
	%	69	31			
	zero frequent	10	11	21	9	0.725
	%	48	52			
All totals	token count	74	155	229		
	% of data	32	68			

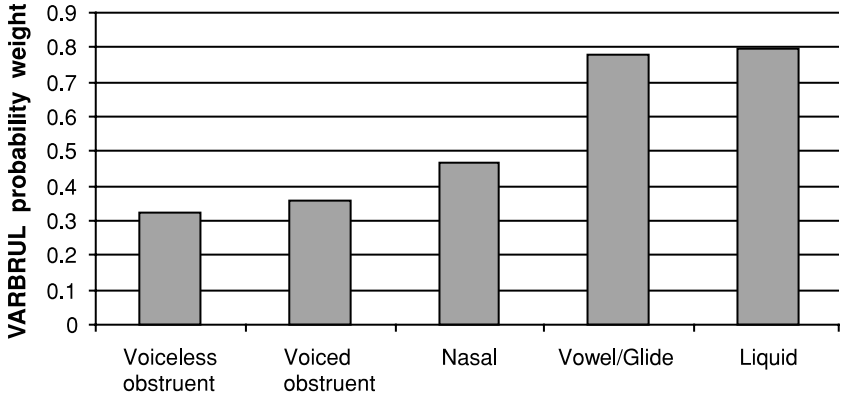

Figure 4.2
VARBRUL weights for following phonetic category as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

Schilling-Estes 1999) have shown that the expected descending order of following phonetic environments favoring monophthongization for *African-Americans* is liquids > nasals > voiced obstruents > voiceless obstruents. Our data for following phonetic context fit the expected pattern, with probability weights as follows (see also figure 4.2): liquids .804 > vowel/glide .799 > nasal .436 > voiced obstruent .384 > voiceless obstruent .320. In this data set, following voiceless and voiced obstruent contexts heavily disfavored the expression of monophthongal /ay/, while a following nasal neither disfavored nor favored the process. Only liquid and vowel/glide following contexts strongly promoted the expression of monophthongized variants. Because Euro-American U.S. Southerners exhibit a different pattern, with high rates of monophthongization before voiced obstruents, we believe that Winfrey's use of /ay/ is indexical of variation in the African-American community.

Preceding phonetic context was not a significant predictor of variation in this data set and was discarded in the final analysis.

4.4.2.2 Word Class As with other reductive processes (Wright 1997), monophthongization may apply at different rates among words depending on their frequency. In an earlier analysis, we tested lexical frequency within the Oprah Winfrey corpus and found that it was highly correlated with monophthongization (Hay, Jannedy, and Mendoza-Denton 1999).

Frequency can be a difficult metric to use because it may be partly confounding a linguistic factor: whether a word belongs to a closed class or an open class. To test this confound, we coded open and closed classes separately from frequency. When run by itself as the only independent variable, word class is highly significant in predicting the patterning of our data. Open class words disfavored the monophthongization process with a probability weight of .397, while closed class words favored it with a weight of .643 (log likelihood = −137.896, $p < .001$). Although both word frequency and word class were significant on their own, the most predictive model of the data was found by using the log-converted CELEX frequency category (see section 4.4.2.4).

4.4.2.3 Raw Frequency in the Corpus One issue when trying to use lexical frequency as a predictive factor in the study of a naturally occurring sample is whether to use the word frequency of the sample itself or some independent metric of word frequency in the language as a whole (because the sample might not be representative of the speaker's overall repertoire). In our case, words that were very frequent in the sample were words like *style* (from a segment of *The Oprah Winfrey Show* called "The House of Style") and *wild* (the descriptor of Tina Turner's "Wildest Dreams" tour).

As a first step toward assessing the importance of frequency, we used the raw frequency within our corpus and divided the words into "frequent" (>5 occurrences in the sample) and "infrequent" (all other words). This distinction also yielded significant results: infrequent words disfavored monophthongization with a probability weight of .329, while frequent words slightly favored it with a weight of .589 (log likelihood = −138.474, $p < .001$). Although significant on its own, raw frequency in this corpus was overshadowed by the log-converted CELEX frequency, which contributed more substantially in fitting the model to the data.

4.4.2.4 Log-Converted CELEX Frequency Another frequency metric that we used was frequency in the English language according to the CELEX corpus. The CELEX database (Baayen et al. 1995) from the Max Planck Institute for Psycholinguistics in Nijmegen incorporates the 17.9-million token COBUILD/Birmingham corpus, and in addition represents more than 90,000 lemmas from dictionary entries. All the sources for CELEX are textual, about 15% coming from U.S. authors.

Despite the differences (oral vs. textual, U.S. vs. composite) between our raw frequency corpus and the CELEX corpus, CELEX codings were better able to account for variation in our data. This strongly suggests that the processes at work in the patterning of our data transcend these particular instances of *The Oprah Winfrey Show* and may well be operating in other contexts as well.

The CELEX ordinal frequency ranking for each token was converted to a log-based frequency code because there is good evidence that humans process frequency information in a logarithmic manner. That is, a frequency difference occurring among the lower frequencies carries more weight than a frequency difference of equal magnitude occurring among the higher frequencies. Since VARBRUL requires discrete independent variables, in order to input the data we created a five-way log value split that provided a near-perfect cline of influence in which the most frequent words ($>$ log 12) strongly favored monophthongization (probability weight .734), while the least frequent words ($<$ log 6) strongly disfavored it (probability weight .063) (see figure 4.3). A binary (median) log value division was also devised. Words of frequency $<$ log 10 strongly disfavored monophthongization (probability weight .370), while words of frequency $>$ log 10 favored it (probability weight .642) (see figure 4.4). We used the binary division to code for interactions between word frequency and ethnicity.

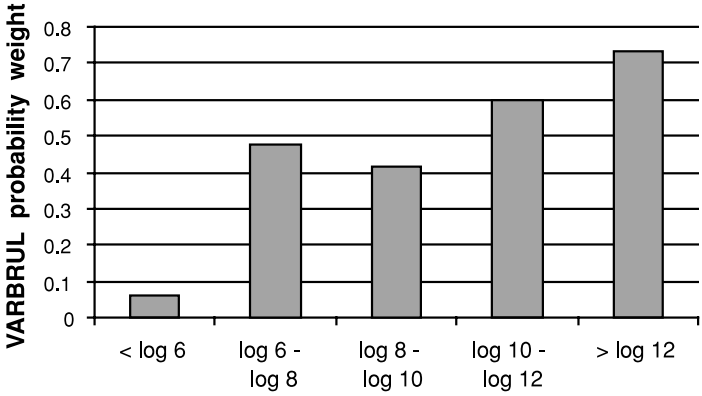

Figure 4.3
VARBRUL weights for lexical frequency as a predictor of monophthongization: results for log-converted CELEX frequency. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

4.4.2.5 Individual Referee To investigate the possibility that Winfrey was treating each referee in an idiosyncratic or individualized way, and not according to gender or ethnicity, we also assigned segments codes that referred to people individually. Nineteen referee codes (including "other" for segments that were not about particular people) were used (the full list is given in the appendix). The "other" code was used primarily when Winfrey spoke straight into the camera without a specific addressee. These segments fulfilled our requirement that the speech have no specific interlocutor, and they included a segment called "Gratitude Moments," where Winfrey spoke about her favorite things, one where she spoke about her birthday, and one where she warned the audience about getting scammed (robbed). One of our initial hypotheses was that the individual codes would be important predictors of variation. However, it was not borne out in the VARBRUL results and was eliminated. In our later analysis using CART trees, the individual codes became an important factor.

4.4.2.6 Referee Gender So-called external or social variables that we coded in the corpus included the referee's gender. By itself, again, gender was significant, but not when included in a statistical run with any other factor. Codings for this factor included male, female, and "other," used for situations where the referent did not have a gender or where gender

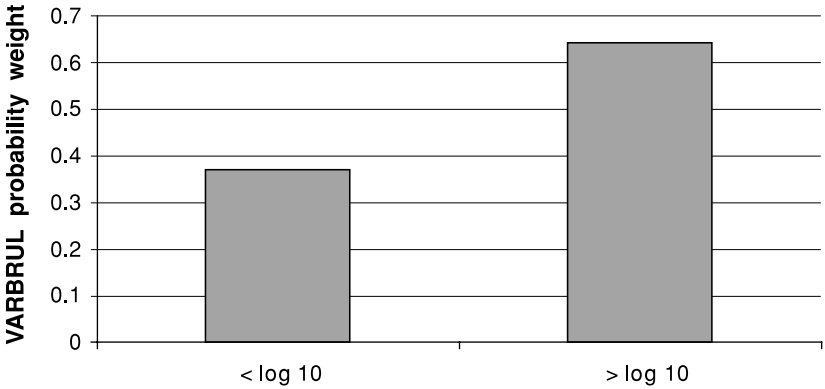

Figure 4.4
VARBRUL weights for lexical frequency as a predictor of monophthongization: results for log-converted CELEX frequency (cutoff at median). Values above 0.5 favor monophthongization, values below 0.5 disfavor.

could not be determined. Oddly enough, there were no statistically significant differences between rates of monophthongization for female and male referees (both of these were neutral, f: .468, m: .435), while the "other" category showed a markedly favoring effect, o: .730 (log likelihood = −139.252, $p < .01$).

4.4.2.7 Referee Ethnicity The ethnicity of the referee was the most important factor group (first selected in the Goldvarb step-up/step-down procedure) in modeling the monophthongization of /ay/. We coded referee ethnicity according to three categories: African-American referees (strongly favoring monophthongization; probability weight .622); non-African-American referees (strongly disfavoring; .336); and zero referee, which favored monophthongization more strongly (.7) than the other categories. These weights are shown in figure 4.5. However, it was also clear from our analysis that ethnicity of referee also interacted strongly with word frequency. And VARBRUL assumes that the different factors included in a single analysis act independently of one another.

4.4.2.8 Ethnicity and Frequency One solution to the problem of assuming factor group independence in VARBRUL is to create an inde-

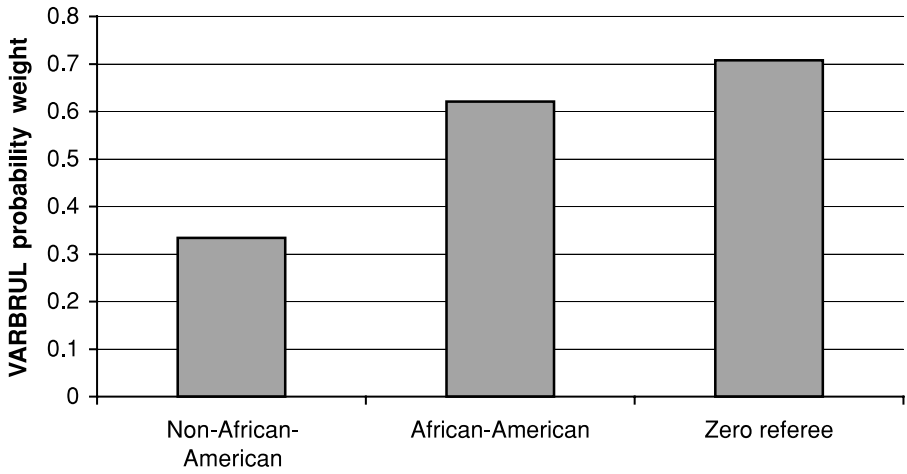

Figure 4.5
VARBRUL weights for ethnicity of referee as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

pendent factor group that combines the interacting factors into discrete possibilities in order to isolate their effects (Bayley 2001). We did this by creating an additional interaction factor group that combined the six possibilities resulting from the interaction of two frequency categories (frequent vs. infrequent words; i.e., $>$ log 10 vs. $<$ log 10 in the binary CELEX coding) and three ethnicity categories (African-American, non-African-American, and zero referee). Our results were most puzzling: they showed a significant interaction in what we had originally coded as two separate predictive factor groups. When combined, the ethnicity of referee/binary CELEX frequency factor group was intriguingly arranged thus (see also figure 4.6): [no ref, infrequent .783 > African-American, frequent .781 > no ref, infrequent .725 >] non-African-American, frequent .576 > African-American, infrequent .437 > non-African-American, infrequent .177. The bracketing around the first three factors indicates that according to the difference-in-log-likelihoods test (Rousseau 1989), these factors are not statistically significantly different from each other and should be collapsed. They are shown separately here for expository reasons.

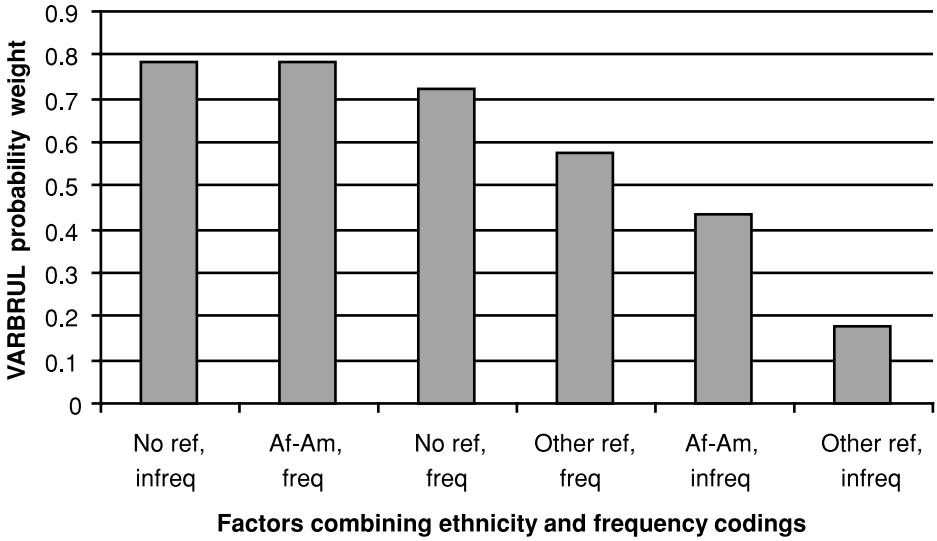

Figure 4.6
VARBRUL weights for the interaction between referee ethnicity and lexical frequency as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

Initially, we found this result—that either a frequent or an infrequent word with a zero referee was as likely to lead to monophthongization as a frequent word with an African-American referee—difficult to explain, especially since there is such a strong distinction in likelihood of monophthongization between African-American and non-African-American referees. What could this mean? Colleagues have suggested to us that zero referee might be Winfrey's baseline style, and that this might be close to a style used for African-American referees. And, as discussed below, much of the zero referee section included frequent self-reference, specifically using the words *I* and *my*. Of course, self-referring words are also frequent words, so it is difficult to disentangle the two effects and precisely identify the locus of the observed patterns. Because the two predictors are highly correlated with one another, one cannot simply include them both in a logistic regression, to see which one is the stronger predictor. The technique we have used assumes strict independence among the factors. In the next section, we explain how we set about investigating the self-reference effect. For now, we return to the interaction between ethnicity of referee and lexical frequency.

From the interaction of frequency and ethnicity of referee, given our understanding of frequent words as the carriers of style (Hay, Jannedy, and Mendoza-Denton 1999) we expected frequency to have a much bigger effect in speech relating to African-American referees (where frequent words should be prone to monophthongization for both stylistic and articulatory reasons) than in speech relating to non-African-American referees (for whom we expect some canceling out of the articulatory tendency to monophthongize frequent words, given that the stylistic setting favors diphthongs). In fact, our results show the opposite: word frequency has a much bigger effect for non-African-American referees than for African-American referees. Upon closer examination, we believe this result does not necessarily contradict our assumptions about frequency and style; rather, it reflects an asymptote for this type of variation. When referencing African-Americans and using frequent words, Winfrey reaches the limit of her range of variation. VARBRUL probability weights around .78 set the upper bound of monophthongization that can be found in Winfrey's speech. Essentially, there is a ceiling effect, indicating that in no speech situation will she go beyond her personal maximum of variation (even just doubling her rate for infrequent words with African-American referees would overshoot this asymptote).

Frequency thus has the biggest effect within the subset of the data that is not otherwise prone to monophthongization (non-African-American referees), while referee ethnicity has the biggest effect within the subset that is not otherwise prone to monophthongization (the infrequent words).

4.4.2.9 Self-Reference, Style, and the Use of Qualitative Analysis To disentangle the effects of lexical frequency and ethnicity, we inspected the show transcripts and found specialized discourse patterns in the use of the highly frequent words *I* and *my*. The segments coded as "zero referee" consisted largely of Winfrey self-disclosing to her audience. The segments "House of Style," "My Favorite Things," and "Oprah's Birthday" are frequently self-referring. This self-disclosure feature of Winfrey's television persona—and the genre of daytime talk shows in general—has received a great deal of attention from scholars (Shattuc 1997; Masciarotte 1991). In terms of our data, a style of conversational engagement through self-disclosure means that Winfrey talks about her own past encounters with the people to whom she refers, sharing her personal history in great detail. Guests she knows well elicit more self-reference, so that a short segment on Michael Jordan, with whom Winfrey has a famously close relationship, included 8 self-referring tokens out of 17 /ay/ tokens, or 47% of the tokens for that segment. The segment "My Favorite Things/Birthday Presents" included 23/35 or 65% self-referring tokens. A segment about Mia Farrow, by contrast, included only 1/20 or 5% self-referring tokens.

The use of highly frequent words as stylistic devices in the genre of talk show hosting may boost overall perceptual saliency of the variable and make it a good candidate for the display of speaker style. Other examples of highly frequent words used as iconic displays of speaker and group style can be found in Mendoza-Denton 1997 and California Style Collective 1993.

When included in a VARBRUL run with ethnicity and with following phonetic context only, self-referring words joined these factors in a set that best predicted variation in the data. Self-referring words correlated positively with monophthongization, exhibiting VARBRUL weights very similar to those of the zero referee category; non-self-referring words had probability weight .398, while self-referring words had weight .680 (log likelihood $= -109.428$, $p < .001$) (see figure 4.7).

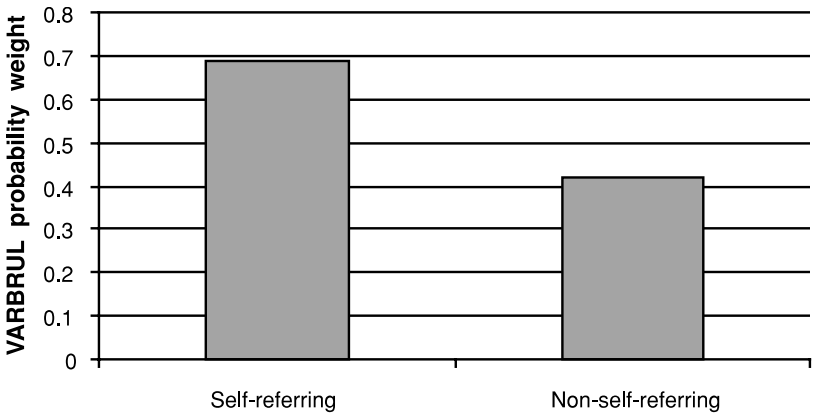

Figure 4.7
VARBRUL weights for self-reference as a predictor of monophthongization. Values above 0.5 favor monophthongization, values below 0.5 disfavor.

We believe both frequency and self-reference are playing a role in the aggregate data set. The observed frequency effects are spread throughout the whole frequency range (see table 4.3, where frequency effects were significant even when split into five frequency categories), and so they cannot be attributed only to self-reference. However, it is consistent with the observed discourse patterns to hypothesize that Winfrey's self-referring speech might be particularly prone to monophthongization—although, as explained above, because of the collinearity of the factors this is best investigated qualitatively. Precisely disentangling the relative contribution of frequency and self-reference would require a larger data set and remains a task for future work.

4.4.3 An Alternative to VARBRUL: Classification and Regression Trees (CART)

In this section, we briefly explore the patterns in our data further, using a different statistical approach.

The construction of classification trees is essentially a type of variable selection. Such trees are a valuable tool for exploratory data analysis and can handle missing values or empty cells with ease, tree construction being based on the cases that do not have missing values. Classification trees are an attractive method of data exploration because they handle interactions between variables automatically. They also have the advan-

tage of being completely nonparametric. No assumptions are made about the underlying distribution of the data. These features make them less powerful for detecting patterns in data, but fairly reliable in terms of the patterns found.

Classification trees do assume that the effect being modeled is organized into discrete factors. An analogous class of models, regression trees, deals with continuous data.

Foundational literature on classification and regression trees includes Morgan and Sonquist 1963, Morgan and Messenger 1973, and Breiman et al. 1984. A good practical guide for their implementation in S-Plus can be found in Venables and Ripley 1994.

A classification tree begins with the data to be analyzed and then attempts to split it into two groups (here, one that maximizes monophthongization, and one that minimizes it). Ideal splits minimize variation within categories and maximize variation across categories. All possible classifications of the independent variables are attempted. Tree construction works one step at a time, so once the first split is achieved, an optimal split is sought for each resultant node. The particular technique used here (that implemented in S-Plus/R) allows only binary splits. At any given node, the maximum reduction of deviance over all possible splits is used to identify the best split. This process continues until either the number of cases reaching each leaf is small or the leaf is sufficiently homogenous relative to the root node.

This process often grows a tree that overclassifies the data. That is, a tree may fit a particular data set extremely well, but may be unlikely to generalize if new data points are added to the analysis. A selection process can then be used (akin to the stepwise procedure used in multiple regression) to determine which divisions should appropriately be included in the model and which are best discarded—a process known as tree pruning (Breiman et al. 1984). There are a number of different methods for choosing where to prune the tree (i.e., for deciding which nodes can best be removed).

One method of tree pruning uses a process of cross-validation. The data set is divided into subsets, and separate trees are grown on the basis of each subset. The trees based on each subset of the data can then be compared with one another. As Venables and Ripley (1994, 44) explain, "Suppose we split the training set into 10 (roughly) equally sized parts. We can then use 9 to grow the tree and test it on the tenth. This can be done in 10 ways, and we can average the results." This process returns an

averaged deviance for trees of each possible size. In the analysis presented below, we used this cross-validation technique—pruning the tree to the smallest tree size with the minimum deviance. This represents a fairly conservative approach to tree building.

When we attempted to build a tree based on the monophthongization data, we allowed for the possible contribution of the following variables: the individual identity, ethnicity, and gender of the referee; the class and frequency of the word; the preceding and following phonetic environment. Of these, only two remained in the pruned tree: the identity of the individual and the following phonetic environment. Because each branch of the tree deals with successively smaller sets of data, a fairly large data set is required to establish the coexisting significance of a sizable number of contributing factors. The power of this technique is therefore slightly limited when dealing with small data sets—especially if these data sets display much variability.

The pruned tree is shown in figure 4.8. The first and most important split is between segments where Winfrey is talking about Tina Turner, Will Smith, Halle Berry, or no one in particular (group (c)), and all other segments. In the former four instances, she was much more likely to monophthongize /ay/ (60% of tokens) than in all others (17%).

These two nodes split further into two subcases. The left branch splits into two more sets of individuals: those who strongly discourage monophthongization (group (a): 2%) and those who are more likely to lead to monophthongization (group (b): 23%). Finally, among group (c) the classification algorithm detects a significant effect of the following environment: monophthongization is more likely preceding liquids and nasals than other phonological segments.

Other variables were included in the full tree (lexical frequency is the next factor to appear), but did not survive the pruning process. Because a classification tree looks for patterns in progressively smaller sets of data, we would likely need a much bigger data set than we currently have in order for it to reveal the full range of complexity in our data. Those factors that do survive the pruning process, however, are ones in which we can have extreme confidence.

The classification algorithm divides individuals into three groups. No African-American referee appears in group (a), 3 African-American referees appear in group (b) (3/8, 38%), and the individuals identified in group (c) are all African-American and are grouped together with the zero referee cases.

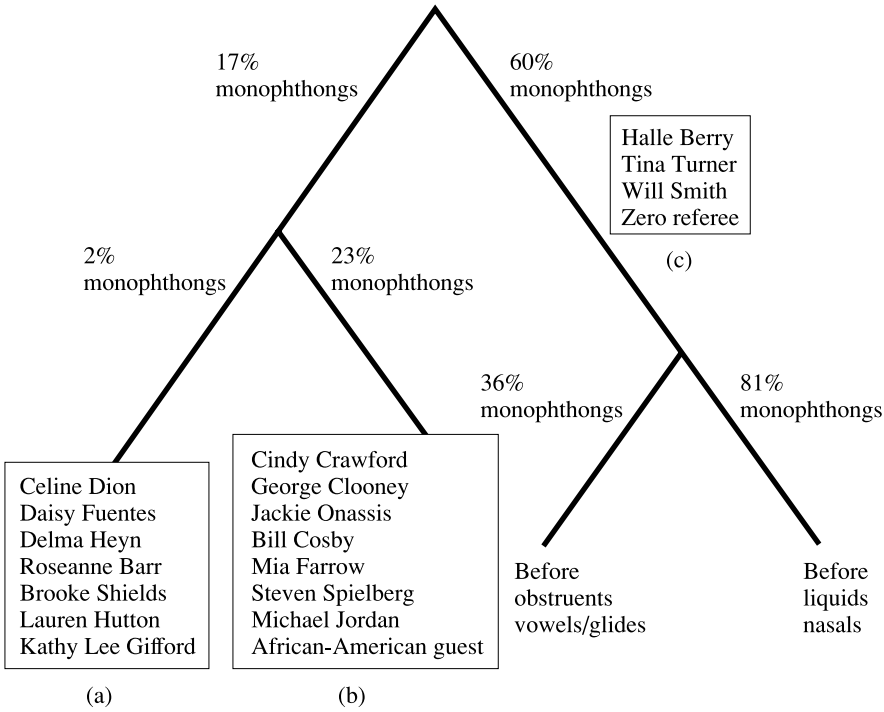

Figure 4.8
CART classification tree for monophthongization

This "individual identity" variable therefore echoes the effects of ethnicity, while imbuing it with an added level of subtlety. It could perhaps be seen as organizing people into groups according to the nature of Winfrey's involvement with them—ethnicity being one component of this (though probably not a binary component), and other dimensions of solidarity perhaps also playing a role. Because the classification tree is an excellent tool for revealing significant groupings in data, it can be used to reveal natural groupings of individuals, which (as inspection of the groups reveals) could not be replicated by any combination of standard social variables.

4.4.4 The Oprah Winfrey Data: Summary

4.4.4.1 Analysis Techniques Using a relatively small data set, we have shown how inferences can be derived from it in different ways. What we

hope to have demonstrated with this exercise is that different types of analysis can assist in the interpretation of results. In our case, we used two types of quantitative analysis (VARBRUL and CART) as well as qualitative analysis (looking at patterns of self-reference). Two basic results emerge unambiguously from our study: Winfrey's style shifting is partially conditioned by the ethnicity of the person she is referring to, and partially by the following phonetic environment.

Subtler nuances of the data—the role of lexical frequency, the presence of interaction effects, the emergence of natural groupings of individuals—are highlighted differently by the different statistical techniques we used.

Each technique has drawbacks. Since classification trees are local optimizers, once an initial split is made it is impossible to ask what the overall effect of a second factor is, given the first one. And in order to examine the effect of a large number of variables using a classification tree, a large data set is required. VARBRUL is not well equipped to easily explore possible interactions, nor is it equipped to deal with continuous dependent or independent variables, although these limitations can be overcome by the use of logistic regression in commercially available statistics packages.

The VARBRUL program effectively implements a binomial stepwise regression analysis. It models a binomial outcome, using discrete factors. In this sense, it is an appropriate tool to use in the formulation of variable rules as they were originally conceptualized—rules that predict which of *two discrete outcomes* will occur on the basis of *discrete factors*, such as the gender or ethnicity of the speaker (or in the case of our data, the referee) or the identity of the phoneme that precedes or follows. Continuous independent factors can be built into the model, by breaking them up into discrete groupings—a technique that imposes artificial category boundaries on the factor.

And yet monophthongization is not really discrete. Different degrees of monophthongization (or diphthongization) exist, and Winfrey exploits the full range of this continuum in her performance of style. Winfrey does not shift between discrete styles (an "African-American referee," a "non-African-American referee," and a "self-disclosure" style); rather, she seamlessly navigates a range of continuous stylistic dimensions, and the degree to which she employs monophthongization signals (together with many other variables) where she is positioned in stylistic space. Monophthongization is not discrete, ethnicity is not discrete, nor is lexical

frequency. And yet we impose boundaries on all of these in order to simplify our analysis and detect statistical patterns that will shed light on language and on society.

We believe one challenge for the future of probability theory in sociolinguistics is to move beyond the limitation of discrete categorization and to work toward understanding how gradient, continuous linguistic variables are conditioned by both categorical and continuous social and linguistic factors. Such analyses have begun to appear, notably Berdan's (1996) study of second language acquisition, where he used the logistic regression module in SPSS to model time as a continuous independent factor. Sudbury and Hay (in press) also model continuous independent factors (time and frequency) in their analysis of rhoticity and /r/-sandhi.

Modeling and understanding the combination of continuous and discrete factors in predicting gradient implementation of sociolinguistic variables will be a major challenge in the future of probability theory in sociolinguistics—one that will require an adjustment in the way data are collected and analyzed, and in the statistical techniques used to explore the patterns of variability within those data.

As illustrated in our Oprah Winfrey data analysis, if one of the predictor variables in the hypothesized model is continuous (such as lexical frequency or age), VARBRUL is unable to model it as a continuous predictor; instead, the researcher must break it up into a number of discrete sets. This does not tend to be a feature of more general implementations of logistic regression, which can unproblematically model continuous variables. Thus, as discussed by Berdan (1996) and Bayley (2001), the VARBRUL implementation may not be the most appropriate for data sets that involve one or more important continuous independent variables.

And while it is possible to encode interactions in VARBRUL by creating hybrid categories (see, e.g., Sankoff and Labov 1979, 204; also the example in section 4.4.2.8), this solution is not straightforward, and it requires that the researcher identify the possible interaction in advance. Other implementations of logistic regression tend to allow possible interaction effects to be explored in a more straightforward way. Sigley (2001) tested for the presence of interactions in seven previously reported data sets and found that about 26% of pairwise tests produced significant interactions. He argues that interaction effects are widespread and are

potentially just as important as main effects when modeling (socio)-linguistic variation. For further discussion of problems associated with interactions in VARBRUL, see Young and Yandell 1999 and Bayley 2001.

Another major challenge for sociolinguistics lies in finding appropriately sophisticated frameworks with which to understand the patterns that probabilistic analyses reveal—frameworks with adequate insight and explanatory power. The patterns revealed by our Oprah Winfrey study need explanation in many areas. Here we address just two: What are the *cognitive* patterns and processes through which such patterns of intra-speaker variation arise? And what are the *social* mechanisms and constructs that condition the observed behavior?

4.4.4.2 The Cognitive: What Is Variation? Our analysis of Winfrey's monophthongization patterns gives a good indication of their characteristics. Her orientation toward the person she is talking about (an important component of which is the person's ethnicity) affects the likelihood (and probably the degree) of monophthongization. Monophthongization is further influenced by the phonetic environment and by the lexical frequency of the word it appears in.

So what are the cognitive implications of these findings? What is Winfrey doing when she style-shifts? Models of speech production do not currently account for sociophonetic variation, even though this is a large part of what people do when they produce speech. One component of speech is clearly the continuous signaling of social identity and orientation. In order to satisfactorily begin to model this process, sociolinguists and those who work on speech will need to combine efforts.

One promising interpretation is that producing a phoneme (or word) involves the activation of a *distribution of phonetically detailed remembered examples* that characterize that phoneme (or word). More prototypical or central exemplars will be easiest to access, because of their central status in the distribution; and particularly frequent examples will also be easy to access, because of their high resting activation level. Exemplar theories of speech production and perception have been developed by, among others, Pierrehumbert (2001a, in press) for production and Johnson (1997b,c) for perception. Exemplar models are promising candidates for modeling sociophonetic effects because they do not treat variation as noise; on the contrary, variation is central and is inherently coded in lexical representations. Such models would appear to provide a

natural explanation for the involvement of lexical frequency in style shifting, as well as for why intraindividual style shifting echoes the interindividual social distribution of variables to which a speaker has been exposed (Bell 1984).

In Pierrehumbert's (2001a) implementation of exemplar theory, the selection of a phonetic target is modeled as random selection from a cloud of exemplars associated with the appropriate category. This models many social effects well, because "although social and stylistic factors may select for different parts of the exemplar cloud in different situations, the aggregate behavior of the system over all situations may be modeled as a repeated random sampling from the entire aggregate of exemplars" (Pierrehumbert 2001a, 145). Pierrehumbert demonstrates how a model with fully remembered exemplars can account for the fact that frequent words lead historical leniting changes and can model the timecourse of certain types of phonological merger.

The implementation is modified in Pierrehumbert, in press, so that production does not involve the specific selection of an exemplar, but rather can be heavily biased by activated exemplars. Exemplars are weighted and can be activated to different degrees in different contexts. Weighting can be affected by sociostylistic register and by contextual and attentional factors.

Goldinger (2000), Kirchner (in press), and Bybee (2001) also advocate exemplar-based models for speech production. And results reported by Goldinger (1997), Niedzielski (1999), Strand and Johnson (1996), and Whalen and Sheffert (1997), among others, provide strong evidence that social and speaker-specific information is not only stored, but also actively exploited in speech perception. Such results are highly consistent with models that include an exemplar-based level of representation, and they are very difficult to account for in models in which detailed exemplars are not stored.

Docherty and Foulkes (2000) have attempted to situate a discussion of sociophonetic variation in an exemplar model of lexical representation. Such a model accounts nicely for other patterns of variance such as coarticulation, connected speech processes, background noise effects, and intra- and interspeaker variability, and so, as Docherty and Foulkes point out, this seems a natural place to start. One of their central questions is "how phonology stands in the face of the variable aspects of a speaker's performance ..." (p. 112). It would certainly seem that modeling sociophonetic variation would be a crucial test of the degree to which any

model of phonetic or phonological production and perception succeeds. However, it is not just models of speech that could benefit from such an understanding. Sociolinguists' understanding of the factors that are involved in style shifting, both linguistic and social, and the potential and possible ways in which they interact, would be deeply enriched by a clear understanding of the mechanisms through which this variation is represented and produced.

Resolving the nature of the cognitive status of probability distributions found in sociolinguistic studies would certainly make researchers' understanding and modeling of these phenomena more sophisticated and open new doors for analysis and explanation. By embedding studies of language variation in an understanding of language perception, production, and reproduction, researchers can start to consider how the observed probability distributions may come about, and how they might propagate, spread, and be manipulated in different social contexts for different social ends.

4.4.4.3 The Social: What Is Style? In the exemplar-theoretic view outlined above, social information that is interpretable by a listener is automatically stored with the exemplar, made more robust with repetition, and crucially linked to the actual instances of use of a particular variant. The proposal that linguistic categories, targets, and patterns are gradually built up through incremental experience with speech is entirely compatible with a view of the social world that relies on gradually built up social categories that emerge from the experiences that surround individuals as social actors. Just as there are no preset categories in phonology, and phonemes are abstracted from statistical patterning of the input (see Pierrehumbert, this volume, for extensive supporting evidence), so are social patterns abstracted and recovered from the same input.

We underscore the importance of interpretability by the listener. Within both the linguistic and the social world, young learners or foreign language speakers may not be equipped to fully understand the category composition of the stimuli to which they are exposed. It is only with repeated exposure that a child or a nonnative speaker can develop a robust enough model to incorporate and interpret new examples.

Because the development of an exemplar-based model proceeds example by example, it is important to look not only at overall distributions and gross statistical generalizations, but also at the micropatterning of individual instances. Understanding the flow of on-line discourse and its

relationship to robustness for both linguistic and social categories is an urgent task for sociolinguistics. Earlier, we mentioned that many of the social categories that researchers assume as given are not discrete, but may be treated as discrete for the purposes of statistical convenience. By supplementing statistical methods with qualitative analysis, we have exemplified one possible way to investigate how categories are built up in naturalistic contexts.

4.5 Conclusion

The use of probabilistic methods has led to important breakthroughs in sociolinguistics and has played an extremely important role in shaping the study of language variation and change. An important challenge for the future will be to move toward a more unified understanding of how subtle, gradient patterns of variation affect and are affected by cognitive, linguistic, and social structures, while always remembering that choices made for the analyst's convenience (such as treating monophthongization or ethnicity as binomial variables) are not pure mirrors of discrete categories in the world. We believe that the strongest theory of the interaction of language and society is a probabilistic theory, yet we encourage probabilistic sociolinguistic scholars to go beyond current methods: uncollapse what has been collapsed, and look for finer-grained social-theoretic explanations within what is uncovered in aggregate patterning.

Appendix

This appendix lists the different referees for the segments analyzed. Individuals were coded as "African-American" or "non-African-American." The "zero referee" cases involve segments in which the discourse is not focused on a specific individual.

Roseanne Barr, F actor	non-African-American
Halle Berry, F actor	African-American
George Clooney, M actor	non-African-American
Bill Cosby, M actor	African-American
Cindy Crawford, F model	non-African-American
Celine Dion, F musician	non-African-American
Mia Farrow, F actor	non-African-American
Daisy Fuentes, F actor	non-African-American
Kathy Lee Gifford, F actor	non-African-American
Delma Heyn, F writer	non-African-American
Lauren Hutton, F actor	non-African-American

Michael Jordan, M basketball player	African-American
Jackie Onassis, F celebrity	non-African-American
Brooke Shields, F actor	non-African-American
Will Smith, M actor/musician	African-American
Steven Spielberg, M movie director	non-African-American
Tina Turner, F musician	African-American
F Guest who dreams of having a house	African-American
"Gratitude Moments"	zero referee
"Oprah's Birthday"	zero referee
"How to Avoid Getting Scammed"	zero referee
"House to Style" (how to have more of it)	zero referee
"Oprah's Favorite Things"	zero referee

Note

The authors would like to acknowledge Rens Bod, Janet Pierrehumbert, and Kie Zuraw for extensive comments and suggestions. Malcah Yeager-Dror provided helpful guidance and Matt Loughren helped with references. All errors and omissions remain our own.

Chapter 5

Probability in Language Change

Kie Zuraw

5.1 Introduction

Why do languages change? If children are able to infer surrounding adults' grammatical systems from their utterances, and if adults adjust their lexicons and perhaps grammars to achieve better communication with their interlocutors, any linguistic innovations that might somehow arise should be quickly stamped out. This is a combination of Weinreich, Labov, and Herzog's (1968) "actuation problem" (how and why does a particular change occur at a particular time?) and what we might call the "continuation problem": what sustains the momentum of a change, causing an innovation to increase in frequency, to spread from word to word, or to spread from speaker to speaker, rather than stalling or receding?

Any answer to the continuation problem must rely on a probabilistic model of the language faculty. If the rise in frequency of an innovation results from snowballing mislearnings, we require a model of how learners respond to their variable environment. If the rise in frequency results from individuals' adopting the speech patterns of some social group, we require a probabilistic model of the speech community, in which individuals probabilistically and incrementally update their grammars and lexicons in response to interlocutors' behavior. Moreover, when the rise in frequency of an innovation involves variation within individuals, as we can often see that it does in written records, we require a probabilistic model of language representation and/or use. Otherwise, we have no way of representing the difference between a generation whose members use a new variant 20% of the time and a generation whose members use a new variant 40% of the time.

The fact that language change happens seems to demand a probabilistic view of the language faculty, in phonology, morphology, syntax, semantics, processing, acquisition, and the social use of language. The probabilistically oriented study of language change therefore relies on probabilistic models of all the areas of linguistics discussed in this book.

This chapter surveys the role of probability in the study of language change. Section 5.2 describes the use of probabilistic tools in establishing language relatedness through vocabulary comparison, an important task when historical and textual records are lacking and inferences about language change must be drawn from the ways in which related languages differ. Section 5.3 examines how the frequencies of linguistic traits change over time in the historical record, and how the timecourse of a change can shed light on its motivation and on the continuation problem. Section 5.4 discusses the role that the frequencies of lexical items and constructions play in their susceptibility to change, and what this can tell us about the synchronic effects of frequency. Section 5.5 asks how language change is directly molded by probabilistic behavior on the part of its participants—speakers, hearers, and learners.

5.2 Probability as a Tool for Investigating Language Relatedness

An important task in historical linguistics is establishing which linguistic changes are possible or probable (the "constraints problem" of Weinreich, Labov, and Herzog 1968). In many cases, we can look to synchronic variation to tell us which changes are in progress in a particular language (see Labov 1994). In rare cases, we have written records of change within a language. But the vast majority of language changes that have taken place in human history have left no trace either in synchronic variation or in the written record. The only way to discover them is through comparison of related languages: if we can reconstruct a protophoneme *p, for example, that became b in some context in a daughter language, then we know that the change from p to b in that context is a possible one; if we find many such cases, then we know that the change is a common one. Moreover, once we have established by reconstruction that a change took place, we can use synchronic evidence to answer questions such as how regular the change was and which types of exceptions were allowed to persist.

But how are we to know if two languages are related in the first place, so that an attempt at reconstruction makes any sense? A common method

for establishing the relatedness of languages is to compare their vocabularies. A list of words is collected for each language, based on a standard list of 100 or 200 meanings (e.g., the lists proposed in Swadesh 1952, 1955) that are expected to have a name in every language. If the languages are related, we expect to find either similarities between the sounds of words with identical or similar meanings (e.g., if the word for meaning *i* begins with a labial consonant in language *A*, then the word for meaning *i* begins with a labial consonant in language *B* too) or consistent correspondences between them (e.g., wherever we see a *t* in language *A*, there is a *k* in the corresponding word of language *B*). Because reflexes of a protophoneme can differ considerably in daughter languages, consistent correspondences are a more appropriate criterion for languages that are not known to be closely related.[1] The more frequent and consistent the correspondences are, the more likely it is that the two languages are connected, whether through descent from a common ancestor or perhaps through borrowing.[2]

The mathematical challenge in using this method is, how sure can we be that the similarities or correspondences found are not merely due to chance? It turns out that the degree of similarity or correspondence necessary to establish relatedness is greater than we might intuit. Ringe (1992) gives a detailed and highly accessible demonstration of this fact using randomly generated and real word-lists. Ringe's method is flawed, as discussed below, but he makes an important point: even though a particular event may be very unlikely to occur by chance, it may be an instantiation of a larger class of events one or more of which is relatively likely to occur.

Suppose, for example, that we hypothesize that two languages are related, and our criterion for relatedness is similarity (rather than regular correspondence). If we find that the word for *eye* begins with *t* in both languages, that is a piece of evidence in favor of the hypothesis, but how striking is it? How likely is it to have occurred by chance if the two languages were not related? The probability that language *A*'s and language *B*'s words for *eye* should both begin with *t* by chance is equal to the proportion of words in language *A* that begin with *t* (A_t) times the proportion of words in language *B* that begin with *t* (B_t). If, in each language, only 5% of words begin with *t*, then $A_tB_t = .0025$, a low probability. But this is a misleading result: the hypothesis being tested is not that both languages' words for *eye* begin with *t*, but that the languages' vocabularies are similar. The probability we should be interested in is the

probability that at least one pair of words on the list would begin with the same sound by chance; this probability will depend on the phoneme distributions in the two languages, but will be much, much higher than .0025. For example, if each language has the same 20 phonemes, each occurring word-initially 5 times in a list of 100 meanings, the chance of obtaining at least one match is nearly 100%.

Because this issue has caused so much confusion in the literature (see Manaster Ramer and Hitchcock 1996 for an attempt to sort out one exchange), it is worth belaboring. Manaster Ramer and Hitchcock call the confusion of a specific event with the class to which it belongs the "birthday fallacy": the chance that two randomly chosen people share the birthday of February 1 is small (1 in $365^2 = 133{,}225$), but the chance that they merely share the same birthday is much greater (1 in 365).[3] Choosing a specific date when calculating the probability of a shared birthday is analogous to requiring a correspondence to involve a particular sound or pair of sounds, or to occur in a particular word.

The same issue arises when we seek correspondences across multiple languages, as suggested by Greenberg and colleagues (Greenberg 1987; Greenberg and Ruhlen 1992). A correspondence seen in any two languages out of a group of, say, 15 languages is not as significant as a correspondence seen in a comparison between just two languages, because there are $\binom{15}{2} = 105$ pairs in which such a correspondence could have occurred, rather than just 1.[4] As Baxter and Manaster Ramer (1996) argue, it should be possible to determine the number of matches across a set of n languages that would be as significant as a match in a two-language comparison, but the determination becomes much more complicated when, for example, each language participating in the correspondence is required to be from a different family (Manaster Ramer and Hitchcock 1996).

Considering just the simpler case of a comparison between two languages, what we would like to do is determine how different a *contingency table* is from what would be expected by chance. The contingency table 5.1 represents how often each word-initial consonant (or $\varnothing$ for vowel-initial words) in a list of 100 English words corresponds to each word-initial consonant in the German word with the same meaning. For example, there are 4 words that begin with *w* in English and with *v* (orthographic *w*) in German. We could construct similar tables for any

Table 5.1
Observed values for initial-consonant correspondences in English and German. (From Ringe 1992, 22–23.)

		German																	
		f	∅	h	b	v	ʃ	k	z	r	l	n	g	m	t	ts	d	pf	Sum
English	s	0	0	1	0	0	5	1	6	1	0	0	0	0	0	0	0	0	14
	b	1	0	0	5	0	1	1	0	1	0	0	1	0	0	0	0	0	10
	h	0	0	6	0	1	0	1	0	0	0	0	0	1	0	0	0	0	9
	∅	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
	n	0	0	1	0	1	0	1	0	0	0	5	0	0	0	0	0	0	8
	f	8	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	12
	w	1	1	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	7
	l	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	5
	m	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	5
	t	0	0	0	1	0	1	0	0	0	0	0	0	0	0	3	0	0	5
	k	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	4
	r	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	4
	d	0	0	1	0	0	1	0	0	0	0	0	0	0	2	0	0	0	4
	g	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	3
	j	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2
	ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	Total	11	9	9	8	8	8	7	7	5	9	5	5	4	2	3	2	1	103

other type of correspondence we were interested in, such as medial consonants or consonant-vowel sequences.

If the two languages were not related, we would expect to see, on average, the values shown in table 5.2, which preserves the row and column totals of table 5.1 but eliminates any row-column interaction. In a given case, the numbers will differ from those in table 5.2 (minimally, they must be integers), so our question is, how unusual is it for a chance-generated table to deviate from table 5.2 as strongly as table 5.1 does?

Ringe proposes calculating the probability, for each cell, that the number of observed matches or more would be seen by chance, by summing binomials. That is, he proposes that the probability of finding exactly one match of x in language A with y in language B in a list of 100 words is the product of three numbers: A_xB_y (the probability that a particular pair shows the correspondence), $(1 - A_xB_y)^{99}$ (the probability that the 99 other pairs do not), and 100 (the number of places in the list where a matching pair could be found). Similarly, the probability of finding exactly two such pairs would be $A_xB_y{}^2 \cdot (1 - A_xB_y)^{98} \cdot 9{,}900$. To calculate the probability of finding n or more matches, we would sum the probabilities of finding n through 100 matches:

$$\sum_{i=n}^{100} (A_xB_y)^i \cdot (1 - A_xB_y)^{100-i} \cdot \binom{100}{i}. \tag{1}$$

For $A_x = B_y = .05$ and $n = 3$, this sum is .20—in other words, at least one correspondence between x and y is a fairly likely event.

The problem with Ringe's method, as pointed out by Baxter and Manaster Ramer (1996), is that it wrongly assumes that the probability of seeing a correspondence in one word-pair is independent of whether the same correspondence occurs in another pair. A_x and B_y are based on frequencies within the chosen word-list. Suppose that $A_t = B_t = .05$: there are five instances of initial *t* in each language's word-list. If the words for *eye* begin with *t* in both languages, then the chance that the words for *cheek* will also both begin with *t* is lowered, because there is one fewer *t* left in the pool from which *cheek* can draw its initial consonant. The probability that the words for *cheek* would begin with *t* in both languages is now not $(5/100) \cdot (5/100) = .0025$, but $(4/99) \cdot (4/99) = .0016$. The values to be summed are not binomials as shown in (1), but hypergeometrics, which are unwieldy for numbers as high as 100.

Table 5.2
Expected values for initial-consonant correspondences in English and German

		German																	
		f	∅	h	b	v	ʃ	k	z	r	l	n	g	m	t	ts	d	pf	Sum
English	s	1.5	1.2	1.2	1.1	1.1	1.1	1.0	1.0	0.7	1.2	0.7	0.7	0.5	0.3	0.4	0.3	0.1	14.1
	b	1.1	0.9	0.9	0.8	0.8	0.8	0.7	0.7	0.5	0.9	0.5	0.5	0.4	0.2	0.3	0.2	0.1	10.3
	h	1.0	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.4	0.8	0.4	0.4	0.4	0.2	0.3	0.2	0.1	9.1
	∅	0.9	0.7	0.7	0.6	0.6	0.6	0.5	0.5	0.4	0.7	0.4	0.4	0.3	0.2	0.2	0.2	0.1	8.0
	n	0.9	0.7	0.7	0.6	0.6	0.6	0.5	0.5	0.4	0.7	0.4	0.4	0.3	0.2	0.2	0.2	0.1	8.0
	f	1.3	1.0	1.0	0.9	0.9	0.9	0.8	0.8	0.6	1.0	0.6	0.6	0.5	0.2	0.4	0.2	0.1	11.8
	w	0.7	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.3	0.6	0.3	0.3	0.3	0.1	0.2	0.1	0.1	6.7
	l	0.5	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.6
	m	0.5	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.6
	t	0.5	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.6
	k	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.2
	r	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.2
	d	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.2
	g	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.0	2.9
	j	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.0	2.1
	ð	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.0	2.1
	p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
	Sum	10.9	9.1	9.1	8.0	8.0	8.0	6.9	6.9	4.8	9.1	4.8	4.8	4.2	2.1	2.9	2.1	0.7	102.4

How, then, can we accurately determine whether a table like table 5.1 is significantly different from the average table we would expect to see if the two languages were not at all related? Statistics such as χ^2 give a measure of how different an observed contingency table is from the expected table,[5] but in order to determine the significance of that difference—how likely it would be to arise by chance—we must rely on lookup tables that are inappropriate to the task. Lookup tables for the distribution of χ^2, for example, assume that the data points are independent and that expected cell values are relatively high (expected frequencies of at least five in each cell)—much too high for a table with dozens of cells and only 100 instances to go around.

Kessler (2001) proposes an ingenious solution to the inapplicability of standard lookup tables, similar in spirit to Oswalt's (1970) shift test,[6] but much more robust. We want to know the distribution of values of χ^2, or some other measure of skewedness, if languages A and B are not related, so that we can see how unusual the observed value of χ^2 is. If A and B are not at all related, and if we have excluded from the word-list words subject to sound symbolism and onomatopoeia, then any lineup of A's words with B's should be equally likely—the particular lineup that occurs in reality is the result of mere chance. Thus, the universe of possible arrangements that should occur by chance, while preserving the individual phoneme distribution of each language, is well represented by keeping the order of A's list constant and permuting B's list in all possible ways. If we calculate χ^2 for each such permutation, we obtain the distribution of χ^2. We can compare the value of χ^2 obtained for the actual word-list to this distribution: if it is larger than 99% of the χ^2 values in the distribution, then we know that a contingency table as skewed as the one we obtained will occur only 1% of the time if the two languages are unrelated.

In practice, however, we cannot consider all permutations of B's list. For a list of 100 words, there are astronomically many permutations: $100! \approx 9.3 \times 10^{157}$. This is too many to consider, even for a computer. Kessler's solution is to instead randomly generate some large number of permutations to get a close estimate of how often the resulting χ^2 values are greater or smaller than the observed one. The more permutations sampled, the more accurate the count; Kessler uses 10,000 permutations. The same method can be used for any other measure: Kessler considers R^2, the sum of the square of each cell entry (minus one if nonzero); various breakdowns by phonetic feature; and matching phoneme sequences

rather than individual phonemes. For Kessler's example data, R^2 seems to work best.

The problem of determining whether a language resemblance is stronger than would be expected by chance is a tractable one, then, at least in simple cases such as correspondences between phonemes. As all the authors cited here agree, however, establishing relatedness is only a starting point. These statistical methods do not replace the work of establishing which words are cognates, determining the contextual determinants of sound changes that lead to inexact correspondences, or reconstructing protoforms. They do, however, give us a tool with which to determine how striking an apparently striking connection really is, so that we can decide whether an attempt at reconstruction is warranted.

5.3 Changes in Probabilities over Time

Language change appears to take place gradually, with innovations being used at different rates in different parts of the speech community and in different linguistic or social contexts, and with an innovation's overall rate of use rising gradually, often over centuries (though see discussion of Shi 1989 below). Changes in observed probabilities in the historical record can give evidence for the nature of the linguistic system underlying variable linguistic behavior, the nature and proximal cause of a particular change, and the way in which changes take hold and spread.

5.3.1 Correlations in Rate of Change

Suppose that a language is observed to undergo a gradual change from an SOV (subject-object-verb) word order to an SVO order; that in texts from intermediate stages, the innovative order is found more frequently in main clauses than in subordinate clauses; and that in the intermediate stages, variation is observed even within each individual writer. How should the linguistic system of an individual living during the middle stages be represented? If it is a grammar that encodes separately the probabilities of employing SOV or SVO in various contexts, then the innovative word order may spread at quite unrelated rates in main and subordinate clauses. If, however, the difference between SOV and SVO is controlled by a single parameter in the grammar—whose setting can be probabilistic to allow variation—and it is some orthogonal force (stylistic, perhaps) that prefers SOV in subordinate clauses, then although the frequency of use of the innovative order may differ according to clausal

context, the rates of change of those contextual frequencies should be the same, assuming that orthogonal forces remain constant. This is the constant rate hypothesis, proposed by Kroch (1989): because changes occur at the level of abstract grammatical parameters, they spread at the same rate in every context, although the base frequencies of use in each context may differ for external reasons.

Kroch and colleagues have tested the constant rate hypothesis by modeling S-shaped language changes with a logistic function. It has long been observed (e.g., Osgood and Sebeok 1954; Weinreich, Labov, and Herzog 1968; Bailey 1973) that language change takes an S-shaped course: a new variant appears rarely for a long time, then quickly increases in frequency; finally, the rate of change slows as the frequency approaches its maximum (100% in the case of a total replacement of the earlier form). There are several mathematical functions that produce an S-like shape. Kroch chooses the logistic function because every logistic function has associated with it a slope, and therefore the slopes of frequency changes that should be linked, according to the constant rate hypothesis, can be compared.

The logistic function takes the form in (2),[7] where P, interpreted here as the probability of seeing some variant in some context that it could potentially occupy, is a function of t, time:

$$P = \frac{1}{1 + e^{-k-st}}. \tag{2}$$

Simple algebra transforms (2) into (3), where now the logistic transform, or logit, $(\ln(P/(1-P)))$, is a linear function of t, with a slope (steepness) s and an intercept (initial value) k:[8]

$$\ln \frac{P}{1-P} = k + st. \tag{3}$$

When frequency changes over time are plotted for the same innovation in different contexts, the logit for each context should have approximately the same slope under the constant rate hypothesis, although they may have different intercepts. Figure 5.1 illustrates two logistic functions whose logits have the same slope, but different intercepts, and one that has a different slope.

The constant rate hypothesis can also be tested using the multivariate analysis performed by the VARBRUL program (see Mendoza-Denton, Hay, and Jannedy, this volume). VARBRUL represents the logit as the

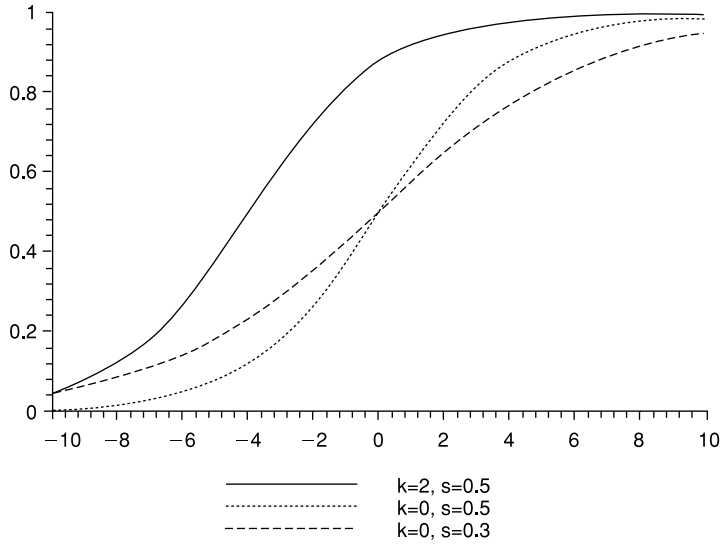

Figure 5.1
Three logistic functions. The solid and dotted lines have the same logit slope (0.5), but different logit intercepts (2 and 0, respectively); the dashed line has the same logit intercept as the dotted line (0), but a different logit slope (0.3).

sum of some contextual weights, representing the positive and negative effects of various features of the context, plus a base rate of use, in this case a linear function f of time:

$$\ln \frac{P}{1-P} = f(t) + a_1 + a_2 + a_3 + \cdots \quad \text{(Kroch 1989, 6)} \tag{4}$$

If the values of the a_j do not change as t changes, then the contribution of the context is constant over time, and only the base rate of use of the innovative variant changes.

Kroch and colleagues have found evidence for the constant rate hypothesis in several cases of language change. Kroch (1989) illustrates how the results of Noble (1985), Oliveira e Silva (1982), and Fontaine (1985) support the constant rate hypothesis in the replacement of possessive *have* by *have got* in British English, the rise of the definite article in Portuguese possessive noun phrases, and the loss of verb-second in French, respectively. Kroch (1989) also reanalyzes Ellegård's (1953) data on the rise of periphrastic *do* in English. Pintzuk (1995) has found that Old English I′-initial Infl rose in frequency at the same rate in main and

subordinate clauses, and Santorini (1993) reports that the rise of a similar I′-initial Infl phenomenon in early Yiddish proceeded at the same rate with both simple and complex verbs, although in this case the intercepts of the logits are also similar, so we cannot be sure that the simple and complex verbs represent two truly different contexts.

These findings suggest that syntactic and morphosyntactic changes do indeed occur at some abstract level of the grammar, affecting all contexts equally, and subject only to independent influences on various contexts. Tabor (1994), using a very different model of grammar, essentially agrees, but views constant rate effects as a special case of frequency linkage effects—related changes proceeding at related, though not necessarily identical, rates.

Tabor's model of (morphosyntactic) language change uses a connectionist network to learn associations between words and the contexts in which they tend to occur and among words that tend to occur in similar contexts. Words, represented by input nodes, are connected to intermediate hidden-layer nodes; words that are strongly associated to the same hidden-layer nodes act as clusters somewhat like traditional grammatical categories (e.g., Noun, Verb), although cluster membership is a gradient property, so that a single word may belong to different clusters to different degrees. Hidden-layer nodes are connected to each other, to represent sequential information, and to output nodes, representing behaviors in various syntactic constructions. How strongly a word is associated to some syntactic behavior is therefore mediated by the hidden units and thereby by the behavior of cluster-mates.

If a network that has been trained on a corpus is exposed to an altered version of the original training data (representing an externally motivated shift in frequency), it adjusts its connection weights in response, but not only those aspects of the language that changed in the training data will be affected: aspects of the language that were strongly linked to the changed aspects will be affected also. In particular, if the frequency with which some word occurs in some context changes in the training data, the network will adjust the word's association strengths with the hidden units in response, thereby altering the word's indirect association to other words; as a consequence, other aspects of the word's behavior will also change, under the influence of the word's new cluster-mates.[9] At the same time, the network must adjust associations between the word's strongly associated hidden units and the output units, so that the behavior of other words that were strongly associated to the same hidden units will change

too. This is where frequency linkage effects come from in Tabor's model: a change in one aspect of the language drags along with it changes in other aspects of the language.

Frequency linkage of the constant-rate variety will be observed when two words or constructions have the same distribution (or nearly so) before the change begins. Tabor demonstrates with an abstract example: two nouns, N1 and N2, behave similarly along five binary contextual dimensions, C1 through C5 (e.g., if C1 is what possessive verb the nouns appear as the object of, they might appear as the object of *have* 96% of the time and as the object of *have got* 4% of the time). A third noun, N3, behaves like N1 and N2 along dimension C1, but shows different frequencies on the other four dimensions. A network is trained on a corpus with these properties, then retrained on a corpus that is the same except that N2 undergoes a frequency change in C1, from choosing option A 4% of the time to choosing it 100% of the time; no examples of N1 and N3 are given for C1 in the new corpus. The point of the experiment is to observe how the frequencies for N1's and N3's choosing option A in C1 change as the network approaches the desired frequency for N2's choosing option A in C1. The slope of the logit for how often N1 exhibits option A in C1 is almost identical to the slope of the logit for N2, but N3's slope is much shallower. Because N3 does not share N2's properties as well as N1 does, N3 is not "dragged along" as much as N1 is. Thus, for Tabor, constancy of rate is gradient; he would predict that SVO would spread at the same rate in main and subordinate clauses to the extent that the behavior of main and subordinate clauses is otherwise similar.

It remains to be seen whether any convincing cases of demonstrably partial frequency linkage exist. Tabor argues that the rise of English periphrastic *do* is such a case, but Kroch (1989) proposes that certain syntactic assumptions can explain why the slopes for some of the contexts for *do* are unequal. If clear cases can be found, then we have evidence that language change is indeed abstract, occurring at the level of structures and categories, but that structures and categories can be fuzzy and membership in them gradient.

5.3.2 Reanalysis and Frequency Change

Besides bearing on the abstractness of language change, rates of use over time can shed light on the relationship between reanalysis and frequency. It seems clear in many cases of morphological, syntactic, and semantic

change that some word or construction has been reanalyzed; that is, its behavior has changed in a radical way, indicating that it has joined a different grammatical category. For example, *be going to*, which once obligatorily indicated motion toward, is now used as an all-purpose future marker in English.

How and why does reanalysis occur? In the (otherwise very different) models of Lightfoot (1991) and Tabor (1994), reanalysis results from frequency shifts that encourage or even force learners to assign a new structural analysis to a form because of the contexts in which it appears. Others argue that reanalysis is a prerequisite for syntactic change: only after two structural options become available can one rise in frequency. Santorini (1993) and Pintzuk (1995), for example, argue that in Yiddish and English, respectively, the availability of an I′-initial position for Infl in both main and subordinate clauses occurs at the beginning of the rise in frequency of medial Infl, not at the end (the alternative analysis is that in the early stages, clause-medial Infl results from movement, and not until later is Infl reanalyzed as potentially I′-initial). In these two cases, the argument for the availability of I′-initial Infl at the early stages is mainly a syntactic one, but it also has a probabilistic element. Surface nonfinal Infl could be the result of base-generated I′-initial Infl or base-generated I′-final Infl, with rightward movement of other constituents. Santorini and Pintzuk both argue that the rate of such rightward movements observed in unambiguous contexts is too low to account for the relatively high rate of nonfinal Infl. Therefore, I′-initial Infl must have been used at least some of the time at the early stages of the change, before it became very frequent.

Frisch (1994) presents another case in which evidence that reanalysis precedes syntactic change comes from frequencies over time. In Middle English, *not* acted like a sentence-level adverb: it could appear preverbally or postverbally, much like modern *never*; it carried an emphatic meaning; and it alone was not sufficient to indicate negation (*ne* was instead the usual marker of nonemphatic negation).

(5) Þat Jesuss *nohht* ne wollde Ben boren nowwhar i þe land, . . .
that Jesus not NEG would be born nowhere in the land
'That Jesus did not (at all) want to be born anywhere in the land, . . .'
(Frisch 1994, 189; Ormulum I: 122)

It is standardly proposed (Kroch 1989; Pollock 1989; Shanklin 1990; Roberts 1993) that *not* was reanalyzed as a sentential negator, losing its

preverbal position and emphatic meaning, because the phonological loss of the clitic *ne* eventually forced *not* to be so interpreted. Frisch demonstrates, however, that the loss of preverbal *not* was well underway before the loss of *ne* began.[10] Frisch argues, therefore, that a semantic reanalysis of *not* as nonemphatic, allowing it to occupy the specifier position of NegP rather than a sentence-level adverb position, caused *ne* to become redundant and be lost. With *ne* gone, *not* was free to occupy either the specifier or the head of NegP.

An important assumption is that the rate of adverbial use of *not* in ambiguous cases can be extrapolated from the behavior of the unambiguous sentence adverb *never*. *Never* is preverbal 16% of the time, and during the first 70 years of Middle English, *not* has the same distribution. Frisch presents the following formulas:

$$\text{number of preverbal } not = 0.16 \times \text{total number of adverbial } not, \quad (6)$$

$$\text{total number of adverbial } not = \text{number of preverbal } not/0.16. \quad (7)$$

Assuming that the rate at which true sentence-level adverbs appear preverbally is constant at 16% throughout the period, Frisch obtains an estimate of how often *not* is used adverbially from 1150 to 1500. The key finding is that this percentage falls drastically before the percentage of negative sentences containing *ne* begins to drop much at all.

We have, then, cases in which reanalysis appears to occur at the beginning of a frequency shift, rather than at the end. Does this contradict Tabor's claim that frequency shift leads to gradient reanalysis, in which a word begins to belong more and more to a different cluster, gradually taking on properties of that cluster? Perhaps not: in Tabor's model, reanalysis and frequency shifts can be gradual and mutually reinforcing. If Tabor's model were extended to include semantics, an increasing use of *not* in nonemphatic contexts (a typical case of semantic bleaching) could cause *not* to be gradually reanalyzed as a nonemphatic negator. The more strongly it was so recategorized, the less often it would appear preverbally. Reanalysis would thus follow one frequency shift, and precipitate another: frequency shifts affect probabilistic learners and in turn are affected by probabilistic speakers.

5.3.3 The Timecourse of Language Change

As mentioned above, it has long been observed that language change proceeds along an S-shaped curve. Why should change begin and end slowly? If changes spread from speaker to speaker, the rate of spreading

depends on the number of interactions between a speaker who has the new variant and one who has the old variant. There will be few such exchanges at first, because there are few speakers who have the new variant, and few such exchanges at the end, because there are few remaining speakers to whom the change has not yet spread (Bloomfield 1933). When there is variation within individuals, as there is in nearly all studies of historical texts, the picture is more complicated, because there are no speakers with 100% use of the new variant at first. We must assume that speakers can slightly increment their use of a variant and that some force (such as group identification or learnability) encourages the change to continue in one direction. The remainder of this section discusses some attempts to derive S-shaped language change mathematically, with limited success.

But first, a cautionary note, based on Shi's (1989) findings. Shi argues that a gradual, S-shaped change that appears to have taken place over 1,000 years is actually an abrupt change that was completed in at most 200 years. The illusion of gradualness comes from the persistence of classical style in modern texts. Shi tracks the rise of the aspectual particle *le* in Mandarin, which derives from the classical verb *liao* 'finish'. When the number of uses of *le* per 1,000 characters is tracked for a corpus from the pre-tenth to the twentieth century, the rate of use rises slowly from the tenth to the twelfth century, then rises quickly until the seventeenth century, and continues to rise slowly (though unevenly) to the present.

Shi finds, however, that *le* seems to be inhibited by classical verbs and hypothesizes that avoidance of *le* by more recent writers is merely an attempt to emulate classical style. Shi uses occurrences of the sentence-final copula or interjective *ye* as an index of classicalness. Classical texts have approximately 8 occurrences of *ye* per 1,000 characters, so if there are n occurrences of *ye* per 1,000 characters in a text, there are approximately $n/8$ classical characters per actual character in the text; the rest can be considered vernacular. When the number of *le*s per 1,000 vernacular characters is plotted, the picture is very different from when raw character count was used: there is now a sharp rise in use of *le* from the tenth to the twelfth century, and the rate of *le* use has not risen since. Shi's study points out an important potentially distorting effect of the unavoidable use of written records: even when a change is abrupt, the conservatism of written styles may cause it to appear gradual.

Assuming, however, that the S-shaped model is accurate (though it may appear artificially stretched in the written record), are there any models that can derive it? Manning (this volume), points out that sto-

chastic Optimality Theory (Boersma 1998; Boersma and Hayes 2001) predicts S-shaped change if one constraint rises or falls at a constant rate through the grammar. In stochastic Optimality Theory, surface forms are chosen according to their satisfaction of constraints whose rankings are normally distributed. Change is therefore slow when constraints' distributions overlap only at the outer edges, accelerates as the centers of the bell curves begin to overlap, and slows as the distributions again overlap only at the edges. The mechanism by which the grammar is transmitted from generation to generation in such a way that a change in ranking is persistent and linear is not known, however. The following paragraphs review some attempts at achieving S-shaped change through modeling transmission of the grammar from adults to children over time.

Niyogi and Berwick (1995) present an abstract simulation of language change that does derive a logistic function for change, among other possibilities. In Niyogi and Berwick's model, different members of the population use different grammars, and learners must decide which grammar to adopt. (Admittedly, this is an unrealistic assumption, as it predicts no variation within individuals.) Each grammar is a series of n binary parameters, and the distribution of sentences produced by each grammar is uniform (all well-formed sentences are equally likely). Learners set parameters on the basis of examples, permanently and without tracking probabilities. Because the learner has limited opportunity to adjust its grammar, mislearning is likely, especially if many utterances are ambiguous, making even homogeneous populations potentially unstable.

Learning proceeds as follows in Niyogi and Berwick's model. The learner draws at random two utterances by members of the surrounding population. If the second trigger utterance unambiguously supports one parameter setting, the learner chooses that setting. If only the first trigger is unambiguous, the learner chooses that setting. And if both triggers are ambiguous, the learner makes an unbiased choice at random. In other words, the critical period is just two utterances, and if they conflict, the more recent utterance prevails.

Niyogi and Berwick investigate by simulation the case of three parameters governing constituent order (yielding eight possible grammars) and find that the distribution of grammars sometimes changes according to a logistic function (S-shaped curve) that varies in steepness. But with some starting distributions and maturation times, the function is not logistic: rapid change can occur right away (the initial tail of the S is cut off), or the function may fall off toward the end rather than continuing to approach an asymptote.

Niyogi and Berwick apply their model to the change from Old French verb-second (V2) to Modern French SVO, using the five binary parameters suggested by Clark and Roberts (1993) to yield 32 possible grammars. If learning time is limited, so that the younger generation does not have full opportunity to acquire the older generation's grammar, then in simulations even a population that begins homogeneously V2 shifts away from V2, though the change is slow and does not proceed very far. But when even small numbers of SVO speakers are included in the initial population (perhaps representing foreign speakers), there is relatively rapid loss of V2.[11]

Niyogi and Berwick's model is deterministic if the population of agents is infinite (and generations do not overlap). Extending the investigation to cases in which the population is finite and small, Briscoe (2000) finds that the results are quite different. For example, if two competing grammars are initially equally distributed and produce equal proportions of ambiguous sentences, in the infinite-population model the two grammars should remain in balance: half the learners will adopt one, and half will adopt the other. In a finite population, however, the probability that exactly half the learners will adopt one grammar on any given trial is low (just as the probability is low that exactly half of a finite number of coin tosses will come up heads). Therefore, one grammar will probably gain ground over the other. As one grammar becomes much more common than the other, however, it becomes less and less likely that it can maintain its advantage. At the extreme, if a grammar is used by 100% of the population, as long as there are some unambiguous sentences, some learners will learn the other grammar. Even a moderate bias such as 75%–25% is untenable if there is a high proportion of ambiguous sentences: if 75% of the population uses grammar *A*, with 50% of sentences from each grammar being ambiguous, and there are 100 learners, the probability that 75 or more of the learners will adopt *A* is only .07. Grammar *A* begins to lose ground, then, falling toward 50%, which we have already seen is itself an unstable state. The proportions of the two grammars will therefore oscillate endlessly.

Clearly, a realistic and complete model of how changes spread remains to be implemented.

5.4 The Role of Frequency in Language Change

We have seen the importance of changes in frequency over time. The individual frequencies of linguistic items also appear to play an important

role in language change. Words' frequencies affect their susceptibility to phonological, morphological, and morphosyntactic change. This fact reinforces the findings elsewhere in this book that not all lexical items are treated alike and that the strength of lexical entries is gradient. These differing strength values are important in the lexical (word-to-word) spread of linguistic innovations.

5.4.1 Frequency and Phonological Erosion

Bybee (1994; see also Bybee 2001) proposes that a usage-based model of phonology can account for two relationships between word frequency and phonological change: frequent lexical items are the first to adopt automatic, phonetic rules, and the last to abandon nonphonetic rules. By "phonetic rules" Bybee means rules, like American English flapping, that involve minimal articulatory or acoustic change. Nonphonetic rules include morphologically conditioned rules, like stress in Spanish verbs or English noun-verb pairs, and lexical generalizations, like the English *sing-sang-sung* and *ring-rang-rung* patterns.

An important assumption for Bybee in explaining the effect of frequency on susceptibility to phonetic changes is that lexical representations do not include only the idiosyncratic aspects of a word. Redundancies and phonetic detail are also included, so that different words may be reliably associated with slightly different patterns of articulatory timing and other subphonemic properties.

Phonetic rules tend to spread gradually through the lexicon, affecting frequent words to a greater extent. For example, in Hooper 1976, Bybee found that medial schwa deletion was most advanced in frequent words like *every* (it is nearly obligatory) and less advanced in less frequent words like *artillery* (it is nearly forbidden). In Bybee's usage-based model, this is because lexical entries are updated by speakers and/or listeners every time they are used. If schwa deletion has some probability of applying every time a word is used, then there is a related probability that the word's lexical entry will be updated to reflect the change. Because there is no reverse rule of "schwa restoration," once the strength of the schwa in a lexical entry is reduced, it cannot later increase; it can only stay where it is or reduce further. The more often a word is used, the more chances it has to drift irreversibly toward schwa deletion. Thus, highly frequent words are the innovators in phonetic change.

Pierrehumbert (2001a; see also this volume), in developing an exemplar-theory-based model of production, derives this finding quantitatively. In exemplar theory, categories are represented mentally as clouds

of remembered tokens (projected onto a similarity map) that are typically densest in the middle. Highly similar tokens are grouped into a single exemplar, whose strength is augmented when tokens are added to the group (and, countervailingly, decays over time). An incoming stimulus is classified according to the number of exemplars from each category that are similar to it, with a weighting in favor of stronger exemplars. Categories are produced by choosing an exemplar at random, but with a preference for stronger exemplars, and with some amount of noise added, so that the actual production may differ slightly from the exemplar chosen. Pierrehumbert shows that when exemplars are chosen in this way and the resulting tokens added to memory, the exemplar cloud gradually becomes more diffuse, but its center does not shift.

When a persistent bias (motivated by some external force) is added, however, drift does occur. If there is a tendency for productions to be slightly hypoarticulated with respect to the exemplar chosen for production (i.e., the articulatory gesture is reduced in magnitude), the center of the exemplar cloud gradually shifts toward hypoarticulation. For example, if an exemplar is chosen whose articulatory effort along some dimension is 0.9, it may be produced with 0.89 effort instead. The 0.89 token is then added as an exemplar, and if it is chosen in a later production, it may be pronounced with 0.88 effort, and so on.

The shift increases as the number of productions of the category increases. This means that if individual words have their own exemplar clouds, then words that are used more often shift more rapidly, as Bybee predicts. Pierrehumbert further shows how an infrequent category that is subject to lenition (or any other persistent bias) is absorbed into a frequent category that is not subject to lenition.

Bybee argues that frequent words are more subject to phonetic rules for an additional reason: phonetic rules tend to be lenition rules, involving reduced articulatory gestures. Frequent words are more likely to be used in prosodically unemphasized positions, which are associated with less articulatory effort. This is because a frequent word is likely to be used more than once in a discourse, and subsequent occurrences of a word in a discourse tend to be less emphasized prosodically than the first occurrence (Fowler and Housum 1987). In addition, frequent words or constructions are more likely to become semantically bleached (see Bybee, in press, discussed below) and thus less likely to be the carrier of important discourse information that is subject to prosodic emphasis.

Frequent words' lexical entries are thus doubly subject to a phonetic rule when that rule is lenitive: not only does the word's more frequent use

give it more opportunities to undergo the change, but the word's tendency to occur in repetitive or semantically bleached contexts disproportionately subjects it to lenition.

5.4.2 Frequency and Nonphonetic Rules

Highly frequent words are conservative, however, when it comes to nonphonetic rules like the English irregular past tenses (Hooper 1976) or English noun-verb stress shifts (Phillips 1998, 2001):[12] when the language begins to lose or gain a rule, they are the last words to change.[13] There are two reasons for this. The first reason is the competition between irregulars (residual archaic forms) and regulars (the innovative form). This competition proceeds differently in different models of regulars and irregulars, but in every case an irregular requires a strong lexical entry in order to resist regularizing. Under the dual-mechanism model of Pinker and Prince (1994), for example, listed irregular words and regular morphological rules compete in the brain: irregular, listed *sang* competes with regular, synthesized *sing+ed*. If the lexical entry of an irregular word is not strong enough, it may not be accessed in time or with enough certainty to win the competition, and the regular pronunciation will win. In a model such as Albright and Hayes's (2000) that encodes both regular and irregular patterns in the grammar, the competition is between very specific irregular rules and more general regular rules; in a connectionist model, it is between patterns in associative memory (Rumelhart and McClelland 1986a; Daugherty and Seidenberg 1994).

Frequent words' lexical entries are strong from frequent use and reinforcement and thus will tend to beat out synthesized, regular pronunciations, whether the pressure for those pronunciations comes from the grammar or from elsewhere in the lexicon. Infrequent words' lexical entries, on the other hand, may not be strong enough to win reliably.

The second, related reason for the retention of nonproductive rules in frequent words concerns transmission from one generation to the next. Infrequent irregulars may fail to be transmitted to the next generation—if a word is too infrequent, the child may never encounter it—and the younger generation will apply regular rules to the word. An abstract simulation performed by Kirby (2001) confirms that this mechanism can have the observed effect. Although the population in Kirby's simulation begins with no lexicon at all, as a lexicon begins to develop, it is only the most frequent words that are able to retain an irregular form; words that are too infrequent to be reliably transmitted fall under a regular compositional rule.

5.4.3 Frequency and the Undertransmission of Morphosyntax

The instability of infrequent irregulars is one type of "undertransmission." Richards (1997) has studied a more drastic type of undertransmission in the morphosyntactic realm that leads to a change not just in particular lexical items, but in the whole grammatical system. Richards compares the word order and verbal morphology of current speakers of Lardil, an Australian language, to data collected by Hale from speakers in the 1960s. Lardil is being replaced in everyday use by English, but Richards argues that the changes observed in Lardil are due not to the linguistic influence of English, but to the scarcity of Lardil data available to learners. (Richards's arguments rest on syntactic sensitivities of the changes that would not be expected if the language were merely adopting English morphosyntax.)

The morphological difference between "Old Lardil" and "New Lardil" that Richards discusses is the frequent absence of inflection on objects in New Lardil (the syntactic difference is the resulting rigidification of word order). Richards's explanation is that in Old Lardil, certain phonological rules could delete object suffixes. New Lardil learners exposed to these apparently unsuffixed forms, and not exposed to enough overtly suffixed forms to learn that suffix deletion is phonologically conditioned, might conclude that overt suffixes alternate freely with null suffixes.

In analyzing the behavior of particular nouns and pronouns, Richards finds that the pronoun on which inflection was most often produced had a highly irregular paradigm in Old Lardil. The pronoun on which inflection was least often produced had a more regular paradigm. Richards suggests that although regular morphophonological rules have been lost in New Lardil because of insufficient evidence, individual lexical entries exhibiting idiosyncratic inflection have been retained when frequent enough. Similarly, Richards finds that the regular morphophonological rules of verb augmentation have been lost, but that certain (presumably) irregular verbs forms of Old Lardil have been retained. High frequency, then, can allow a word to retain various idiosyncratic properties in the face of a more general language change.

5.4.4 Frequency and Grammaticalization

Frequency may also have an effect on which words or morphemes will undergo morphosyntactic change. Grammaticalization, the process by which content morphemes or morpheme sequences become function elements, tends to be correlated with an increase in frequency (see Traugott

and Heine 1991, Hopper and Traugott 1993, for overviews and many case studies of grammaticalization). Is this increase merely the result of grammaticalization, as the morpheme becomes needed in more contexts, or could it also be a cause?

Bybee (in press) argues that it can. Bybee traces the evolution of English *can* from a content word meaning 'have mental ability/knowledge' to a function word meaning 'possibility exists'. Following Haiman (1994), Bybee views grammaticalization as a form of ritualization, whereby repetition of a frequent act (in this case, the uttering of a word or construction) bleaches the act of its significance, reduces its (phonological) form, and allows it to become associated to a wider range of meanings. Bybee shows how *cunnan*, the ancestor of *can*, which first took only noun phrase objects, began to take as its object the infinitives of verbs relating to intellectual states and activities, communication, and skills. Bybee argues that because *cunnan* with a noun phrase object was already common, additional mental verbs began to be added to "bolster the meaning," creating seemingly redundant expressions like *cunnan ongitan* 'know how to understand'. This use of *cunnan* further weakened it semantically: presumably, a learner who encounters the phrase *cunnan ongitan* is likely to attribute all the meaning to *ongitan* and treat *cunnan* as merely grammatical.

The token frequency of *can* increased greatly from Old to Middle English, partly as *can* came to be used with a larger number of verbs, partly as some of the *can+VERB* combinations became more frequent. The increase in both type and token frequency, Bybee argues, further bleached *can* semantically, and its verbal objects expanded to include emotional states, nonmental states, verbs that take as object another person, verbs indicating an action (rather than merely a skill). A few instances of inanimate subjects also began to occur. Eventually, as the 'possibility' meaning became more common, the use of inanimate subjects increased. Thus, increasing frequency and semantic bleaching reinforce each other.

Bybee further notes (citing crosslinguistic findings in Bybee, Perkins, and Pagliuca 1991, 1994) that grammaticalized morphemes tend to be shorter and more phonologically fused with surrounding material, for reasons discussed above: frequent morphemes (including grammatical morphemes) are more susceptible to erosive lenition rules, which can cause loss and overlap of gestures. Bybee proposes that the units of lexical storage are not only morphemes or words, but also highly frequent phrases or sequences. When grammatical morphemes enter into

high-frequency sequences such as *going to*, those sequences too are subject to erosion (*gonna*). As these sequences gain their own lexical representations, they can also develop idiosyncratic meanings and syntactic functions.

Tabor's (1994) connectionist model, described above, similarly views frequency as a driver of syntactic change. Tabor focuses not on the overall type or token frequency of a lexical item, but on the frequency with which it occurs in a particular context. Tabor performed a series of experiments, simulating real changes that occurred in English, in which a network was trained on a corpus, then trained on a frequency-altered version of that corpus, and a word or sequence of words consequently changed its categorical affiliation, exhibiting new behaviors that were previously ungrammatical (i.e., below some probability threshold). The cases simulated include the rise of periphrastic *do*, the development of *sort of/kind of* as a degree modifier, and the development of *be going to* as a future auxiliary.

Tabor, like Bybee, argues that the changes in frequency that often precede a reanalysis can be the cause of the reanalysis: the more a word appears in the same context as words from some other category, the more it is pushed to take on the characteristics of that category. For example, in the *sort of/kind of* case, sentences like (8) would have been parsed only as (8a) until the nineteenth century ('It was a type of dense rock'), but can currently also be parsed as (8b) ('It was a somewhat dense rock'). Tabor argues that a high frequency for sentences like (8)—where *sort of/kind of* is followed by an adjective+noun and therefore appears in a position that the degree modifies *quite* or *rather* also can appear in—caused *sort of/kind of* to become affiliated with the degree modifiers and therefore become able to appear in unambiguously degree-modifying contexts, like (9).

(8) It was a sort/kind of dense rock. (Tabor 1994, 137)
  a. It was $[_{NP}$ a $[_{N'}[_{N}$ sort/kind$]$ $[_{PP}$ of $[_{NP}$ dense rock$]]]]$.
  b. It was $[_{NP}$ a $[_{N'}[_{AdjP}[_{DegMod}$ sort/kind of$]$ $[_{Adj}$ dense$]]$ rock$]]$.

(9) We are sort/kind of hungry. (Tabor 1994, 137)

Tabor finds a sharp rise in the late eighteenth century in how often *sort of/kind of* is followed by an adjective (crucially, preceding the rise of sentences like (9)). The simulation showed that increasing the frequency of ⟨a sort/kind of Adj N⟩ noun phrases does lead unambiguously degree-modified utterances like (9) to rise above the threshold of grammati-

cality (i.e., become more frequent than a near-grammatical control sentence type that is never attested in the corpus and does not become more likely over the course of the training) and continue to rise in frequency. Thus, again we see that reanalysis and frequency change are mutually reinforcing.

5.5 Language Agents in a Probabilistic Environment

Speaker-hearer interactions, whether involving adults, children, or a combination, are the atoms of language change. What we call a language change is not a single event, but a high-level description of millions of individual interactions over time, with early interactions influencing later ones. If a participant, child or adult, comes away from an interaction with her grammar or lexicon slightly changed, then her altered behavior in a subsequent interaction may cause a change in the grammar of her interlocutor, and so on.

The mathematics of a model built up from many probabilistic interactions of agents can be unwieldy, however. Rather than trying to calculate directly how the system will behave, researchers often use computer simulation as an experimental tool. Artificial agents with the desired properties and behaviors are left to interact and change, and the results observed. Through such simulations, the effects of probabilistic learning and behavior on language change can be explored, and we can determine under what conditions a change will continue or accelerate, and under what conditions variation is stable.[14]

5.5.1 The Adoption of New Words

Elsewhere (Zuraw 2000), in presenting a model of exceptions and regularities in the phonological grammar, I show that listeners' probabilistic updating of their lexicons can shape the integration of new words into a language. The puzzle I attempted to address is that even though there is a resistance to applying semiproductive phonology to new words, as words become integrated into the lexicon they begin to undergo semiproductive phonology at rates similar to those found in the established lexicon.

In the proposed model, semiproductive phonology is encoded in a stochastic optimality-theoretic grammar (see Boersma 1998; Boersma and Hayes 2001) by low-ranking markedness constraints. Existing words' behavior is determined by high-ranking faithfulness constraints that require the preservation of idiosyncratic properties encoded in lexical

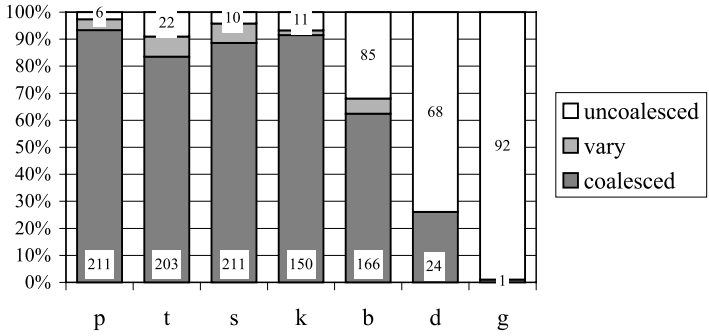

Figure 5.2
Rates of nasal coalescence in the native Tagalog lexicon, broken down by stem-initial consonant

entries. These constraints do not apply to new words, however, because those words lack lexical entries. The "subterranean" constraints emerge, therefore, to determine the pronunciation of new words.

The example examined in greatest depth is Tagalog nasal coalescence, which fuses a prefix-final nasal with a stem-initial obstruent.

(10)

Stem	*Nasal-coalesced*
bákat 'mark, scar'	ma**m**ákat 'to leave a scar'
Stem	*Prefixed but not nasal-coalesced*
bajáni 'hero, helper'	ma**mb**ajáni 'to offer cooperation'

As shown in figure 5.2, nasal coalescence appears to be distributed in the lexicon according to a pattern—voiceless obstruents are much more likely than voiced to undergo it, and obstruents with fronter places of articulation are somewhat more likely than those with backer places to undergo it—but the pronunciation of individual words is unpredictable and must be memorized. I argue that words with nasal-coalescing prefixes (or at least some of them) have their own lexical entries, and thus high-ranking faithfulness constraints against coalescing, splitting, or inserting segments within a lexical entry ensure that they are pronounced correctly. An additional constraint, USELISTED, which prefers inputs to be a single lexical entry, ensures that if a lexical entry exists, it is used as the basis for evaluating faithfulness.

When no lexical entry exists, as when a prefixed form is created for the first time, USELISTED cannot be satisfied, and the faithfulness constraints do not apply, so it falls to low-ranked constraints to decide probabilisti-

cally whether nasal coalescence should apply. I further assume that the strength of a lexical entry grows gradually as instances of the word are encountered, and that a lexical entry with strength of 0.5, for example, is available for use only half the time. Thus, in 50% of utterances, USE-LISTED and the faithfulness constraints will enforce the memorized pronunciation of such a half-strength word, but in the other 50% of utterances, the lower-ranked constraints will decide, because the lexical entry has not been accessed.

Boersma's (1998) Gradual Learning Algorithm is shown to be able to learn the distribution of nasal coalescence from exposure to the lexicon and encode that distribution in the ranking of subterranean constraints, preferring nasal coalescence on voiceless obstruents and dispreferring nasal coalescence on back obstruents (crosslinguistic motivations are suggested for both). The behavior of the resulting grammar in generating and assigning acceptability ratings to new morphologically complex words is shown to be a fair match to experimental results with speakers. The part of the model that I will describe here concerns the grammar and lexicon's effects on the adoption of new words by the speech community.

The grammar is biased against applying semiproductive phonology to new words (this is just the definition of semiproductivity in this model: an unfaithful mapping from input to output is productive to the extent that the ranking values in the grammar allow it to apply to new words). This is consistent with experiments in several languages, finding that speakers are reluctant to apply semiproductive phonology to new words; although various aspects of the experimental design can increase apparent productivity, it is always less than what might be expected from looking at the lexicon (Bybee and Pardo 1981; Eddington 1996; Albright, Andrade, and Hayes 2001; Suzuki, Maye, and Ohno 2000). It has also been observed in many cases, however, that words eventually tend to conform to existing lexical patterns after they have been in the vocabulary for some time. In the Tagalog case, prefixed forms of Spanish loan-stems undergo nasal coalescence at rates similar to those seen in the native vocabulary, as shown in figure 5.3. This phenomenon seems counterintuitive, if we expect that the more frequent pronunciation early in a word's life (i.e., without nasal coalescence) should take over and become the conventionalized pronunciation as the word establishes a lexical entry in the minds of speakers.

I propose that the solution lies in probabilistic interactions between speakers and hearers, specifically in probabilistic reasoning on the part of

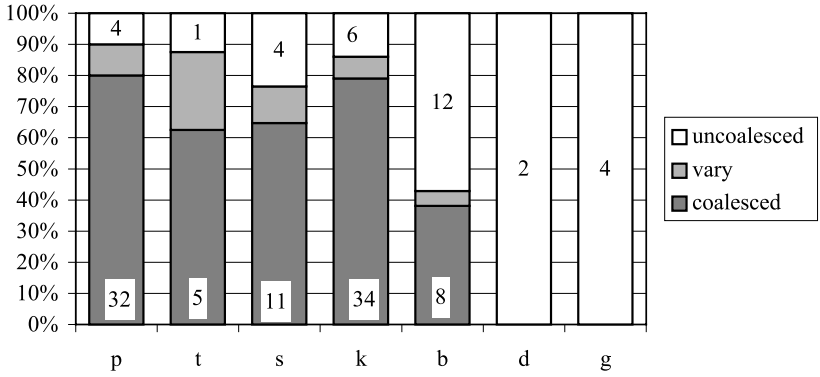

Figure 5.3
Rates of nasal coalescence in Spanish loan-stems

the listener. Because morphologically complex words can be either drawn directly from their own lexical entries or formed synthetically by morpheme concatenation, the listener must decide whether a morphologically complex word that she hears was lexical or synthesized for her interlocutor, assuming that she wants to maintain a lexicon that is similar to her interlocutor's. For example, if a listener hears *mambulo*, she must guess whether the speaker was using a lexicalized word *mambulo* or merely concatenating the prefix *maN-* with the stem *bulo*. Factors that should enter into the calculation include how strong the listener's own lexical entry for the word is (if she has one at all) and how likely it is that a lexicalized or concatenated input, respectively, would produce the observed pronunciation. The listener can apply Bayes' rule:

$$P(\textit{synthesized}\,|\,\textit{pronunciation}) = \frac{P(\textit{pronunciation}|\textit{synthesized}) \cdot P(\textit{synthesized})}{P(\textit{pronunciation})}, \tag{11}$$

$$P(\textit{lexicalized}\,|\,\textit{pronunciation}) = \frac{P(\textit{pronunciation}|\textit{lexicalized}) \cdot P(\textit{lexicalized})}{P(\textit{pronunciation})}. \tag{12}$$

The grammar influences that calculation, because the probabilities $P(\textit{pronunciation}|\textit{synthesized})$ and $P(\textit{pronunciation}|\textit{lexicalized})$ depend on the grammar. $P(\textit{pronunciation}|\textit{lexicalized})$ is always close to one, because of the high-ranking faithfulness constraints. $P(\textit{pronunciation}|\textit{synthesized})$

is higher for non-nasal-coalesced pronunciations than for nasal-coalesced pronunciations—recall that the grammar somewhat disfavors nasal coalescence on new words. Therefore, there is a bias toward classifying non-nasal-coalesced words as synthesized and nasal-coalesced words as lexical. Intuitively, the low productivity of a phonological rule encourages speakers to interpret words that do display the rule as exceptional and therefore listed.

The lexicon also influences the calculation, by contributing to $P(\textit{synthesized})$ and $P(\textit{lexicalized})$. $P(\textit{synthesized})$ depends on the construction's productivity, determined by how many morphologically and semantically eligible words participate in the construction. $P(\textit{lexicalized})$ depends on the candidate word's similarity to existing words. Thus, pronunciations that are similar to existing, lexicalized words (e.g., nasal-coalesced voiceless front obstruents and non-nasal-coalesced voiced back obstruents) are more likely to be interpreted as lexical.

If a hearer does decide that a word was lexicalized for her interlocutor, she will create a weak lexical entry for it. The existence of this weak lexical entry means that when it is the hearer's turn to speak, she has some small probability of using it. The bias toward recording nasal-coalesced words as lexical, especially when they resemble existing nasal-coalesced words (and ignoring non-nasal-coalesced words as synthesized) results in stronger lexical entries for nasal-coalesced pronunciations, which in turn results in an increase in the number of nasal-coalesced productions, leading to further strengthening of lexical entries for nasal-coalesced pronunciations.

The model was implemented in a computer simulation with 10 agents of varying ages who from time to time "die" and are replaced by agents with empty lexicons. To avoid undue influence from young speakers with immature lexicons, in each speaker-hearer interaction the hearer probabilistically decides, as a function of the speaker's age, whether to let her lexicon be affected by the speaker's utterance.[15] Different pronunciations for the same word (nasal-coalesced and not) do not directly compete, but if they are not reinforced, lexical entries decay.[16] Therefore, if two pronunciations remain prevalent, agents can have two strong pronunciations for the same word. This is a desirable result, because there are certain nasal-coalesced words whose pronunciation is variable within speakers. But if one pronunciation becomes much more common than the other, the lexical entry for the uncommon pronunciation will gradually decay.

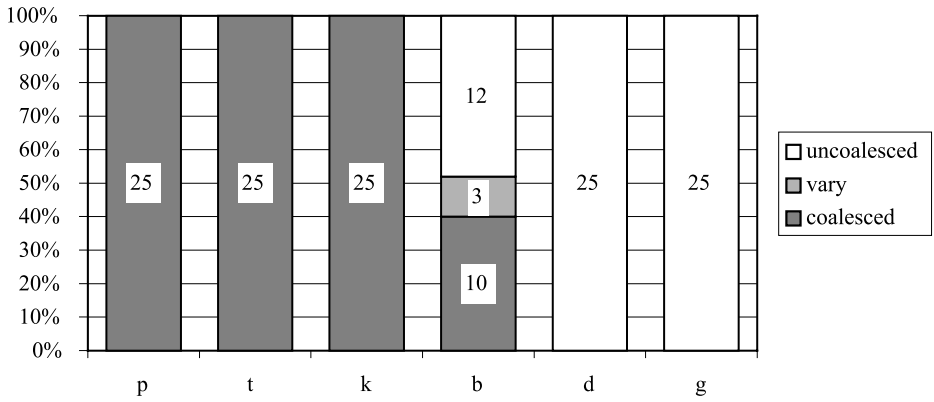

Figure 5.4
Rates of nasal coalescence in new words in a simulated speech community

The result of the simulations, shown in figure 5.4, was that new words were incorporated into the lexicon in a pattern similar to that seen among Spanish stems (though somewhat more extreme—this could be because some of the Spanish-derived words have not been in use long enough for their pronunciations to reach their final state). That is, for voiceless obstruents, the final rate of nasal coalescence is nearly 100%; for front, voiced obstruents, it is around 50%; and for back, voiced obstruents, it is close to 0%.

Probabilistic reasoning by adults, then, could explain the maintenance of lexical regularities over historical time. Such reasoning requires speakers to have a probabilistic grammar, so that there is variation in the treatment of new words, and it requires listeners to have access, whether direct or indirect, to the statistical characteristics of the lexicon.

5.5.2 Learners' Response to the Probabilistic Environment

If language change is a shift in the distribution of competing variants, what causes that distribution to change from one generation to the next? Why don't children mimic the frequencies of their elders? We have seen that in some cases—generally cases of morphosyntactic change—the shift in frequency appears to reflect a reanalysis (e.g., Frisch's *not* case, Tabor's *sort of* case): younger speakers use a construction at a different rate because they assign a different interpretation to it. In other cases—generally cases of phonological change—younger speakers may use the older

or the newer variant according to the requirements of the social setting (i.e., formal vs. informal), indicating that they control a grammar that is qualitatively similar to their elders', but that assigns different probabilities to different variants. Biber and Finegan (1989) argue that stylistic shifts in written English over the last four centuries similarly reflect sociological, rather than structural, motivations.

Another possible source of frequency shift that has been proposed is the conflict between frequency and learnability: some variable situations could be inherently unstable, depending on learners' bias in dealing with ambiguous utterances. Yang (2000) and Briscoe (1999) both explore this idea within a principles-and-parameters framework (Chomsky 1981b), where acquisition is the process of parameter setting.

For Yang, no parameters are preset—all settings must be learned. The learner has a finite number of grammars to choose from, each having an associated weight that the learner maintains.[17] In each learning trial, the learner receives an input sentence and probabilistically selects one grammar, with higher-weighted grammars more likely to be chosen. If the grammar selected can parse the sentence, then the learner augments its weight and decrements the weight of the other grammars. If the grammar selected cannot parse the sentence, then the learner decrements its weight and augments the weights of all the other grammars.

The final weight that each grammar will attain depends in part on the distribution of grammars among the adults providing the learning data, but also on how many ambiguous sentences occur and what the learner does with them. For example, adults using a V2 grammar will produce a high proportion of sentences that are compatible with an SVO grammar.

Yang shows how this system can cause a drift in grammar probabilities. Suppose that the learning environment contains two grammars G_i and G_j, and that a proportion α of G_i's sentences are incompatible with G_j (this is G_i's *advantage*—the proportion of G_i-generated sentences that unambiguously lead the learner to strengthen the weight of G_i), and a proportion β of G_j's sentences are incompatible with G_i (G_j's *advantage*). These proportions vary according to the specifics of the grammars and according to the likelihood of various utterances; for example, the likelihood that an unambiguously V2 sentence is uttered given a V2 grammar may be quite different from the likelihood that an unambiguously SVO sentence is uttered given an SVO grammar. At generation n, the linguistic environment contains some proportion p of adult utterances from G_i and some proportion q of adult utterances from G_j ($p + q = 1$).

The probability that grammar G_i will have its weight incremented in any learning trial is αp, and the probability that G_j will have its weight incremented is βq. The learners will therefore tend to converge on new weights $p' = \alpha p/(\alpha p + \beta q)$ for G_i and $q' = \beta q/(\alpha p + \beta q)$ for G_j. This means that the weights have been learned unfaithfully ($p' \neq p$ and $q' \neq q$), except in the special case of $\alpha = \beta$.

For G_j to overtake G_i, q needs to grow at p's expense. This means that $p'/q' < p/q$ (the ratio of p to q decreases), or, coalescing the equations for p' and q' obtained above, $\alpha p/\beta q < p/q$, or $\alpha < \beta$: G_i's advantage must be smaller than G_j's. Yang examined corpora in two case studies to see if $\alpha < \beta$ does indeed cause G_j to overtake G_i.

The first case study was the change from Old French's V2 grammar to Modern French's SVO grammar. The SVO grammar must have generated more sentences not analyzable as V2 (i.e., SXVO and XSVO sentences) than the V2 grammar generated sentences not analyzable as SVO (i.e., XVSO and OVS sentences). Certain sentences would have been ambiguous: SVO, SVOX, SVXO. To get an idea of how many unambiguous sentences an SVO grammar would generate, Yang looked at modern SVO English and found that 10% of sentences are SXVO or XSVO (SVO's advantage). Looking at modern V2 languages, Yang found that the combined proportion of XVSO and OVS sentences (V2's advantage) is 30%. If these advantages also held for competing SVO and V2 grammars in transitional French, then SVO should not have been able to overtake V2. Yang proposes that the solution lies in Old French's null-subjecthood: null-subject XVS sentences would be produced XV, which is also compatible with an SVO analysis (the XV sentence would be interpreted as XSV). Taking null subjects into account, V2's advantage is only 5%–18%. If it fell below about 10%, then SVO would begin to take over.

The second case study was the change from V2 in Middle English to SVO in Modern English. The problem is similar to that in the French case: why would SVO take over? Yang proposes that the answer here is Middle English's pronominal proclitics, which resulted in some XSVO and OSV sentences ("V3"). When cliticization waned and these pronouns had to be reanalyzed as real DPs, the V3 sentences would have been compatible only with an SVO grammar, adding to its advantage.

In Briscoe's (1999) model, the instability of a variable grammar comes not from the frequency of ambiguous sentences, but from the (overturnable) presetting of certain parameter values.[18] On the one hand, a more frequent variant has more opportunities to shape the learner's

grammar; but on the other hand, the more learnable variant—the one that uses more default parameter settings—has an advantage from the start. Briscoe simulates changes in word order, using a generalized Categorial Grammar framework, in which the syntactic rules are weighted so that different well-formed sentences have different probabilities of being uttered under a particular grammar. The parameters of the grammar include the default head-complement order, the order of subject and verb, the order of verb and object, and several others.

Certain parameter settings are associated with prior probabilities, intended to reflect innate markedness. Parameter settings are changed only when the learner's current grammar fails to parse an incoming sentence. The learner tries changing some settings, and if this makes the input sentence parsable, those potential new settings are strengthened, though not adopted right away. If the strength of a setting exceeds a threshold value, the new setting is adopted, though it can be changed back if contrary evidence is encountered later. Even after a setting is adopted, its strength continues to be updated; this determines how easy it will be to reverse the setting later. Thus, during learning each parameter has an innate prior probability, a posterior probability derived from learning, and a current setting.

Briscoe's approach differs from Yang's in that, although the learner keeps several grammars under consideration, only one grammar is used at a time, and thus individual adults' outputs will not be variable. The learner chooses the most probable grammar according to Bayes' rule. Letting g be a grammar, G the space of possible grammars, and t_n a triggering input (= the set of sentences seen so far), the probability of a grammar g given a triggering input t_n is given in (13):

$$P(g \in G|t_n) = \frac{P(g) \cdot P(t_n|g)}{P(t_n)}. \tag{13}$$

The prior probability of $P(g)$ is equal to the product of the probabilities of all its parameter settings. $P(t_n|g)$, the probability that a given grammar produces the set of sentences seen, is derived from the rule weights of each grammar. The denominator $P(t_n)$ of (13) is unimportant, because it is the same in all grammars being compared. The grammar that the learner uses to try to parse incoming sentences, and that the learner will use when speaking, is simply the most probable $g \in G$.

Briscoe's model, like Yang's, is sensitive to the amount of overlap in triggers (e.g., surface *SVO* as evidence for either an SVO grammar or a

V2 grammar). Briscoe found that in a population of mainly SOV+V2-speaking adults ("German") and some SVO-speaking adults, learners reliably converged to SOV+V2 as long as the percentage of unambiguously SVO triggers did not exceed 15% (the number depends on the strength of the default settings, if any).[19] If the percentage exceeded 15%, a drift toward SVO could begin.

The learner's response to a variable environment is crucial to language change. But in order to ensure that learners do not merely replicate the frequencies around them, there must be some persistent bias at work, whether it comes from social motivations, from learnability, or from ambiguity.

5.5.3 Language Change under Competing Forces

We have seen several forces that may be at work in probabilistic language change: innate parameter settings, frequencies of ambiguous utterances, frequencies of individual lexical items or constructions, variation due to language or dialect contact. In real cases of language change, however, it can be difficult to tease apart these factors to determine which are necessary or sufficient triggers for various changes. As Dras, Harrison, and Kapicioglu (2001) note, simulation studies provide a way to perform diachronic experiments on language, altering the strength of each force and observing the effects.

Few such simulations have been undertaken that attempt to model real changes, but this line of inquiry seems promising. This section concludes by reviewing preliminary results from one simulation project (Dras, Harrison, and Kapicioglu 2001) that investigated the effects of various forces on changes in vowel harmony. Dras, Harrison, and Kapicioglu collected corpus data on Old Turkic—in which 100% of words were harmonic for palatality and backness—and several of its descendants, from the ninth century to the present. Some of the contemporary languages have maintained very high rates of harmony, while others' rates of harmony have fallen drastically. Dras, Harrison, and Kapicioglu identified internal and external factors that could affect rates of vowel harmony—vowel coarticulation, inherent markedness of certain vowels, consonantal effects, merger of vowels (collapsing harmony pairs), the introduction of disharmonic loanwords, and language contact—and modeled several of them.[20] Agents in the simulation exchanged words with randomly selected neighbors, updating their lexical entries to reflect what they had heard. When speaking, an agent might mispronounce a word or coarticulation; when

listening, an agent might mishear, ignore a coarticulation, or adjust an interlocutor's pronunciation before adding it to the lexicon. Agents might also mutate their lexical entries at an agent-specific rate; if a vowel was to be mutated, there was an agent-specific probability that it would be made harmonic.

If factors favoring harmony were strong enough, harmony could increase, following roughly an S-shaped curve. Dras, Harrison, and Kapicioglu found that vowel merger alone was not sufficient to eliminate harmony, nor was the addition of disharmonic loanwords. Though the authors emphasize that their results are preliminary and that the model needs to be enriched with several more factors, this study shows a promising direction for future work: using probabilistic simulation tools and real historical data to model the effects of a variety of internal and external factors on language change. Such simulations should help us determine which factors, alone or in conjunction, are strong enough to cause and continue language change.

5.6 Conclusion

Many linguists are interested in language change because of what it can tell us about synchronic language. For example, the types of reanalysis that are common may tell us about the learning mechanism, and the way a change spreads through the speech community may tell us about the social function of language.

Because the study of language change draws on all areas of linguistics, and particularly on probabilistic approaches to all areas of linguistics, the study of language change also has something to contribute to the study of probabilistic linguistics: models of synchronic acquisition, representation, and use of language must be consistent with observed diachronic facts.

We have seen that the probabilistic behavior of learners, speakers, and listeners can shape language change and that simulation studies can help us explore how this happens. Simulation studies can also support or undermine models of how agents represent and use linguistic knowledge: if a model yields a good match to the known facts concerning language change, it is to be preferred over one that does not. In Tabor 1994, for example, a connectionist model of syntactic knowledge is supported by its ability to model frequency linkage and the relationship between frequency changes and reanalysis. In Zuraw 2000, a probabilistic model of

knowledge of lexical regularities is supported by its ability to model the incorporation of new words into the lexicon.

Language change can be a testing ground, then, for probabilistic models of learning, speaking, and listening. It is to be hoped that current advances in our understanding of the probabilistic nature of the language faculty will have much to contribute to the study of language change over the coming years, and that the study of language change can return the favor.

Notes

Many thanks to the editors, Rens Bod, Jennifer Hay, and Stefanie Jannedy, and to Bryan Zuraw, for their substantial feedback. The greatest thanks go to Norma Mendoza-Denton and Chris Manning, whose comments significantly shaped the form and content of this chapter.

1. When a group of languages is known to be related, we may wonder how closely. Although answering this question requires quantitative techniques, it generally does not involve probabilistic tools. Determining degree of relatedness involves establishing a similarity metric and a clustering technique to group the most similar languages most closely. See Embleton 1986 for a review of quantitative techniques in tree reconstruction, and Guy 1980a,b, for some interesting computer experiments. Current work in comparative dialectology (e.g., Kessler 1995; Nerbonne and Heeringa 2001) similarly explores similarity metrics and clustering techniques, although this literature generally does not seek to establish facts about genetic relatedness but rather seeks to quantify the degree of current similarity between dialects in a way that might, for example, be useful to language planners and educators.

2. The tools of probability can do little to help us identify loans; obvious loans could be excluded from word-lists, but, as Kessler (2001) observes, borrowings that took place in the distant past may be impossible to detect.

3. An analogous nonlinguistic example is the "lottery fallacy": even though you will almost certainly not win the lottery this week (even if you buy a ticket), it is quite likely that someone will win. It is perhaps the high likelihood of the general event that makes the specific event seem within reach.

4. $\binom{p}{q}$, "p choose q," is the number of ways that a subset set with q elements can be chosen from a set of p elements. $\binom{p}{q} = p!/((p-q)! \cdot q!)$, where $n!$, "n factorial," is equal to $n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot 3 \cdot 2 \cdot 1$.

5. χ^2 is the sum of $(O-E)^2/E$ for each cell, where O is the observed value and E is the expected value (the value that the cell would have if the proportion of entries per row were the same for each column and vice versa).

6. In the shift test, for a list of length n, only n permutations are considered: shifting list B down by 0 lines, by 1 line, by 2 lines, and so on.

7. Or, equivalently,

$$P = \frac{e^{k+st}}{1 + e^{k+st}}.$$

8. *ln* stands for *natural logarithm*. $\ln(x) = y$ means that $x^y = e$, where $e \approx 2.72$ is the so-called natural number.

9. This is the source of Tabor's "Q-divergence": changes in a word's category are accompanied or even preceded by changes in the frequency with which it appears in an ambiguous context.

10. A potential counterargument is that if *ne* was lost for phonological reasons, it might have been preserved in the written record for a deceptively long time.

11. Niyogi and Berwick make the point that only the five-parameter system, not a similar three-parameter system, tends toward loss of V2. The three-parameter system actually produces a tendency toward increasing V2. Therefore, Niyogi and Berwick argue, diachronic simulations such as theirs can be a way of investigating the plausibility of substantive proposals in linguistic theory.

12. Phillips found that the stress shift illustrated by *convíct* (noun or verb) becoming *cónvict* (noun)/*convíct* (verb) affected infrequent words first. She also found, however, that final stress on verbs ending in *-ate* in British English developed first on frequent words. Phillips suggests that the *-ate* shift is not really a morphological rule in Bybee's sense. Applying the stress shift does not require analyzing the morphological category or morphemic structure of a word; rather, it involves ignoring *-ate*'s status as a suffix.

13. Bybee (in press) proposes that just as frequent words can retain archaic phonology or morphology, so frequent words and constructions can retain archaic syntax. Bybee proposes this as the explanation for why the English modal auxiliaries (*can*, *might*, *should*, etc.) retain their archaic behavior with respect to negation and question formation: *He **should** not*, ***Should** he?* versus *He does not **drive***, *Does he **drive**?* The frequency effect could be thought of as complementary to, or a driving force for, the traditional explanation that only modals occupy Infl.

14. A growing literature simulates "language evolution," that is, change whose starting point is a speech community that lacks any shared linguistic knowledge. See Steels 2000 for a very brief overview; for further reading, see Kirby 1999 and many of the papers in Hurford, Studdert-Kennedy, and Knight 1998 and Knight, Studdert-Kennedy, and Hurford 2000.

15. Like most agent-based simulations of language change, the model lacks social or spatial structure: each agent has an equal probability of interacting with any other agent. Because leadership in a language change seems to be correlated with the characteristics of the speaker's social network (see Milroy 1980; Labov 2001), a simulation that attempts to model the spread of a sociolinguistically correlated change through the speech community would require a more realistic social structure. Social structure is probably irrelevant, however, to merely demonstrating the transmissibility of a lexical pattern.

16. In the simulation reported in Zuraw 2000, pronunciations do compete directly: when one is augmented, the other is diminished. The results reported here are from a later version of the simulation.

17. In many situations, it is unrealistic for the learner to maintain a full list of possible grammars. In a principles-and-parameters model, the number of possible grammars is v^p, where v is the number of values that each parameter can take, and p is the number of parameters. Even with binary parameters ($v = 2$), this number grows very quickly as the number of parameters grows. In Optimality Theory, the situation is even worse: the number of grammars is $c!$, where c is the number of constraints. A model in which individual parameter settings or constraint rankings are probabilistic is more tractable (see Boersma 1998; Boersma and Hayes 2001 for probabilistically ranked constraints within a single grammar).

18. This paper also includes some interesting simulations of creole genesis.

19. The source of the variation could be dialect contact, social prestige, or random mislearning by the older generation.

20. Coarticulation, inventory structure, lexical patterns, vowel merger, fixed nonharmonic suffixes, and disharmonic loanwords.

Chapter 6

Probabilistic Phonology: Discrimination and Robustness

Janet B. Pierrehumbert

6.1 Introduction

Phonology deals with the implicit knowledge of language sound structure as used contrastively to convey meaning. The ability of humans to use this knowledge productively shows the need for an abstract generative theory of phonology. A fluent adult speaker of a language can produce new words with native word-level allophony, produce novel combinations of words with native phrase-level allophony, accommodate borrowings to native phonotactics, and create morphological neologisms involving phonological modifications of the component parts. Baayen (2001) presents calculations suggesting that new words are continually created. No matter how large a corpus a researcher is working with, a substantial increase in its size will uncover additional words. Thus, phonology is productive for the same reason that syntax is: to express novel meanings, people construct new combinations of familiar parts.

The productivity of phonology is widely believed to constitute evidence for a theory in which the phonology of any particular language has the character of a formal grammar, and the universal theory of phonology delineates the types of grammars that are available for use in individual languages. For example, the phonology of Finnish includes a set of terminal elements, such as features or phones. It includes principles for combining these elements into well-formed syllables, metrical feet, and phonological words. These principles permit /h/ in coda position (as in the word /kahvi/ 'coffee'). They permit the double attachment of a phoneme to coda and onset position (as in the word /hɛlp:o/ 'easy'). They preclude the kind of complex onsets found in English /strit/ *street*. They set up an alternating word pattern, described with metrical feet. They enforce initial word stress absolutely. Analyses of this sort are very

familiar. The assumption I want to emphasize about them is that they involve synoptic knowledge that exploits abstract variables, such as C (consonant), μ (mora), and σ (syllable). If the knowledge were less abstract and synoptic, it would not generalize to novel cases in the specific ways it does.

This conception of phonology as a formal grammar (with abstract variables) is often assumed to stand in opposition to the idea that phonology involves statistical knowledge. However, this opposition is spurious, because probability theory requires us to assign probability distributions to variables. Without variables, there would be no way for a statistical learning model to tabulate any statistics about anything. Once we have variables, they can be as abstract as we like; in principle, we can assign probability distributions to any variable of any nature that we may care to define in response to scientific findings.

Introducing probability distributions into the model provides obvious and well-established tools for handling variable data and gradient outcomes. However, it in no way precludes a rigorous treatment of phenomena that prove to be highly categorical. Such phenomena can be handled in probabilistic models by assigning extreme probabilities to particular outcomes: namely, a probability of 0 to a nonoccurring event and a probability of 1 to an outcome that is certain. In short, within a probabilistic model, nonprobabilistic fragments of the grammar are readily treated as limiting cases. (For complete formal development of the intrinsic connection between probability theory and logic, see Carnap 1950 and Adams 1998.) In the remainder of this chapter, I will assume that the ultimate and true theory of phonology will involve both probability distributions and abstract variables, because the abstract variables have probability distributions over other levels of representation. My discussion will focus on questions of which distributions over which variables. I will argue that careful consideration of how statistical distributions are established and distinguished from each other enables us to reach important conclusions about the nature of human language, conclusions that would elude us in a nonprobabilistic framework.

In viable theories of phonetics/phonology, there is a ladder of abstraction, each level having its own representational apparatus. Thus, the theory as a whole must delineate both the available representation at each level and the principles relating one level to another. In the remainder of this chapter, it will be important to distinguish the following levels. This list represents a minimal rather than definitive list of levels of representa-

tion, and representational distinctions within each level on the list have been glossed over when they are not central to the goals of the chapter.

1. *Parametric phonetics*. The parametric phonetic representation is a quantitative map of the acoustic and articulatory space. In speech perception, it describes the perceptual encoding of the speech signal on each individual occasion. In speech production, it describes the articulatory gestures as they unfold in time and space.

2. *Phonetic encoding*. The phonetic encoding system of a language abstracts over the parametric phonetic space, defining the inventory available in the language for encoding word forms (phonological representations of words). In traditional phonology, these categories are taken to be phonemes. Phonemes are minimal contrastive units and they can be equated across different structural positions; the same phoneme occurs at the beginning of *pat* and the end of *step*. However, much recent evidence (reviewed below) suggests that phonetic categories are considerably less abstract than phonemes. They are not minimal in the sense that they include redundant information. Nor can they be equated across contexts. Thus, they can be viewed as peaks in the total phonetic distribution of the language (e.g., areas of the parametric space that the language exploits preferentially) or as positional allophones. Phonetic encoding, as used here, includes not only speech segments but also aspects of prosody and intonation that are defined with respect to the speech signal.

3. *Word-forms in the lexicon*. Each word in a speaker's lexicon has a representation of its sound structure that allows it to be recognized despite variation in its phonetic form resulting from speaker differences and context. The same representation presumably mediates between perception and production, making it possible for speakers to repeat words they have acquired through perception. Given this description, it is clear that word-forms are also abstractions over the phonetic space. A given word is learned through repeated exposure to that word in speech. Pervasive word frequency effects in psycholinguistics show that a word's frequency of occurrence affects the long-term representation of that word. Connections between word frequency and detailed quantitative aspects of pronunciation are documented in Bybee 2001 and Jurafsky, Bell, and Girand, 2002. Such phenomena strengthen the point that words are generalizations over speech.

4. *The phonological grammar*. The phonological grammar, encompassing both prosodic structure and phonotactics, describes the set of possible words of a language. (Phrasal phonology is beyond the scope of

this chapter.) The grammar is revealed by well-formedness judgments as well as neologisms and borrowings. It is also revealed by the procrustean adjustments that morphologically complex forms can undergo if they become lexicalized as units.

Phonology represents generalizations over the word-forms in the lexicon, which are in turn generalizations over speech. Hence, phonology does not abstract over speech directly, but rather indirectly via the abstraction of word-forms. This jump in level has important consequences. First, we can find discrepancies between the strongest patterns in the lexicon and the strongest patterns in running speech. Discrepancies occur because a pattern that is common in running speech may be rare in the lexicon if its occurrence in running speech results from just a few different words. Second, the lexicon is small in comparison to the total number of word tokens a person encounters. An adult speaker of English knows on the order of 10,000 monomorphemic words and 100,000 words total (see discussion below). Any generalization in the word-level phonology must be learnable from a data set of this size. In contrast, the 200-million word corpus used in Baayen's (2001) calculation cited above would correspond to about 18,000 hours of speech, about the amount of speech native speakers have encountered by the time they reach adulthood, assuming they hear speech two to three hours per day. Language learners have at their disposal a vast amount of data for learning generalizations over speech, such as native details of pronunciations. The consequences of this vast discrepancy between lexicon size and number of encountered word tokens will be discussed further below in connection with statistical robustness.

5. *Morphophonological correspondences*. A given stem or affix can assume phonologically different forms in different, related words. Many of the differences are attributable to general constraints of the phonological grammar. For example, in a standard analysis the /n/ pronounced in the word *hymnal* (before a vowel-initial suffix) is not realized in the bare form of the stem *hymn*. This fact may be attributed to sequencing constraints within the syllable. The vowel-initial suffix rescues the consonant by parsing it into onset position. A general line of development in phonological theory, launched by Kisseberth's (1970) work on conspiracies, aims to show how morphophonological alternations arise when contextual factors make a given stem subject to different surface constraints. This enterprise, insofar as it succeeds, minimizes the need for explicit statements about alternations. Optimality Theory is a recent manifesta-

tion of this trend. However, no current theory explains all morphophonological alternations on the basis of general phonological patterns, because many such alternations are not, in fact, general. Many alternations need to be learned on the basis of specific relations between words or within paradigms. I will refer to such alternations as *correspondences*.

Morphophonological correspondences will not be the primary topic of this chapter, since they are treated independently by Baayen (this volume). The cases that Baayen analyzes in depth are linking elements in Dutch nominal compounds and final voicing/devoicing in the past tense of Dutch verbs. Parallel cases in English include irregular voicing in the English plural, as in *thief*/*thieves*, and (perhaps surprisingly) English flapping, as in the contrast between *capitalistic*, which shares the flap of *capital*, with *militaristic*, which maintains the aspiration of *military* (Withgott 1983). Such patterns involve generalizations over pairs or sets of words. This point is brought home by an effort to estimate the plural of an unstable form. Is the plural of *roof rooves*, on the analogy *thief* : *thieves* :: *roof* : *rooves*? Or is it *roofs*, on the analogy *cuff* : *cuffs* :: *roof* : *roofs*? The AML model (Analogical Modeling of Language) developed by Skousen (1989) acknowledges this fact in its very name; for detailed discussion of this model, see Baayen, this volume. In Optimality Theory, output-output correspondence constraints, or sympathy constraints, acknowledge the same kind of relationship (see McCarthy and Prince 1995; McCarthy 1999). Morphophonological schemas in the usage-based theory of Bybee (2001) also involve generalizations over word relationships.

The levels of representation just introduced imply an organization of probabilities in the theory. Specifically, phonetic categories have probability distributions over the parametric phonetic space. Word-forms, viewed as sequences of phonetic categories, also have probability distributions over temporal sequences of events in the phonetic space. (These distributions may be deducible from the categories that comprise the word-form. Or nearly so.) The prosodic and phonotactic templates that define the phonological grammar have probability distributions over the word-forms in the lexicon. Morphophonological relationships also involve probability distributions over a universe of pairings or collections of word-forms.

In the following sections, I will first review empirical evidence that implicit knowledge at all levels of representation is probabilistic. Then, I will show how probabilistic reasoning predicts limits on the inferences

that can be made about language, either by the scientist or by the language learner. These limits lead to level-specific effects, because the phonetic encoding system and word-forms are acquired from vast exposure to speech, whereas the phonological grammar is abstracted over the lexicon. The lexicon provides a much smaller number of data points than running speech. Language is statistically robust, and this robustness forms its character at all levels. Finally, I will discuss correlations across levels and their implications for phonological theory.

6.2 Probability at Various Levels of Representation

6.2.1 The Phonetic Space and Categories over It

The phonetic inventory of a language is a set of labeled probability distributions over the phonetic space. By "phonetic space," I mean the acoustic and articulatory parameterization of speech as a physical event. For example, to a first approximation, vowels can be viewed as probability distributions over F1-F2 space, as shown in figure 6.1. Figure 6.1 pools data for the same vowel across speakers. Each vowel occupies a continuous region of the space. The regions for different vowels are quite distinct, even ignoring F3 and any other characteristics that distinguish them. Each vowel is more frequently instantiated by values near the center of its distribution than by values near the edges of its distribution.

The F1-F2 space is continuous, because the parameters vary continuously. It is possible to define a vowel location that is *between* any two particular vowel tokens, no matter how close these two may be to each other, and it is possible to define a distance metric between vowels that integrates the separation in the F1 and F2 directions. The same claims would apply to other physical parameters as well, such as spectral tilt, jaw opening, or activation of the cricothyroid muscle. If we conceive of the phonetic space in terms of all controllable or noticeable articulatory and acoustic dimensions, then it is a very high dimensional space indeed. A complex shape within this hyperspace describes the combinations of articulatory and acoustic values that human anatomy and physiology permit us to achieve.

Any individual language exploits as categories a reasonably small number of regions in hyperspace, with no language using all regions of the phonetic space equally. In fact, a rather sparse selection of regions is used by any individual language, as compared to the universal capabilities evidenced by the union of phonetic outcomes across all languages.

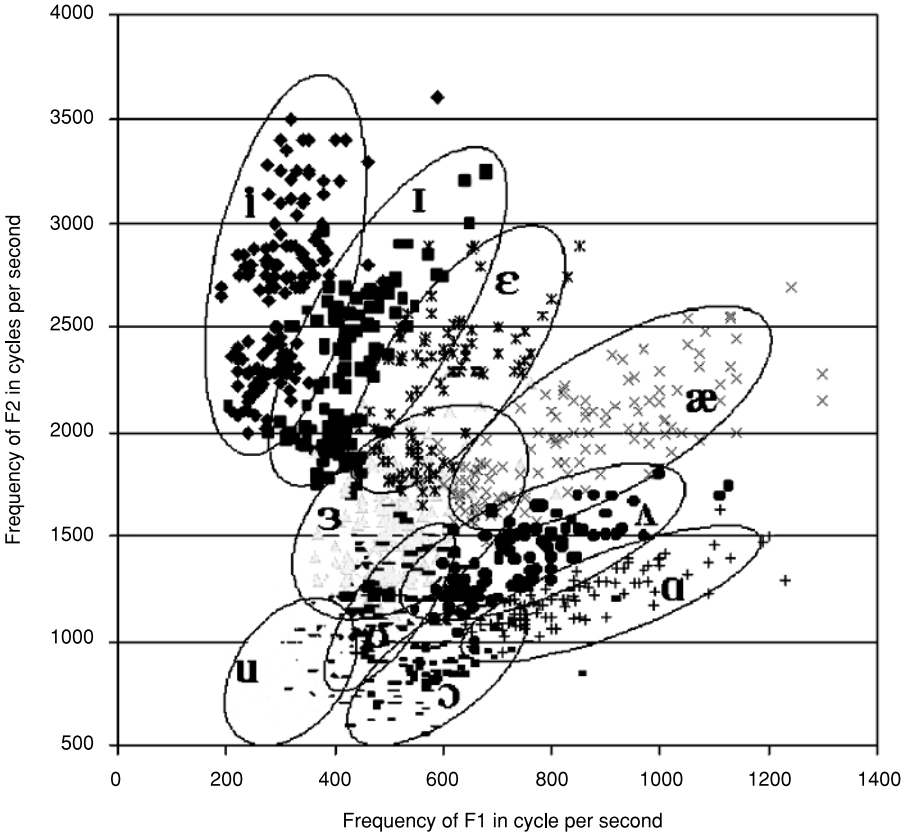

Figure 6.1
The F1–F2 vowel space. Data, taken from Peterson and Barney 1952, combine data for children, women, and men. Regions corresponding to each vowel are indicated. Only tokens that were unambiguously identified by listeners are included. In cases of overlap between regions, F3 or some other feature disambiguated the percept.

The largest phoneme inventories reported in Ladefoged and Maddieson's (1996) typological survey are still small compared to the complete IPA (International Phonetic Alphabet), viewed as an estimate of the total set of sound segments available for use in any language.

The claim that languages use regions of the phonetic space—as opposed to points in the space—is supported by the fact that the phonetic realization of any given element is *always* variable. Even repeated recordings of the same speaker saying the same word in the same context will yield some variability in the measured values of physical parameters. When we consider the amount of variation relating to differences in vocal tract anatomy, speech style, and idiolect that the speech perception system encounters, it is clear that the system has an impressive capability for coping with variation. Indeed, the whole point of abstract levels is to cope with variation. If productions of the same word were acoustically identical across speakers and situations, then the recognition system could work just by matching spectral templates, the mathematical equivalent of laying one spectrogram on top of another and holding the pair up to the light to see whether they were identical or not. Lexical entries could link spectral templates directly to meanings, and no more abstract phonetic encoding would be necessary.

Acquiring the phonetic encoding system of a language involves acquiring probability distributions over the phonetic space. Evidence that these distributions are acquired comes both from language typology and from studies of language acquisition. As discussed at more length in Pierrehumbert 2000 and Pierrehumbert, Beckman, and Ladd 2001, there is no known case of a phoneme that has exactly the same phonetics in two different languages. Even the most closely analogous phonemes prove to be systematically different when examined quantitatively in analogous contexts, and patterns of variation across contexts reveal even more quantitative differences. For example, Caramazza and Yeni-Komshian (1974) show that voice onset times for stops in Canadian French differ systematically from those in both American English and Continental French. Experiments reported by Flege and Hillenbrand (1986) show that vowel lengthening before a voiced fricative is quantitatively different in English and French, and that English and French listeners are attuned to this difference as a perceptual cue for the voicing of the fricative. Beddor and Krakow (1999) explore language-specific details of nasal coarticulation, and Beddor, Harnsberger, and Lindemann (in press) present similar results for patterns of vowel-vowel coarticulation in different languages.

Studies of language acquisition show that phonetic knowledge is acquired gradually. Although children show attunement to the encoding system of their language toward the end of the first year of life (Werker and Tees 1994), it takes a long time to reach adult levels of competence. In production, elementary-school children still lack adult levels of accuracy in durational and spectral patterns (Lee, Potamianos, and Narayanan 1999) and in control and perception of coarticulatory patterns (Nittrouer 1992, 1993). Hazen and Barrett (2000) present similar results for phoneme perception, showing that categorization boundaries for minimal pairs such as *coat*, *goat* sharpen gradually from ages 6 to 12, but at age 12 they still differ from those of adults. Such findings are readily modeled by assuming that the probability distribution corresponding to any encoding unit is incrementally updated through experience. In contrast, the (still common) assumption that phonological acquisition involves selecting items from a universally available categorical inventory (along the lines of the IPA) provides no way of representing the fine phonetic differences that define a native accent. It also fails to explain why children take so very long between positing a category in their language and using it with adult levels of precision in perception and production.

As discussed by Johnson (1997c) and Pierrehumbert (2001a), the perceptual learning involved in the gradual acquisition of detailed phonetic categories is readily modeled using exemplar theory. In exemplar theory, labels are associated with a distribution of memory traces in a parametric space, in this case a cognitive representation of the parametric phonetic space. These traces are the exemplars that give the model its name. Empirical distributions of exemplars associated with each label are gradually built up as speech tokens are encountered and encoded. This concept is illustrated in figure 6.2, in which a single dimension, F2 (the second formant value), is selected for the purposes of exposition. The two labels involved are /ɪ/ and /ɛ/, which are generally but not completely separated in F2. Vertical lines represent remembered instances of /ɪ/ and /ɛ/, with higher vertical lines representing the strong memory trace resulting from a pileup of recent examples. The empirical distributions are not individually normalized, and so /ɪ/, being more frequent than /ɛ/, is more abundantly represented. An incoming token of an unknown vowel is indicated by the asterisk, located at a point on the F2 axis that reflects its actual F2 value. According to a standard model of perceptual classification, the labeling of this token is determined by a statistical

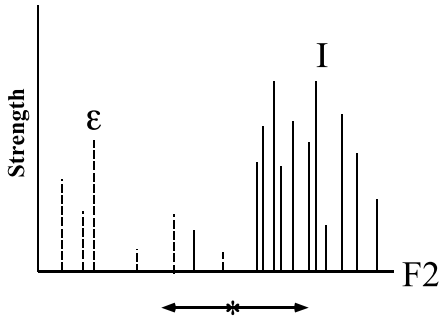

Figure 6.2
Classification of an unknown stimulus token, in the framework of exemplar theory. The asterisk indicates the F2 location of the unknown stimulus. The arrows show the window over which the evaluation is made. (From Pierrehumbert 2001a.)

choice rule that assigns the most probable label based on the location of the unknown stimulus in relation to the density of the competing labels in the neighborhood of the stimulus (see Luce, Bush, and Galanter 1963a; Kruschke 1992.) The relevant neighborhood is indicated by the arrows around the stimulus. Since the distribution for /ɪ/ shows a higher density in this neighborhood, the stimulus is classified as /ɪ/. This classified speech event would induce an additional labeled memory trace at that location. Through continual updating in this fashion, the mental representation of the distribution for each label is gradually built up with experience.

For details of production to mirror perception, it is necessary to run the model backward. Pierrehumbert (2001a, in press) provides an exact proposal for doing so. Once a label is selected (presumably, through a decision to say a word that involves that label), a production goal is established by sampling the empirical distribution for that label. Specifically, an averaged neighborhood around a randomly sampled point in the distribution serves as a production goal. Bias terms on the initial sampling and on the computed production goal are available to model systematic shifts in the use of the phonetic space. Such shifts can come about through choices of speech style and historical changes in progress.

Johnson (1997c) and Pierrehumbert (2001a) leave open the question of what causes a phonetic category to be initiated in the first place. Categorization could be tied to functionality at another level (in particular, to functionality in meaning contrasts). However, increasing evidence that

infants initiate categories of phonetic encoding before they know meanings of words has encouraged researchers to look for bottom-up factors in categorization. A series of experiments reported by Maye and Gerken (2000) and Maye, Werker, and Gerken (2002) contrasts encoding of the same phonetic continuum under two conditions, one in which the distribution of tokens over the continuum is bimodal and one in which it is unimodal. These researchers found that both adults and infants interpret the bimodal continuum as involving two categories, even in the absence of any information about meaning. These experimental results are substantially anticipated by calculations presented by Kornai (1998). Kornai carried out an unsupervised cluster analysis on vowel formant data from Peterson and Barney 1952. The clusters identified through this analysis are extremely close to the mean values for the 10 vowels of American English, according to Peterson and Barney. The success of this analysis reflects the fact that utilization of the vowel space is not uniform, as discussed above. Utilization is more intense in regions close to the prototypical values for vowels of the relevant language, and it is sparse in between. Results of this character are not confined to speech segments. Mermelstein (1975) launched a line of research on bottom-up detection of syllables by obtaining strong results on the basis of the amplitude integral of the speech signal. Pausing, lengthening, and voice quality changes may effectively cue major intonational boundaries. Thus, it appears to be a hallmark of natural language that units of phonetic encoding correspond to distributional peaks over the parametric phonetic space. I will return to the significance of this finding in section 6.4.

6.2.2 Word-Forms and Lexical Neighbors

People learn new words by encoding instances of those words encountered in speech. Recent results reviewed in Pierrehumbert, in press, indicate that these long-term representations of individual words are surprisingly detailed. Evidence for long-term encoding of subphonemic detail includes morphological transfer effects, as discussed by Steriade (2000), historical entrenchment of sociostylistic features, as discussed by Yaeger-Dror and Kemp (1992) and Yaeger-Dror (1996), and unconscious imitation of specific voices, as revealed in Goldinger's (2000) experiments. It may also be noted that infants acquire their first words before they abstract phonemes, and that their phoneme inventory is only gradually developed on the basis of recurring patterns in a growing lexicon (see

review in Vihman 1996). Results such as these dictated the present treatment of word-forms as generalizations over the speech stream that use the resources of the phonetic encoding system.

Psycholinguistic results show that a speaker's total store of words is not an unstructured list. Instead, the words are organized in a lexical network in which words are connected to each other by virtue of phonological and semantic relationships. During language acquisition, the similarities and contrasts among words in the early vocabulary play an important role in the subsequent evolution of the system (see Beckman and Edwards 2000 for an incisive review). In adults, each word's competitiveness in perception and production is strongly affected by the number and frequency of word-forms that are similar to it. Luce and Pisoni (1990, 1998) explored the consequences of such similarities for CVC syllables, using the term *lexical neighbor* for a CVC syllable that differs minimally from a given CVC syllable through the addition, deletion, or substitution of a single phoneme. They found that other things being equal, it takes longer to recognize words that have many lexical neighbors than words that have few lexical neighbors. The delay is particularly great if the lexical neighbors have high average frequency relative to the target; the hardest words to recognize are infrequent words with many frequent lexical neighbors, and the easiest are frequent words with few and infrequent lexical neighbors. However, it is not as widely acknowledged that these same factors affect both allophonic outcomes and well-formedness judgments, phenomena traditionally viewed as the purview of phonology.

Wright (1997) measured formant values in a set of "hard" and "easy" words produced in citation form. He found that the "hard" words were more fully articulated in citation form than the "easy" words, as evidenced by a somewhat expanded vowel space. In a similar vein, Scarborough (2002) reports that hard words have stronger nasal coarticulation than easy words, facilitating perception. Already in 1964, Greenberg and Jenkins showed that invented words that are highly similar to existing words receive high ratings as possible words of a language. Consider a new English word *canvle*. It is minimally different from several existing words, as shown in table 6.1. The existence of these real words means that *canvle* is a very good nonsense word. Bailey and Hahn (2001) undertook a detailed experimental and computational study of the effects of phonotactics and lexical neighbors on "wordlikeness" judgments of nonsense words. They found that both of these factors have identifiable effects, and

Table 6.1
Lexical neighbors for a nonsense word

Orthography	IPA
canvle	kænvəl
canvas	kænvəs
anvil	ænvəl
candle	kændəl
Campbell	kæmbəl

together they explain wordlikeness judgments better than either one does alone. Notably, words with improbable phonotactics were rated better if they had more lexical neighbors than if they had fewer.

Both lexical neighborhoods and phonotactics reflect general knowledge of the lexicon. Conceptually, both involve searching the lexicon for words that match some phonological template. However, there are important differences between these two concepts, which relate to differences in the level of abstraction involved. The lexical neighborhood is found by identifying the few words in the lexicon that nearly match a given word in its entirety. The lexical neighbors of a word can mismatch it in various, unrelated regards. In the example above, *canvle* mismatches *anvil* in initial position, whereas it mismatches *candle* in the onset of the second syllable. Phonotactic generalizations are made over all words in the lexicon, but involve far less detailed templates. For example, the fact that /nt/ is a phonotactically splendid nasal-obstruent cluster in English relates to the huge, diverse set of words containing /nt/ as either a heterosyllabic or a final cluster, including words such as *intrepid*, *pantry*, *encounter*, *poignant*, *tyrant*. There are 650 such examples among the CELEX monomorphemes. (The large on-line dictionary CELEX, as described in Baayen, Piepenbrock, and Gulikers 1995, was the basis for the inventory of monomorphemic words used in Hay, Pierrehumbert, and Beckman, in press, and Pierrehumbert 2001b.) For /nf/, which is phonotactically worse than /nt/, there are only 38 examples among the CELEX monomorphemes. Examples include *influence* and *enforce*. For any of the words just cited, the nasal-obstruent template covers only a small part of the word. Lexical neighborhoods are understood to arise on-line during processing as words that are highly similar to the current word token are activated during perception or production. Thus, there is no evidence

that they involve any more abstraction than the abstraction involved in encoding and storing the word-forms themselves. Phonotactic generalizations, in contrast, may plausibly involve abstraction of a phonological grammar based on knowledge of all the words. Thus, Bailey and Hahn's (2001) results may be understood as showing that wordlikeness judgments do not tap a single level of representation in the system. Instead, the decision about whether a neologism is a possible word of English is reached by combining information from two levels of representation. The input from the lexical level is the extent to which the target word activates highly similar existing words. The input from the phonological level is the well-formedness of the word as determined by a stochastic phonological parse integrating its subparts.

Distinguishing lexical neighborhood from phonological effects is difficult, requiring the kind of exacting calculations and experiments that Bailey and Hahn (2001) have performed, because there is a high correlation between the size of a word's lexical neighborhood and its phonotactic likelihood as computed from its subparts. If a word has numerous neighbors, these neighbors will often have numerous shared subparts that all contribute to the counts for the subparts in a (stochastic) phonological grammar. However, the correlation is not perfect. A phonotactically probable word may have few neighbors if its various subparts are instantiated in nonoverlapping sets of other words. Even though each subpart would be found frequently in this case, no substantial fragment of the word would match a substantial fragment of any other word. Similarly, an improbable word may have fairly many neighbors if its rare subparts happen to co-occur in several other words.

Because of the difficulty of making this distinction in concrete cases, the existence of a phonological grammar as an explicit level of abstraction is in fact still controversial in psycholinguistics. Bailey and Hahn (2001) is one study that provides evidence for a phonologically abstract level. Experiments reported by Vitevich and Luce (1998) and Vitevich et al. (1999) show evidence separating lexical and phonotactic effects. Though phonotactic goodness and lexical neighborhood density are correlated in the lexicon, these researchers find that they have opposite effects in perception. Phonotactic goodness facilitates perception whereas dense lexical neighborhoods delay perception by creating competition. Furthermore, the role of the two levels is task dependent. The influence of the lexicon is reduced in tasks that are carried out at high speed and for which only sound pattern information is functionally relevant.

Such an outcome can only be modeled if the lexicon is distinguished from the phonotactics. Data on the psychological reality of obligatory contour principle (OCP) effects discussed by Berent and Shimron (1997) and Frisch and Zawaydeh (2001) also provide evidence for phonological abstraction of this particular constraint. In the light of such results, I will assume, following mainstream thought in linguistics, that an abstract phonological level is to be distinguished from the lexicon proper. I now turn to the properties of this level.

6.2.3 Words and Word-Level Phonology

Words are made up of phonological units, and the way in which units combine differs across languages. Phonotactics deals with how units combine to make words. Because the lexicon is an open set, as discussed in the introduction, the goal of phonological description is not the actual word list, but the larger set of possible words. A number of sources of evidence converge to indicate that there is a scale of possibility, which relates to the perceived likelihood of a whole word as a function of the frequency of its subparts and the specific way they were combined. All of the studies I will cite exemplify probabilistic knowledge of phonological constraints; however, not all authors eliminate lexical neighborhood effects as a possible artifactual source of their findings.

In an experimental study of nonsense CVC syllables, Treiman et al. (2000) constructed stimuli in which phoneme frequencies were controlled and the frequency of the rhyme (the VC) was varied. They found that both well-formedness judgments and outcomes in a blending task reflected the rhyme frequency. Frisch, Large, and Pisoni (2000) collected data on the wordlikeness of two- to four-syllable nonsense words made up of either high-frequency or low-frequency CV syllables. The ratings were a cumulative function of the frequencies of subparts. Hay, Pierrehumbert, and Beckman (in press) investigated the perception and judged acceptability of trochaic nonsense words containing medial nasal-obstruent clusters. They found that transcription errors disproportionately corrected infrequent clusters to (acoustically similar) more frequent ones. Acceptability was a gradient function of the likelihood of the (statistically) best morphosyntactic parse of the words. Nasal-obstruent constraints and an OCP constraint on strident fricatives both affected acceptability, and these effects interacted cumulatively to determine a given form's score. Munson (2001) found that phonotactic likelihood affects judgments by both adults and elementary-school children. It

affects error rates in production by children but not by adults. Zamuner, Gerken, and Hammond (2001) replicated Munson's finding for adults and extended the production results to infants.

In the literature on the psychological reality of phonotactic regularities, effects by now have been found for a wide variety of phonological templates. Treiman et al. (2000) report probabilistic effects relating to the VC rhyme of CVC syllables. Hay, Pierrehumbert, and Beckman (in press) report effects relating to a C.C syllable juncture in English, and Cutler and Otake (1996, 1998) report effects relating to the same position in Japanese and Dutch. Frisch and Zawaydeh (2001) and Frisch et al. (2001) have demonstrated the reality of OCP effects on the triconsonantal roots of Arabic, the consonants in these roots being separated by various vowels in the various words in which the roots appear. (In these studies of Arabic, phonotactic effects and neighborhood effects are distinguished.) Cutler and Butterfield (1992) have demonstrated effects in speech perception of the strong but not absolute constraint that English words begin with a stressed syllable. Frisch, Large, and Pisoni (2000) report a cumulative effect of the likelihood of CV syllables when combined with each other in two- to four-syllable nonsense words.

All of these phonotactic patterns are readily described using the apparatus of autosegmental and metrical phonology. A combined autosegmental/metrical formalism (as is developed in Pierrehumbert and Beckman 1988) permits all of them to be understood simply as fragments of phonological description. These is no single privileged level of analysis, and the fragments crosscut each other in the sense that they do not stand in any fixed hierarchical relationship. Taking the syllable as a kind of mental reference point, note that the list I have just given includes patterns that are bigger or smaller than a syllable (the word stress and the syllable rhyme); syllables that happen to be diphones; syllable junctures, containing just the end of one syllable and the beginning of the next; and consonantal projections that abstract across variations in syllable structure.

These observations are at odds with a thread in psycholinguistics that has sought a privileged unit of analysis. For example, in the (subsequently updated) speech production model of Levelt (1989), the syllable is the unit of production. This idea makes it hard to understand why some complex syllables such as /zɪmp/ are readily produced and judged acceptable even though they do not occur in the lexicon at all, at least as indicated by a search of CELEX. However, the observations fit in readily

with the framework of Data-Oriented Parsing (DOP) as described by Bod (1998). In DOP, as developed for automatic grammar acquisition in syntax and semantics, all partial formal descriptions up to some threshold of complexity are projected from any experienced utterance. Frequency counts for each descriptor are incremented as more and more utterances are encountered, and these frequency counts are exploited in scoring alternative parses of novel incoming forms. Compared to syntax, in fact, phonology seems like a very attractive forum for DOP because the phonological grammar lacks recursion and the known constraints appear to have relatively simple formal structure. In section 6.4, I will return to the question of what constraints are possible in relation to the issue of what constraints are statistically learnable from a lexicon of realistic size.

6.2.4 Morphophonological Correspondences

The last level of representation introduced above encompasses generalizations about relations between or among words. This is the level of morphophonological correspondences. Experimental studies of morphophonological generalizations indicate that their psychological reality and productivity depends on type frequency—for example, the number of different word-pairs that instantiate the generalization. Cena (1978) reports that the productivity of the English Great Vowel Shift (as in *serene* : *serenity*, *cone* : *conic*) depends on the specific vowel pair involved, being productive only for the more frequent vowel pairings. A study on conjugation of novel verbs in Spanish by Bybee and Pardo (1981) found that conjugation patterns exhibited only in a few high-frequency verbs failed to generalize; those exhibited in more than six different mid-frequency verbs did generalize. Bybee (1995b) discusses the role of type frequency in the formation of German participles for verbs and plurals for nouns. She shows that the default participial form of German is the most frequent one (contra Clahsen and Rothweiler 1992) and that type frequency also affects the productivity of the default German plural ending /s/.

It is known that children need to have acquired a critical mass of examples before they project a morphophonological generalization (Marchman and Bates 1994). However, alternations are not necessarily acquired in order of frequency. Since morphophonological generalizations are generalizations over word pairs, the cognitive availability of a given generalization depends not merely on the existence of the two words separately

in a child's lexicon, but also on the perception that they are related. For example, the words *bicycle* and *biscotti* both contain (diachronically) a common morpheme meaning 'two'; biscotti are twice-cooked. The relationship between these words goes unnoticed by many adult speakers, however. For a young child, the relationship between *imagine* and *imagination* might not be established, even if the child knew both words separately. In a similar vein, Raimy and Vogel (2000) argue that the notoriously late acquisition of the distinction between compounds such as *BRICK warehouse* (a warehouse for bricks) and phrases such as *brick WAREhouse* (a warehouse made of bricks) is actually due to the complexity and irregularity of the semantic relationships expressed.

The scientist seeking to model frequency effects in morphophonology faces a similar challenge in determining what word relationships should figure in the universe over which probabilities are established. There is no comprehensive formal treatment of what word pairings or word sets are relevant. Many of the clearest results have been obtained with inflectional morphology, in which the paradigmatic organization of the forms is less controversial than for derivational morphology.

6.2.5 Interim Summary

In summary, probabilistic effects are known to exist at all levels of representation of sound structure. These effects are perhaps most widely acknowledged in the area of phonetic encoding. Statistical classification models that originate with the foundational works of mathematical psychology can be adapted and extended to model how different languages utilize the phonetic space in different ways. A fairly large body of experimental work also reveals the existence of probabilistic implicit knowledge of word-forms in the lexicon. This knowledge is revealed both in speech processing and in tasks closer to the traditional purview of phonology, such as well-formedness judgments and allophonic outcomes. Following results in psycholinguistics, we can distinguish two aspects to this knowledge. One is the epiphenomenal knowledge of the lexical neighbors of a given stimulus, being the set of words that are so similar to the given word that they are activated in speech processing. The other is the long-term abstraction of the phonological grammar over the entire lexicon. Frequency effects are found at both levels. Probabilistic effects of word-form relationships are also observed in the area of morphological alternations and productivity.

6.3 Expected and Observed Frequencies

The productivity of phonology indicates that humans have internalized an abstract system for making complex forms from simpler parts, and linguists have the scientific goal of characterizing this system. Comparing alternative characterizations of the system reveals an inverse relationship between grammatical complexity of the system and the productivity it can describe. At one extreme of this relationship lies the simplest possible grammar, an arbitrary cross product of the phonological inventory (e.g., any combination of elements, in any order and of any length). This cross product provides the maximum number of forms that could be generated with the inventory of elements. At the other extreme lies the list of word-forms actually observed in the lexicon. That is, we could take the words in the lexicon as a set of positive grammatical constraints that license all and only the stated combinations of elements. The lexical list would provide the minimum number of outputs consistent with our data set to date, and it supports zero productivity in the sense that any thus far unattested word is taken to be impossible.

Obviously, the true state of phonology is somewhere between these extremes. Researchers attempt to propose specific grammars whose productivity agrees well with the extensions that humans make, and whose restrictions agree well with the forms observed to be absent. It is just as important to explain the impossible forms as to explain the possible ones. This enterprise requires a way to determine what forms are systematically absent; systematically absent forms reflect complications of the grammar, in comparison to the simpler grammar that would generate them. Comparing observed frequencies to expected ones provides the means for making deductions of this kind.

6.3.1 Learning, Inference, and Underrepresentation

The language learner is in many ways similar to a scientist, an analogy developed at length in the "theory theory" of Gopnik, Meltzoff, and Kuhl (1999). The learner is attempting to construct a synopsis of encountered forms that is general enough to be productive yet restrictive enough to rule out impossible forms. If the grammar is not general enough, the learner will not be able to process novel forms generated by other people or to express new meanings. If it is too general, the learner will generate novel forms that no one else can understand. Although the

learner may have specific perceptual and cognitive characteristics that differ from those of the scientist, this broad analogy means that general mathematical bounds on inferring grammars from data constrain the scientist and the learner in the same way.

A notorious issue for the language learner is the lack of negative evidence. Linguistic scientists face exactly the same difficulty. Although we can obtain negative judgments about well-formedness, these are a very limited tool. As discussed above, the cognitive status of well-formedness judgments is under dispute, since they are known to conflate effects at different levels. As sociolinguistic studies have shown, well-formedness judgments are vulnerable to perceived social norms and do not always conform to more naturalistic data. Even if the judgments were valid, there is no way we could obtain enough of them. No matter how many such judgments we obtain, they are a sparse sampling of the set of forms the grammar should rule out. Finally, linguists can't reliably produce forms that are impossible in their own language, and synthesizing the forms using a computer leaves no way to ensure that they exhibit the allophony they would have had if they were possible. Therefore, there is no reliable way to present impossible words to subjects in a way that guarantees they are encoded with the phonological representation the researcher has in view.

This situation has the result that statistical underrepresentation must do the job of negative evidence. Recent studies indeed show that children are quite sensitive to the statistics of sound patterns. By using statistical tools, linguists can also surmount the problem of negative evidence. It is also the case that underrepresentation has limits as a tool for deciding among alternative formal theories. We can turn these limits back on the theory and infer that there are some questions the language learner must not be asking, because they could not in principle be answered.

A phonological configuration that is systematically underrepresented is one that appears less frequently than would be expected if the grammar had no constraint disfavoring it. Thus, any claim about underrepresentation must be made within the context of a comparison between two grammars: a simpler grammar that provides the null hypothesis for the calculation, and the more complex and restrictive grammar whose status one wishes to evaluate. A choice for a simpler grammar that is always available is the generalized cross product described above. Under this grammar, each element has a frequency (estimated by its count relative to the size of some corpus), and elements combine at random. The expected

frequency E of any combination of elements P_1 and P_2 is accordingly the product of their two frequencies:

$$E(P_1P_2) = P(P_1) * P(P_2). \quad (1)$$

In practice, phonologists computing expected frequencies acknowledge that phoneme frequencies differ with position (in the syllable and/or in the word). That is, a partial phonological grammar is presupposed, and competition is between this presupposed grammar and a possible complication of it.

A major application of this reasoning is work on the OCP as it affects homorganic consonants in Arabic and other languages. Combinations of consonants at the same place of articulation are disfavored, the degree of underrepresentation being a function of the similarity and proximity of the consonants as well as of the specific language. McCarthy (1988) and Pierrehumbert (1992) both present results on this effect as it applies to the triconsonantal roots of Arabic verbs. In their calculations of E, positionally correct phoneme frequencies are used (C1, C2, or C3 position). Berkley (1994, 2000) presents similar results for English, French, and Latin. She overturns Davis's (1991) claim that /t/ is somehow exempted from OCP effects that apply to other stops in /sCVC/ configurations in English. Davis notes that forms like /spɛp/ are bad, whereas words such as *stats* actually exist. However, Berkley is able to show that the *stats* case is also underrepresented, suggesting a more uniform grammar than Davis proposed. Frisch (1996) also presents further calculations on Arabic that relate the degree of underrepresentation of consonant pairs in Arabic to the degree of similarity between the consonants as computed from a psychologically motivated metric. A gradient relationship is found, in which the proscription against totally identical consonants emerges at one end of a similarity continuum. These results obviate the need to posit two different grammatical mechanisms to capture the fact that the proscription against identical consonants is statistically stronger than the constraint against similar but nonidentical consonants.

Another application of the idea of expected value is found in Pierrehumbert 1994. Pierrehumbert computed expected frequencies of occurrence for all medial clusters of three or more consonants in monomorphemic English words (such as the /lfr/ in the word *palfry*), by assuming that these arise as random combinations of codas and onsets. The positional frequencies for codas were roughly approximated by the frequencies of the various consonants in word-final position once

appendices were stripped off; the positional frequencies for onsets were approximated by frequencies of onsets in word-initial position. Pierrehumbert showed that the vast preponderance of long medial clusters are expected to occur less than once in any lexicon of realistic size. For example, although a medial cluster such as /lskr/ in the nonsense word *pelskra* is syllabifiable and violates no sequential constraints (cf. *dell*, *else*, *screw*), it is unattested in any monomorphemic words of English found in the Collins on-line dictionary distributed by the Linguistic Data Consortium. Because of its relatively rare subparts, its expected count in a dictionary of this size is less than one. Thus, its absence is expected under a syllable grammar that includes positional probabilities for phonemes.

In a probabilistic model, no complication of the grammar is necessary to explain such cases. If the grammar lacked positional probabilities, then the absence of these forms would need to be modeled by complicating the grammar with additional constraints. The same argument can be made for more than 8,400 additional long clusters that are unattested in monomorphemic words. Pierrehumbert also identifies a number of syllable contact constraints, with observed counts less than the values expected from random combination of codas and onsets. These include an OCP effect leading to underrepresentation of forms with a C1C2C1 cluster (e.g., the hypothetical *palfly*) and a constraint disfavoring coronal obstruents in word-internal coda position (e.g., **pirtfy*). An experiment involving two different judgment tasks showed that these systematically underrepresented patterns relate to psychologically real constraints.

6.3.2 Sample Size

Both the statistical analyses of the Arabic OCP and Pierrehumbert's analysis of the long medial clusters of English were carried out using large on-line dictionaries yielding data sets that are roughly similar in size to the total mental lexicon of an adult speaker (see also discussion below). An individual Arabic speaker probably does not know a great many more triconsonantal verb stems than the 2,674 found by Cowan (1979). The Collins on-line dictionary of English contains 69,737 phonologically distinct entries and is similar in its coverage of monomorphemic words to the CELEX dictionary. Data sets such as these can be viewed as random samplings of the larger hypothetical data set in which all the latent potential for lexical additions is actually instantiated. The constraints governing the larger data set must, however, be inferrable from a sampling of this set. This is the case because adult speakers have learned the

phonology of their language, and learning the phonology cannot require more word-forms than individual adult speakers know.

In general, the bigger the sample, the more confidently patterns can be established. As the sample size increases, increasingly complex patterns can be established with reasonable confidence. The samples for monomorphemic words (around 10,000 items) are relatively small in comparison to the complexity of hypotheses phonologists might entertain. This is even more true for constraints such as the Arabic OCP that are active on only a subpart of the vocabulary (e.g., the verbs, which participate in nonconcatenative morphology and hence have a highly salient projection of the consonantal tier). To appreciate the force of this observation, let us consider how large a sample is required to be confident about two different constraints of English, differing in their statistical force.

The first is the constraint that a triphonemic monosyllable must contain a vowel or vocoid in English (unlike in Berber, well known for permitting even obstruents in nuclear position). For the sake of exposition, I make the simplifying assumption that these monosyllables are all (in fact) of the form CVC, VCC, CCV. The probability that a phoneme is a vowel/vocoid in this universe is 1/3; the probability that it is not is 2/3. (Probabilities sum to one and in this case there are only two alternatives.) Our null hypothesis will be that phonemes combine at random, and so positional probabilities will not figure in the null hypothesis. Under the null hypothesis, the expected probability of a CCC form is $(2/3)^3 = .296$. Under the null hypothesis, we would need to look at three or four forms in order to expect to find one of form CCC. But if such a small item set fails to have any CCC forms, we would have very weak evidence that this outcome is impossible; on the average, one expects to get a 6 once on every six rolls of a die, but rolling a die six times without getting a 6 is not good evidence that the die has no 6. Specifically, the probability under the null hypothesis of obtaining no CCC forms in a sample of size n is

$$\left(1 - \left(\frac{2}{3}\right)^3\right)^n = \left(\frac{19}{27}\right)^n. \tag{2}$$

The bigger n gets, the smaller this probability is and the more confident we become that the null hypothesis can be rejected. In this particular case, the null hypothesis becomes rapidly disfavored as n increases. For $n = 9$ (a sample of nine triphonemic words, of which none prove to have the form CCC), the probability that the null hypothesis is true is already less than .05. For $n = 14$, $p < .01$ and for $n = 20$, $p < .001$.

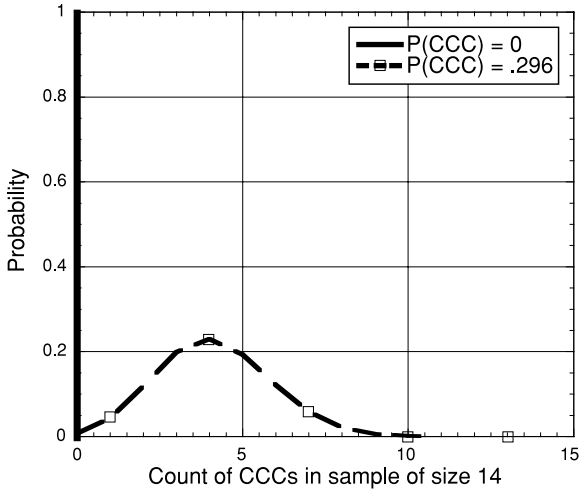

Figure 6.3
Probability distributions for $P(CCC) = 0$ and $P(CCC) = .296$, for a sample size of $n = 14$. Distributions computed directly by a binomial calculation.

Figure 6.3 illustrates graphically the comparison between the null hypothesis and the actual state of affairs, for the case of a 14-word sample. There are two probability distributions on this figure, the one on the right having the common lumpy shape and the one on the left consisting of a spike located on top of the y-axis. The distribution on the right shows the distribution of counts of CCC words under the null hypothesis. Obviously, this distribution peaks around 4. In a 14-word sample, we expect to find $(0.296 * 14) = 4.144$ CCC words. However, we could find more or fewer by the luck of the draw, as the figure shows. The distribution on the left represents the distribution of counts of CCC words under the hypothesis that the probability is 0. This is a degenerate situation in which the distribution has zero variance, because luck of the draw will never provide more or fewer than zero CCC words. The two distributions in this figure are well separated, exhibiting only a barely visible overlap at zero, where the left-hand distribution has a value of 1.0 (because all the probability is at a single location) and the right-hand distribution has a value of .007. The nearly perfect separation of the two hypotheses is what allows them to be distinguished in practice with such a small data set.

In this example, a very small amount of data causes confidence in the null hypothesis to deteriorate rapidly. That is because there was a huge

difference between the probability of a CCC word according to the null hypothesis, and its actual probability. When the quantitative difference between the two hypotheses is smaller, the rate at which increasing the sample size clarifies the inference is correspondingly less.

Fox example, let us consider how large a sample is needed to infer that the probability of /s/ in word-initial position differs systematically from its likelihood in word-final position in English. I will assume (idealizing somewhat) that the positional frequencies of /s/ in the CELEX monomorphemes are the true underlying frequencies for this comparison. (Counts in morphologically complex words could be misleading, because some words and word-level affixes, such as *-less* and *-ness*, show up repeatedly in complex words.) The probability of word-initial /s/ is .1153 in this set, and the probability of word-final /s/ is .0763. Note first that for a sample of nine words, we expect to find $0.1153 * 9 = 1.06$ cases of word-initial /s/, and $0.0763 * 9 = 0.6867$ cases of word-final /s/. Rounding to the nearest integer, a sample of nine words yields the same expected count of 1 word-initial /s/ and 1 word-final /s/. Thus, a sample size that already was decisive (on a typical significance threshold of $p \leq .05$) for the case of the *CCC constraint is too small for one to expect any difference for the positional /s/ frequencies. The situation is explored further in the three panels of figure 6.4, constructed along the same lines as figure 6.3.

Figure 6.4 illustrates two probability distributions: one for the counts of word-initial /s/ (on the left), and one for word-final /s/ (on the right). Because we are now varying the counts by two orders of magnitude across the panels of the figure, the x-axis has now been normalized by the sample size n. In figure 6.4(a), the sample size is 14 (the sample size for achieving significance level $p \leq .01$ for the *CCC constraint). The heavily overlaid distributions reveal how poorly discriminable the two cases are with such a small sample. In figure 6.4(b), the sample size is 140, and the overlap of the distributions is reduced. In figure 6.4(c), corresponding to a sample size of 1,400, the distributions are well separated, but still show some degree of overlap.

Thus, it requires at least two orders of magnitude more data to evaluate positional effects on /s/ than to discover the underrepresentation of CCC monosyllables. This situation comes about through two factors. The *CCC constraint is very general, since C is a broad category whose overall probability in the system is 2/3. /s/ is a more specific category that is instantiated less frequently. In general, the more specific a phonological

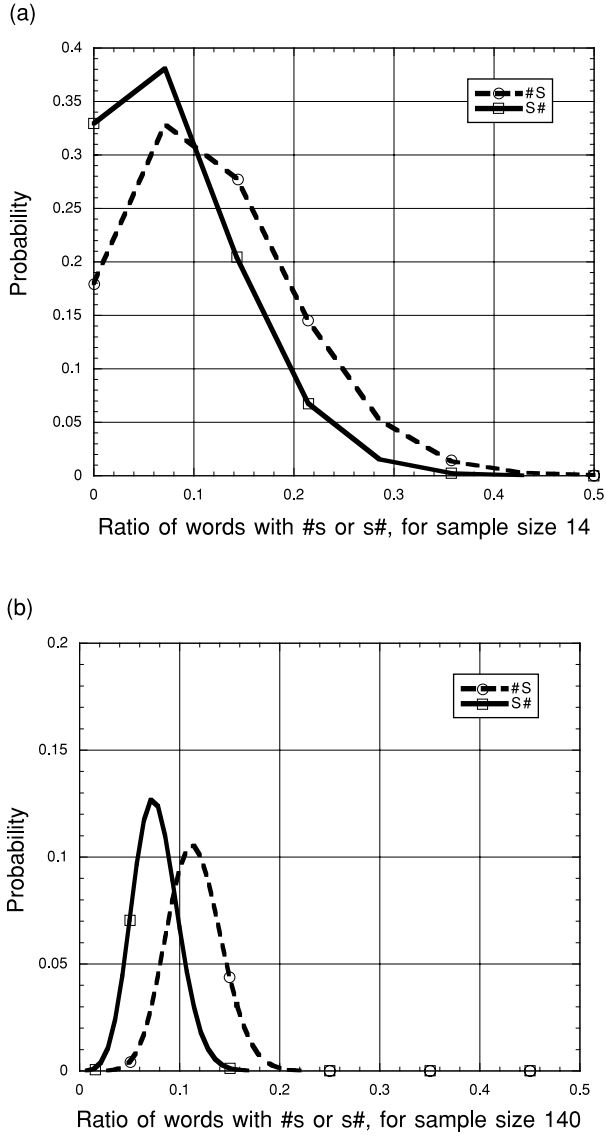

Figure 6.4
Probability distributions for $P(\#s) = .1153$ and $P(s\#) = .0763$, for sample sizes of (*a*) $n = 14$, (*b*) $n = 140$, and (*c*) $n = 1{,}400$. Distributions computed directly by a binomial calculation.

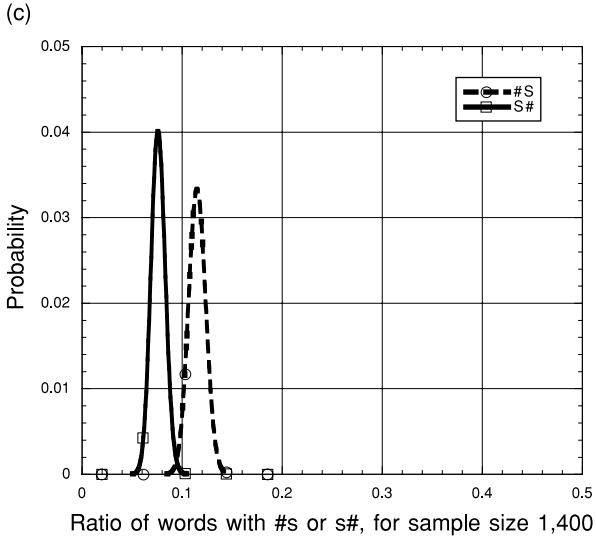

Figure 6.4 (continued)

description is, the fewer forms it will be true over and the more data will be needed to distinguish its probability from the probability of any of its close formal relatives. Also, in the *CCC example, the null hypothesis of $P = .296$ differs greatly from the actual state of affairs to which we compare it. The positional effect on /s/ was relatively slight, the probability for final position being about 2/3 of the probability for initial position. A slight difference is harder to evaluate than a massive one.

It is important to keep in mind that the probability of an anomalous run of events under the null hypothesis will never be zero; it will only get smaller and smaller. The probability of throwing all heads on n throws of a fair coin is $(1/2)^n$, so a coin tosser could get a million heads in a row with a probability of one-half to the millionth power, which though small is greater than zero. If we imagined that people had vocabularies of infinite size, then no amount of data would suffice to prove that CCC monosyllables are truly impossible; we could only show that they are vanishing rare.

6.3.3 Vocabulary Size

Since people do not have vocabularies of infinite size, the sizes of actual vocabularies can be used to place a bound on the statistical resolution

that is either possible or useful. Estimation of vocabulary size is a difficult issue. Productive and receptive vocabularies differ. It is unclear which morphologically complex words are stored and which are derived on the fly. Also, undersampling is a problem even with large data sets. However, some useful reference points can be deduced. One important count is the number of monomorphemic (or root) words in an individual's vocabulary. This set provides the basis for phonological generalizations about within-word phonotactics and prosodic patterns. Lexical items containing internal word boundaries, such as *peppermint* or *matchless*, show word-boundary phonotactics even if they have idiosyncratic semantics. Data from Anglin (1993), as graphed by Vihman (1996), indicate that English-speaking first graders (age 6 to 7) have a receptive vocabulary of about 2,500 root words, and by fifth grade (age 10 to 11) this number has grown to approximately 6,000. According to a tabulation made in the Phonetics Laboratory at Ohio State University, the number of monomorphemic words in CELEX is 11,381. As discussed in Pierrehumbert 2001b, this is a quite comprehensive list which probably exceeds the monomorphemic inventory of any single individual. For phonological constraints over monomorphemic words, probabilities of less than .00005 are effectively zero (since the expected count in a sample of size 10,000 would be zero). Any phonological constraints that are learned early in life need to be learnable on much smaller data sets, a point to which I return below.

For other phonological processes, such as stress patterns and morpho-phonological alternations, a larger vocabulary is relevant. Anglin (1993) estimates that English-speaking fifth graders know 25,000 root words plus derived words and idioms. (Productive and unproductive derivation do not appear to be distinguished in this estimate.) A detailed study by Nagy and Anderson (1984) yields the estimate that printed school English includes approximately 89,000 word "families." This number includes as distinct entries words that are morphologically complex but that have moderately to extremely opaque semantics. For example, in this study, *collarbone* is distinct from *collar*. However, the authors assume that *senselessly* can be productively derived from *senseless* and *stringy* can be derived from *string*. The existence of 89,000 word families in the entire corpus of printed school English does not mean that any individual child knows all the words. Examples provided include words such as *solenoid*, *hornswoggle*, and *ammeter*, and it appears improbable that any single school child knows all of them; discussion draws attention to the ability to read materials containing some unknown words. The authors estimate

that the total number of words known by a high school senior ranges from 25,000 to 50,000. From figures such as these, we can infer that an adult's total vocabulary is an order of magnitude larger than the adult's vocabulary of monomorphemic words, but probably less than two orders of magnitude larger. If a statistical inference is so delicate that it cannot confidently be made with a lexicon of around 100,000 words, there is no way for learners to make it and its cognitive status is highly dubious. If a probability over this set is $<.000005$, it is effectively zero.

6.4 Discrimination and Robustness

We have just seen that a full-blown adult lexicon, though large, is still not large enough for grammatical constraints of arbitrary subtlety to be deduced from it. In practice, the limits on grammatical subtlety are much more severe. First, children learn much of phonology at a young age, before they have large vocabularies. Second, different children have different vocabularies, and adults likewise. The commonplace observation that the grammar can be learned from poor levels of language exposure amounts to saying that it can be learned from a severe downsampling of the maximal data set, and it can be learned equally well from different downsamples of the data set. The fact that people with different language exposure can end up with essentially the same grammar also means that language acquisition is resistant to outliers (or statistically anomalous events). A statistically anomalous hurricane may destroy a large city, but early exposure to the word *Ladefoged* does not destroy a child's stress system. These marvels of the acquisition system amount to the claim that language is statistically robust. In this section, I develop the idea of robustness further and show how it constrains the nature of the phonological system.

6.4.1 Robustness in Categorization

Statistical robustness has been most explored in relation to categorization of the phonetic space. Vowel inventories have provided a particularly fruitful subtopic because of the ease with which the relevant phonetic dimensions can be conceptualized and manipulated. Accordingly, I resume discussion of the perception and production of phonetic categories in connection with the /ɪ/-/ɛ/ example of figure 6.2.

In figure 6.5, idealized distributions for /ɪ/ and /ɛ/ have been plotted on the assumption that the underlying distributions are Gaussian and that

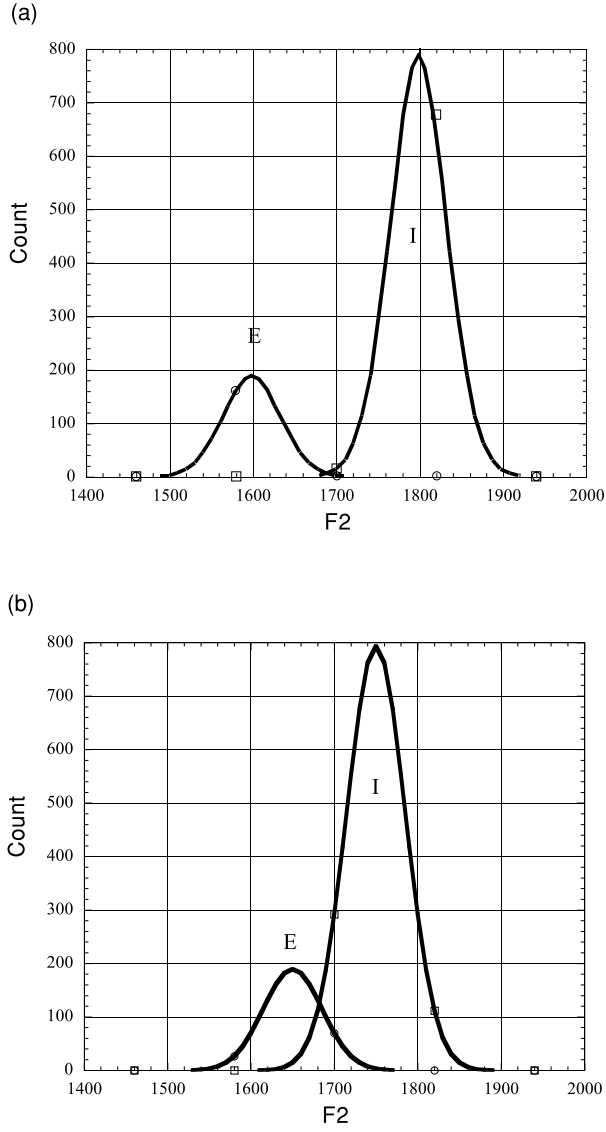

Figure 6.5
Relationship of two competing categories having idealized Gaussian distributions over the phonetic space. (*a*) 200-Hz separation of means. Low variance leads to little overlap of the distributions. (*b*) 100-Hz separation leads to more overlap. (*c*) Greater variance also leads to more overlap.

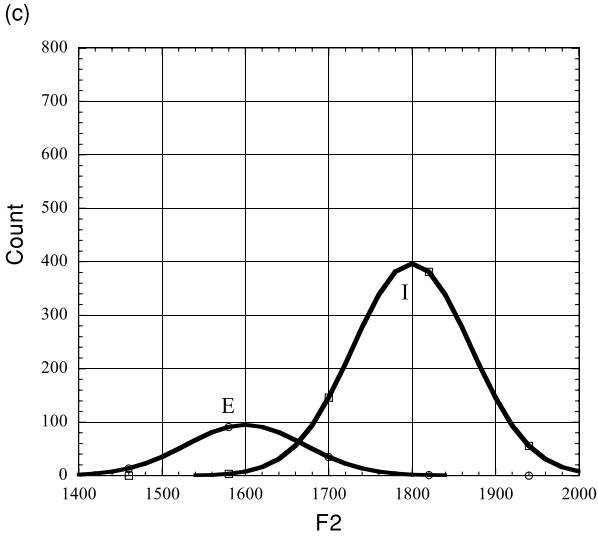

Figure 6.5 (continued)

both categories have been so abundantly experienced that the mental representations approach the true underlying distributions. Thus, the distributions in figure 6.5 are smooth instead of showing individual spikes for memory traces. Three cases are compared. In figure 6.5(a), the means of the distributions differ by 200 Hz and there is slight overlap between the two vowels. In figure 6.5(b), the distributions have been shifted so they differ by only 100 Hz. In figure 6.5(c), the means are as just separate as in figure 6.5(a), but each category has more variability so there is more overlap.

Since the distributions are smooth, the classification of an incoming token at any point can be read off the graph by seeing which distribution line lies above the other at that location on the x-axis. Figure 6.6 shows how this works by reproducing figure 6.5(c) with additional annotations. The listener is attempting to classify a stimulus at the F2 location indicated by the arrow. The line for the label /ɪ/ is above that for /ɛ/, so /ɪ/ is the most probable label. It is immediately obvious that given the distribution type under consideration, there is some F2 value such that any stimulus with an F2 value above this value will be classified as /ɪ/, and any stimulus with an F2 value below this value will be classified as /ɛ/. This cutoff is defined by the intersection of the two distributions and is

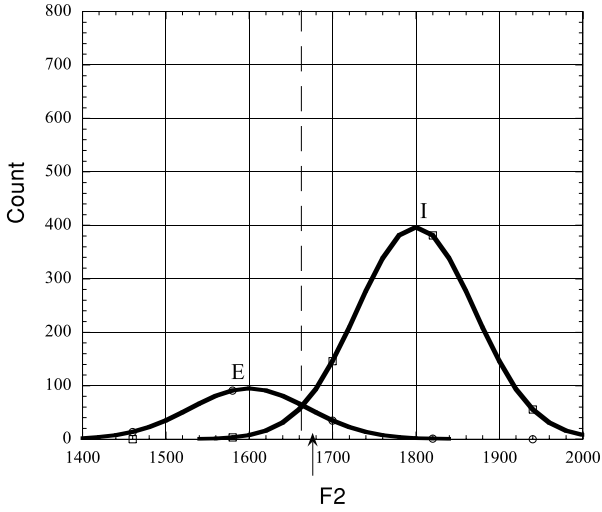

Figure 6.6
Figure 6.5(c), with the discrimination threshold indicated

shown with a vertical dashed line in the figure. Luce, Bush, and Galanter (1963a) provide a mathematical treatment of classification situations in which a threshold or cutoff is well defined.

For the situation shown in figure 6.6, the result is that a sizable region of the /ɛ/ production space will be heard as /ɪ/, and a certain region of the /ɪ/ production space will be heard as /ɛ/. The extent of such confusions would be exactly the same for figure 6.5(b). Although the separation of the means is half as much in figure 6.5(b) as in figure 6.6, the distributions are also narrower and the degree to which the distributions overlap is (by design) exactly the same. To achieve reliable discrimination, in which the intention of the speaker can be recovered by the listener with very little chance of error, we need the situation displayed in figure 6.5(a). Here, the separation of the means of the distribution is large *in relation to their width*.

Discrimination is not the same as statistical significance. As the quantity of data contributing to the distributions in figure 6.5(b) or 6.5(c) increases toward infinity, the difference in the means of these distributions can become as statistically significant as one could wish. That is, if the two labels are given, we can become more and more sure, scientifically speaking, that the distributions for these labels are not the same. But no

amount of certainty on this point improves the situation with respect to discrimination of the categories when classifying an unknown incoming token. In a similar vein, thanks to millions of test takers, it is extremely certain that the distribution of scores on the verbal portion of the SAT is slightly higher for women than for men. (The SAT is a standardized test administered to high school students applying to university in the United States.) However, since this difference is slight compared to the variation of scores for each gender, knowing someone's verbal SAT score does not permit one to deduce that person's gender with any reliability.

Statistical discrimination plays a central role in theories of the phonetic foundations of phonology, in particular pioneering work by Lindblom and colleagues on adaptive dispersion theory for vowels (see Liljencrants and Lindblom 1972; Lindblom 1986; Lindblom 1990). The model is predicated on the understanding that the language system requires robust discrimination of categories in order to guarantee the integrity of communication. Given that speech production and perception are intrinsically variable, this characteristic places bounds on the number of categories that can be maintained over the same space, as well as constraining their optimal arrangement with respect to each other. The effects are seen both in segmental inventories and in the way speakers manipulate the vowel space in continuous speech, making the minimal effort needed to maintain sufficient contrast. One consequence is that total phonetic space will be underutilized in any particular language, since accuracy of discrimination depends on statistical troughs like that in figure 6.5(a). Other researchers who have applied adaptive dispersion theory include Engstrand and Krull (1994), who demonstrate reduced durational variance for vowels in languages in which vowel length is distinctive, and Miller-Ockhuizen and Sands (2000), who use the model to explain phonetic details of clicks in two related languages with different click inventories.

The equations of adaptive dispersion theory treat perceptual discriminability (contrastiveness) synoptically as a direct pressure on productions. The same effects can also be approached in terms of how category systems evolve over time, because discrimination is related to the stability of the systems as they evolve. Assuming that distributions are used in a perception-production loop, as discussed in section 6.2, the classification of stimuli in perception provides data for the probability distributions controlling production. (The productions of one individual provide perceptual data for another individual, so that the perception-production

loop goes through members of the speech community.) Situations such as those depicted in figures 6.5(b) and 6.5(c) are intrinsically unstable over time, because in any region of the phonetic space where the distributions overlap, productions of the less likely label will be misclassified in perception.

Assuming only that the probability distributions are updated with perceptual data (an assumption needed to model how phonetic learning can occur in the first place), there are two major outcomes, which depend on the degree of distributional overlap in relation to the random variation in the perception and production processes. One outcome is that the two distributions sharpen up, each on its own side of the discrimination cutoff. The other is that the distribution for the less frequent label gets entirely eaten up by the distribution for the more frequent label, as each fresh misclassification enhances the latter's frequency advantage. This results in a collapse of the category system. Pierrehumbert (2001a) presents in detail the case of loss of contrast through a historical lenition process. A small but systematic lenition bias in production of the marked (less frequent) member of a contrast eventually results in its collapse with the more frequent member. Note that both of the major outcomes are characterized by a distinct distributional peak for each category label. Thus, consideration of the stability and robustness of categories over time provides a basis for the observations of adaptive dispersion theory, as well as observations made by Kornai (1998), Maye and Gerken (2000), and Maye, Werker, and Gerken (2002). Well-defined clusters or peaks in phonetic distributions support stable categories, and poor peaks do not. The same line of reasoning is convincingly applied in De Boer's (2000) modeling of vowel systems.

Such observations also have important implications for our understanding of the abstractness of phonetic encoding. The calculations presented by Lindblom, Miller-Ockhuizen, Kornai, Maye, Gerken, and Werker all deal with phonemes in a fixed context. Thus, they do not clearly differentiate phonemes (which are equivalenced across different contexts) from positional allophones. However, distributions of phonetic outcomes in continuous speech have revealed many cases in which the realization of one phoneme in one context is extremely similar or even identical to that of some other phoneme in another context. For example, Pierrehumbert and Talkin (1992) show that the distribution of vocal fold abduction (or breathiness) for /h/ overlaps that of vowels when both are tabulated out of context; when tabulated relative to context, there are just

a few sporadic cases of overlap. Overlap of devoiced allophones of /z/ with /s/ is discussed in Pierrehumbert 1993. Cases such as these mean that parametric distributions for phonemes are not always well distinguished from each other, if they are tabulated without regard to context. Within each given context, the distributions are much better distinguished. Thus, positional allophones appear to be a more viable level of abstraction for the phonetic encoding system than phonemes in the classic sense. Further implications follow from cases in which the phonemes are well distinguished in one context but not in another. These situations favor positional neutralizations. Steriade (1993) explores this concept in connection with stop consonants, demonstrating neutralization in contexts in which the stop release is missing. Flemming (1995) extends this line of reasoning in the framework of Optimality Theory.

Category discriminability also has important ramifications in sociolinguistics. An important set of studies (Labov, Karen, and Miller 1991; Faber and DiPaolo 1995) deals with cases of near-merger, in which the phonetic distributions of two categories have become heavily overlapped (owing to historical changes and/or dialect variation) but a statistically significant difference is still observable in productions. A case in point is the near-merger between *ferry* and *furry* in Philadelphia English, as discussed by Labov, Karen, and Miller (1991). A surprising behavioral finding is that subjects whose productions display an acoustic difference between the vowels of such word pairs are unable to distinguish these words at above chance levels. This is true even if they are listening to their own speech.

In the class of model I have been discussing, it is to be expected that discrimination in perception will be worse than the discrimination analysis that the scientist can carry out on objective acoustic measures. This is the case because of intrinsic noise in the perceptual system. The scientist can carry out a statistically optimal analysis on a microphone signal of the highest quality, whereas the perceptual system is limited by the critical bands of the auditory encoding, by the background noise from blood rushing through the ears, and so forth. However, it is also the case in this model that it is impossible to acquire a phonetic distinction that one cannot perceive. The subjects in the near-merger study must have been able to perceive a difference between *ferry* and *furry* at the time their production patterns were being established. Otherwise, the labeling and phonetic distributions needed to produce such a difference would never have been acquired. The model predicts that whenever a contrast becomes truly

imperceptible in a language, the contrast will collapse and this collapse will be irreversible. This prediction is borne out by Labov's (1994) assertion that total neutralizations are never reversed. Apparent cases of reversal can be traced to the survival of the distinction in some context. This can be a social context, as in cases in which a contrast is reimported into the speech community from a dialect or an influx of borrowings. Or the contrast may survive in some other context within the cognitive system, as when a contrast is reimported from a morphologically rich orthographic system. Similarly, detailed phonetic studies reveal that morphological relatives can serve this role through paradigm uniformity effects at the allophonic level. A case in point is Port, Mitleb, and O'Dell's (1981) study of incomplete neutralization of obstruent voicing in German.

The failure of Labov, Karen, and Miller's (1991) subjects to perceive a distinction in their own speech thus requires more explanation, assuming that this failure is more severe than perceptual encoding noise alone would predict. The levels of representation I have outlined provide an avenue of explanation. The task carried out by Labov, Karen, and Miller's subjects was a word judgment task, hence involved access to the lexicon. Thus, one must consider not only the phonetic encoding in perception, but also the relationship of this encoding to the stored word-form. If subjects have learned, from exposure to the varied dialect community of Philadelphia, that vowel quality information does not reliably distinguish *ferry* and *furry*, then they can downweight this information as a perceptual cue in lexical access. They learn not to pay attention to a cue they cannot rely on. This interpretation of the situation is borne out by a related study by Schulman (1983) on the perception of a *sit*, *set*, *sat*, *sot* continuum by bilingual Swedish-English speakers from Lyksele. The speakers were unable to hear the *set*, *sat* distinction when experimental instructions were delivered in Swedish, a language in which dialect diversity renders the distinction unreliable. However, they could hear the distinction when instructions were delivered in English. In short, this study indicates that the attentional weighting of different phonetic cues is not an entrenched feature of the cognitive system; instead, it can be varied according to the speech situation.

6.4.2 Robustness of Phonological Constraints

Statistical robustness is a well-established and major theme of the literature on phonetic encoding. Less widely acknowledged is the fact that it

also plays an important role at more abstract levels of representation. In fact, the small size of the lexicon, in comparison with the huge body of experienced speech, places severe limits on statistical inference over word-forms as well as correspondences between word-forms.

This issue is taken up by Pierrehumbert (2001b), in a set of pilot calculations addressing the issue of why known phonological constraints are so coarse grained. Though a DOP approach would permit us to construct a huge proliferation of phonological templates using the logical resources of phonology, the constraints that appear to be psychologically real by any test are relatively simple and nondetailed. Note in particular that they are simple in comparison to the lexical representations of individual words, as we have already seen from the comparison of lexical neighborhood effects with phonotactic effects.

The method used in this study was a Monte Carlo simulation of vocabulary acquisition; hypothetical vocabularies of different individuals at different levels of development were estimated by random sampling of the 11,381 monomorphemic words in CELEX. Since the goal was to discover the constraints on within-word sequences, morphologically complex words liable to contain internal word boundaries are excluded from this training set. The sampling was weighted by word frequency, on the assumption that an individual is more likely to learn a frequent word than an infrequent one. Inventories for 20 "individuals" at each vocabulary level—400, 800, 1,600, 3,200, and 6,400 words—were computed. The vocabulary level needed to learn a given constraint reliably was determined by inspecting these individual vocabularies to see whether the constraint was actually manifested for most or all individuals. The study compared the learnability of two real phonological constraints of English to that of two unrealistic constraints that are considerably more detailed.

The realistic phonological constraints of English are the constraint on foot alignment in the word and a constraint set governing word-medial nasal-obstruent clusters. The first refers to the fact that the basically trochaic foot structure of English is extended to trisyllables by aligning the foot to the left edge rather than the right edge of the word. The result is a 100 stress pattern (as in *parity*) rather than the 010 stress pattern found in some other languages. The treatment of nasal-obstruent clusters is taken from the detailed experimental study by Hay, Pierrehumbert, and Beckman (in press). This study, also discussed above, showed that the nasal-obstruent constraints are part of the speaker's implicit knowledge and

that perceived well-formedness is gradiently related to the frequency of the different clusters.

One of the unrealistic constraints evaluated in this study is a hypothetical constraint that combines the 100 stress template with the nasal-obstruent regularities. That is, the study asks whether it is reasonable for a language to have different nasal-obstruent constraints for trisyllabic words with a specific stress pattern. The other unrealistic constraint involves the statistical pattern studied by Moreton (1997). Moreton's experiment looked for a phonemic bias effect as a corollary of a difference in probability, comparing word-final stressed /gri/ as in *degree* and word-final stressed /kri/ as in *decree*. /gri/ is much more common than /kri/ in running speech, chiefly because of the high frequency of the word *agree*. There is also a contrast in type frequency between the two patterns, but it is much smaller. This example was selected because Moreton obtained a null result in his experiment—one of the few cases anywhere in the experimental literature for a negative finding on the psychological reality of a statistical phonotactic pattern.

The calculations showed that the 100 stress pattern is extremely robust, learnable by all 20 individuals from a mere 400 words. A rank order of five nasal-obstruent combinations could be reliably learned from 3,200 words, a vocabulary level realistically achievable by an elementary-school child (see above). The two unrealistic constraints were not reliably learnable even at a vocabulary level of 6,400 monomorphemic words. The fact that Moreton's subjects had not (apparently) learned the /gri/-/kri/ constraint is therefore unsurprising.

Pierrehumbert (2001b) views vocabulary levels in terms of stage in language acquisition. The same line of reasoning, however, leads to the prediction that adult speakers with different vocabulary levels should have somewhat different grammars. In particular, adults with large vocabularies should be in a better position to estimate low, but nonzero, probabilities. This prediction is borne out by Frisch et al. (2001), who found that adults with large vocabularies have a more generous threshold for viewing nonce forms as acceptable, and who present calculations showing that this effect, for their stimuli, cannot be attributed to lexical neighborhood effects.

A new series of calculations extends this line of reasoning by exploring phonemic *n*-phones more systematically. The nasal-obstruent sequences contain two phonemes, hence are diphones. Even disregarding the positional information, /gri/ and /kri/ are triphones. Bailey and Hahn's

(2001) study of lexical neighborhood densities and phonotactic statistics, discussed above, systematically evaluated the role of diphone statistics in comparison to triphone statistics. Diphone statistics were an important predictor of wordlikeness judgments, but triphone statistics did not add any predictive power beyond that provided by diphone statistics. The goal of the present calculations was to determine the extent to which diphone and triphone constraints on the internal form of words can or must be learned from the lexicon. If triphone constraints are not learnable as such, then the well-formedness of a triphone must be estimated from the well-formedness of its subparts. Under this assumption, the well-formedness of the /str/ sequence in *street* would follow from the well-formedness of /st/ and /tr/; the ill-formedness of /stl/ in **stleet* would follow from its ill-formed subpart /tl/. If triphones are generally and necessarily parsed in terms of their subparts, then triphone statistics would add no predictive power in a model of perceived well-formedness, just as Bailey and Hahn report.

The calculations were again made on the CELEX monomorphemes. The transcription set contains 37 phonemes, disregarding 11 examples of distinctively nasalized vowels in French borrowings. The transcribed distinction between syllabic and nonsyllabic sonorant consonants was also disregarded on the grounds that it is predictable from sonority sequencing. When these are sonority peaks in their sequence (as in *apple*), they are syllabic; otherwise not (as in *mill*). This means that the full cross product of the phonemes, providing the baseline against which constraints must be evaluated, generates 1,369 different diphones and 50,653 triphones.

Basic counts show the general feasibility of training word-internal diphone statistics from the inventory of extant monomorphemes. The 11,381 CELEX monomorphemes display 51,257 tokens of diphones. Thus, the data set is 37 times bigger than the constraint set that the learner is attempting to parameterize. Downsampling the data set to 3,200 words (the number around which nasal-obstruent constraints began to be reliably learnable) still yields a ratio of approximately ten data points per diphone, on the average. This is an order of magnitude more data than constraint set parameters, and it is equivalent to the count that permitted *CCC to be detected at $p < .05$ in the tutorial example above.

The learnability situation is quite different when we look at the triphones. There are fewer triphones per word than diphones. For example, a four-phoneme word contains two triphones (namely, those starting at

the first and second phonemic positions) but three diphones. Thus, there are only 39,877 examples of triphones in the training set. Meanwhile, the candidate constraint set has exploded to 50,653 different forms. This leaves an average of less than one data point per constraint parameter, even on the implausible assumption that the whole vocabulary is available to the young language learner. Even the number of triphones that actually occur at least once in the monomorphemic inventory—namely, 5,414—is large in relation to children's level of lexical knowledge. In short, it is mathematically impossible to assign probabilities to triphones generally, on the basis of knowledge of monomorphemic words. If the general assumptions presented here are valid, then it should be possible to predict the well-formedness of triphones from more general constraints, which are learnable from the available data.

This raises the issue of how well triphones may be predicted from their diphone subparts. We can look at this issue in two ways. First, we compute the expected count of each triphone in an inventory the actual size of the training set (see table 6.2). To do this, for each triphone P1P2P3, we append P3 to P1P2 with the conditional probability of P3 following P2 in a diphone. If this expected count (rounded to the nearest integer) is at least one, then the triphone is expected to occur. If it rounds to zero, the triphone is expected to be absent. The table breaks out these counts according to whether the triphone is or is not exemplified in the data set. This prediction is quite successful. Out of 45,239 triphones that fail to occur, 42,776, or 95%, are predicted to be absent from the rarity of their subparts. Incorrect predictions constitute only 6.6% of the data. Not revealed by the table is the fact that these cases are overwhelmingly ones in which the count was predicted to be zero and is actually one, or vice versa—in short, cases that sit right on the threshold applied in setting up the table.

A way of viewing the data that overcomes the thresholding issue and also provides more insight into relative well-formedness is to plot the actual rate of occurrence of triphones against their expected rate of occurrence, as shown in figure 6.7. In figure 6.7, the triphones have been ranked by expected count and 10 bins of expected count are established. Within each bin, the median actual rate of occurrence is plotted along with upper and lower quartiles. Overall, the actual rate of occurrence is shown to be strongly related to the expected rate of occurrence. The fact that the relationship is not perfect arises both from sampling issues (the actual lexicon can be viewed as a sampling of the potential lexicon) and,

Table 6.2
Existence and absence of triphones in a data set in relation to expected counts

	Absent	Exist
Predicted to be absent	42,776	892
Predicted to exist	2,463	4,522

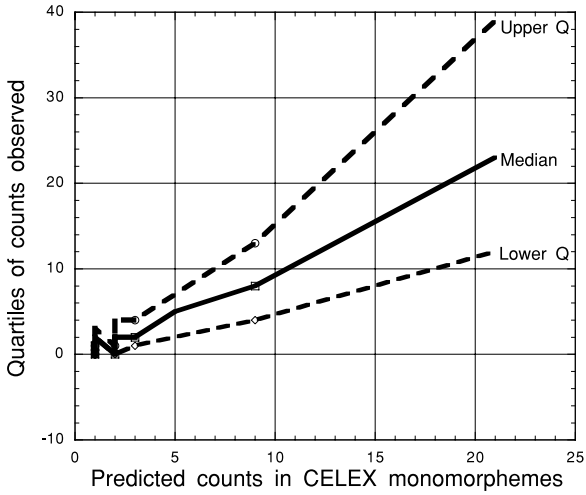

Figure 6.7
Quartiles of observed counts for triphones, plotted against median count expected from diphone frequencies

more importantly, from the fact that triphones are subject to additional constraints beyond the biphone conditions. An example is the OCP constraint disfavoring C1C2C1 combinations, as discussed above.

Comparing the diphone situation to that of triphones reveals that it is not only possible, but also necessary, to learn constraints on diphones rather than estimating their likelihood from their (phone) subparts. Table 6.3, constructed along the same lines as the triphone table 6.2, shows that only a few of the absent diphones are predicted to be absent on the basis of phone frequencies alone. Of the 1314 diphones that are predicted to exist under the null hypothesis, 44% are unattested. This outcome is not surprising, since the null hypothesis does not take into account the pattern of sonority alternation that imposes a syllabic rhythm on the speech stream. One consequence is that a pair of identical high-frequency

phonemes is predicted to be frequent: because of the high frequency of /ɪ/, the sequence /ɪɪ/ is expected to be the most frequent combination of all, whereas in fact it is impossible. Most of the numerous unexplained absences in this table arise from factors that a linguist would view as systematic.

For attested diphones, phoneme frequencies are also rather poor at predicting the rate of occurrence. This is indicated in figure 6.8, constructed along the same lines as figure 6.7. Though diphones containing frequent phonemes are on the average more frequent, the spread around this trend is much greater than in figure 6.7. In particular, the lower quartile remains almost at zero all the way up to the highest expected count.

In the absence of additional evidence, it is not possible to conclude that all triphones are necessarily evaluated via their subparts. The highest-frequency triphones have counts in the three hundreds, amply sufficient to support an evaluation of their observed frequency relative to the expected frequency from a diphone grammar. It is possible that some triphones do acquire probabilities in their own right. The initial triphone of *advent* occurs in 7 words despite having a predicted rate of occurrence of zero. The final triphone of *raft* occurs in 13 words despite having a predicted rate of occurrence of one. Possibly, people may have some implicit knowledge of these facts; but it would be difficult to demonstrate that any such knowledge goes beyond lexical neighborhood effects. However, in inspecting the list of predicted and attested triphones, it was surprisingly difficult to find these examples, as cases of blatantly overrepresented triphones are few. The mere assumption that the lexicon is a random sampling of a bigger universe of possible words means that it would exhibit some random variation in the over- and underrepresentation of particular combinations. It is not known whether cases of this type are treated as within-the-noise by the cognitive system, or whether they are remembered as important.

The fact that triphones are an unpromising field for large-scale phonological constraints also points up the importance of investigating large-scale constraint candidates that are closer to the fruits of linguistic scholarship. In the linguistic literature, constraints with a larger temporal scale (such as vowel harmony constraints and foot structure constraints) generally have the property of referring to classes of phonemes rather than to individual phonemes, if they refer to phonemes at all. As far as trainability goes, an increase in temporal scale is offset by a decrease in

Table 6.3
Existence and absence of diphones in a data set in relation to expected counts

	Absent	Exist
Predicted to be absent	48	7
Predicted to exist	582	732

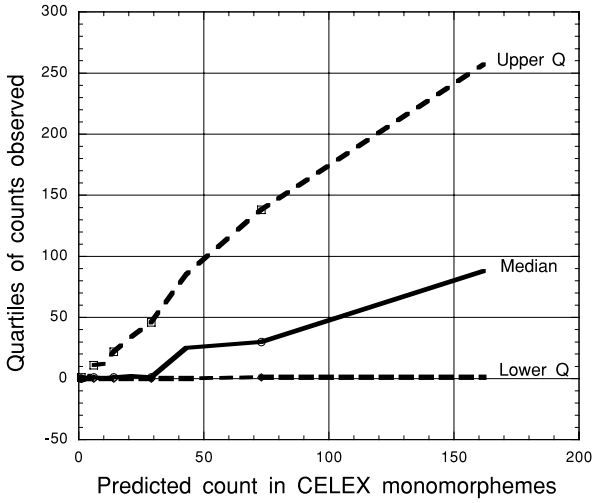

Figure 6.8
Quartiles of observed counts for diphones, plotted against median count expected from phoneme frequencies

featural specificity. As the results on 100 versus 010 stress patterns show, such constraints can be extremely robust.

The nonlearnability of triphones at the level of the phonological grammar contrasts sharply with the amount of detail people learn about specific words. Even the most categorical description of a good-sized word, such as *ambrosia* or *budgerigar*, is highly complex compared to any known phonological constraint, indicating that brute cognitive limits on representational complexity are not the issue. Moreover, cases discussed above in relation to word-specific allophonic detail and subphonemic morphophonological paradigm effects show that the representations of words can be extremely detailed. This detail can be acquired because reasonably frequent words are encountered so many times in speech. Counts in Carterette and Jones's (1974) tabulation of conversational

speech by adults and children show that the most frequent triphonemic function words (such as *was* and *what*) are up to 50 times as frequent in speech as triconsonantal syllable onsets such as /spr/. A word with a frequency of six per million (a threshold above which word frequency effects are experimentally established, even for morphologically complex words) would show up approximately once per 15 hours of speech, assuming an average rate of three words per second. Athough people can learn the existence of new words from very few examples, once they have learned a word, they have plenty of evidence to fine-tune its representation.

6.5 Correlations across Levels

The previous discussion emphasizes the need to distinguish levels of representation, each with its own statistical effects. This is reminiscent of the modularity claims of classical generative theory. However, there are important differences. In the generative approach, the task of setting up modules is viewed as a task of partitioning the system into maximally independent components. Each module is as stripped down as possible. In evaluating specific theoretical proposals, any observed redundancy between information encoded in one module and information encoded in another is viewed as a weakness. However, it has proved impossible to eliminate redundancy across modules. The present approach, in contrast, uses rich representations (as Baayen, this volume, also discusses). Correlations across levels of representation are viewed as a necessary consequence of the way that more abstract levels are projected from less abstract levels. Some of these correlations are extremely obvious and would show up in any current approach. For example, in essentially all current approaches, the phonological entities manipulated in the grammar and morphophonology are viewed as contentful; their behavior reflects their general phonetic character as it percolates up the system through successive levels of abstraction.

Some other examples of confluence across levels are far less obvious and reveal subtle properties of human language.

One nonobvious correlation concerns the relationship of the phonetic encoding system to the phonological grammar. It is known that children can use statistical cues to decompose the speech stream into chunks at a very early age, before they know that words refer to objects and events (see Jusczyk, Luce, and Charles-Luce 1994; Mattys et al. 1999). Low-frequency transitions are taken to be boundaries, and these transition

frequencies must evidently be estimated using surface statistics, since type statistics would depend on a lexicon, which the child has not yet formed. Luckily, these surface (or token) statistics on diphones are highly correlated with type statistics in the lexicon. In general, phoneme combinations that are infrequent in running speech are also infrequent word-internally in the lexicon. This means that a low-frequency phoneme transition is a viable cue to the possible existence of a word boundary. As Hay (2000) points out, it is possible to design a formal language in which this correlation does not obtain. Imagine, for example, that a language had an invariant start phoneme for words (say, /t/) and an invariant stop phoneme (say, /k/), so that every word began with /t/ and ended with /k/. In this case, the combination /kt/ would be extremely common in running speech, but would nonetheless be the cue to the presence of a word boundary. Human languages do not have this property. Instead, they show maximal contrast sets in word-initial position, and the proliferation of alternatives at word onsets leads to low-frequency combinations across the word boundary. The confluence between type frequency and token frequency thus supports bootstrapping of the system during language acquisition and smooths the communication between levels in the mature system.

A second important case of confluence brings together the issues in phonetic categorization and *n*-phone statistics that were discussed above. A substantial body of work in phonetics shows that the acoustic landmarks provided by transitions between the physical regimes of the vocal tract play a key role in forming discriminable categories. For example, at an obstruent-vowel transition, an aerodynamic factor (the buildup of pressure during the stop closure) conspires with an acoustic factor (a jump in the number of resonances as the vocal tract moves from a closed to an open position) and a psychoacoustic factor (the way the ear adjusts its sensitivity to changes in amplitude). These factors together mean that obstruent-vowel transitions are perceptually salient and reliably distinguished from each other, a point developed in more detail in acoustic landmark theory (Stevens 1998). Steriade (1993) provides a typological survey showing the ramifications of this phonetic situation with regard to stops. Transitions between segments are also important in production, since coproduction of adjacent segments can result in substantial modifications of their realizations, including occlusions and gestural reorganizations. Such phenomena are one of the main topics of Articulatory Phonology as developed by Browman and Goldstein (1986, 1992).

The connection made in Steriade's work between phonetic information and phonological sequencing is made in a more broad-based way by the importance of diphone statistics in phonotactics. Any phonological diphone implicitly defines a phonetic transition, which has the potential to be a dynamic cue in speech perception. Some of these transitions are very robust from an encoding standpoint, and others are fragile. Fragility can arise either because the transition lacks perceptual landmarks or because it lies in an unstable region of the phonetic control space and yields excessively variable outcomes. Thus, there is as much reason for a language to selectively utilize the universe of possible transitions as to selectively utilize the phonetic space in its phoneme inventory. The learnability of diphone statistics thus delineates a confluence among the phonetic space, the inventory size, the size of the lexicon, and the complexity of the grammar.

Much recent work in stochastic Optimality Theory takes a different stance on correlations across levels of representation. For example, Flemming's (1995), Kirchner's (1997), and Boersma's (1998) response to correlations has been to dissolve the boundaries between levels by treating all effects through a single Optimality Theory grammar with a single constraint ranking. The Gradual Learning Algorithm (GLA) of Boersma (1998) and Boersma and Hayes (2001) is the most mathematically elaborated model of this type, and so it is the one I will discuss here.

The conceptual core of stochastic Optimality Theory is morphophonological alternations that are by-products of general phonological constraints. A landmark paper by Anttila (available in 1994 and published in 1997) undertook to explain variability in the form of the Finnish genitive plural as a response to constraints on metrical prominence, word stress, and metrical alternation, all of which are active in the language generally. Anttila hypothesizes that the variability in outcome for the genitive plural arises because some constraints are unranked. On each individual instance of word production, a specific constraint ranking is imposed through a random selection of ranking for unranked constraints. Depending on which constraint ranks ahead of the others on that specific trial, one or another of the surface goals takes priority in selecting the form on that particular occasion.

A major extension of this approach is the GLA model of Boersma (1998) and Boersma and Hayes (2001). In this model, constraints are ranked on a real-valued scale. Each individual constraint has a Gaussian distribution of ranking values on this scale. The mean of the distribution

is incrementally learned through language exposure, and the variance is taken to be constant. The exact nature of the training algorithm has important repercussions for the overall behavior of the model. After a careful evaluation of statistical robustness and stability, Boersma proposed a training algorithm in which each individual extant word causes downward adjustment of all constraints it violates. The result is that the model is much closer to a standard stochastic grammar than would at first appear; phonological templates that are abundantly instantiated in the training set end up being highly favored by the grammar, and those that are poorly instantiated end up being disfavored. Thus, the rankings of constraints closely track the frequency values that would be assigned to the same constraints in some stochastic grammars.

This model permits much finer tracking of probability distributions for different outcomes than Anttila's model was capable of. Indeed, the model has such general mathematical power that the perceptual classification phenomena introduced in section 6.2 can be expressed in the model by positing large (in fact, arbitrarily large) constraint families that describe arbitrarily fine regions of the parametric space. Fine-tuning of constraint rankings has the effect of performing the metric calculations of a standard psychoacoustic model. Conceptually, then, phonetic encoding is raised into the phonological grammar by treating the parametric phonetic space as if it were a very large set of categories.

Morphophonological correspondences are also folded down onto the same constraint ranking. The need to express paradigm uniformity effects is generally accepted in Optimality Theory and is addressed by correspondence and sympathy constraints (see McCarthy and Prince 1995; McCarthy 1999). A more challenging case is presented by morphophonological correspondences that are unnatural in the sense of Anderson (1981). A case in point is the alternation of /k/ with /s/ in English word pairs such as *electric*, *electricity*, *electricism*. Though this alternation (velar softening) has a historical basis in a series of phonetic natural reductions, it is not natural as it stands. If /s/ were in any sense an unmarked pronunciation of /k/ before a schwa, then it should be more frequent than /k/ before schwa generally. According to CELEX, however, /s/ is less than half as frequent as /k/ before schwa in the lexicon as a whole. Furthermore, any phonetic generalization that pressured /k/ in the direction of /s/ should also tend to fricate /t/. However, in the cases in which /t/ appears before one of the triggering affixes, it remains a stop: *magnet*, *magnetism*, *Jesuit*, *Jesuitism*, *mute*, *mutism*.

The /k/∼/s/ alternation is nonetheless productive, as we would predict from statistics over the relevant universe. CELEX shows 72 word-pairs involving a base form ending in /k/ with a related form ending in *-ism* or *-ity*. Velar softening is found in all of them. Such a perfect regularity is expected to be extended, and a recent pilot experiment indicates that this is the case. Subjects were led to believe that the experiment concerned the (semantic) choice among various affixes that turn adjectives into abstract nouns. They completed 18 discourse fragments such as the following, involving affixation on a novel stem. Baseline sentences and fillers were also included.

(3) Janet is criotic about environmental issues. Her ????? manifests itself in avid involvement in environmental groups.

All six subjects softened the /k/ to /s/ in every single case in which they selected the suffix *-ity* or *-ism* over the semantic competitor *-ness*.

This outcome can be captured in the GLA model in the following way. (I am grateful to Paul Boersma (personal communication) for suggesting the specifics of this analysis.) A constraint disfavoring /k/ before *-ity* and *-ism* becomes extremely highly ranked as the learner encounters words such as *electricity*. At the same time, Universal Grammar is presumed to supply a universal set of constraints disfavoring replacement of any phoneme with any other phoneme (e.g., disfavoring phonemic changes for a full cross product of the phonemes). The constraint disfavoring replacement of /k/ with /s/ comes to be ranked low as the learner encounters words in which the replacement has occurred. /t/ is unaffected, since the key constraint targets /k/ only, not voiceless stops in general. Once the constraint rankings have been learned, the same replacement will occur in any novel form.

Thus, the price of folding unnatural morphophonological correspondences into the phonological grammar is splitting the correspondences into unrelated constraint pairs, which are ranked separately. Probabilities are not directly encoded on correspondences; rather, they indirectly affect the state of the grammar through incremental training.

Thus, the GLA model is very powerful. It can encode statistical regularities at all levels, from phonetic encoding up through morphophonological correspondences. For regularities that are either more or less abstract than the level of its core strengths, some researchers might find the encoding to be indirect and inperspicuous. In particular, the treatment

of phonetic encoding appears to eschew the well-established resources of mathematical psychology.

In the GLA model, all constraints are at the same level and all constraint rankings are trained on the same data set. Thus, the connection drawn above between the granularity of constraints and the size of the effective training set does not appear to be available. In the model presented above, the probability distributions for different parts of the system are established directly from experience, and thus non-Gaussian distributions will be automatically discovered. Phonetics, like many other physical processes, provides examples of skewed distributions relating to physical nonlinearities and saturations (see, e.g., Duarte et al. 2001, where it is shown that the distribution of consonantal interval durations in speech is skewed and obeys a gamma distribution, for many different languages). In contrast, distributions arising from repeated independent decisions (as in coin flipping or a forced-choice experiment) tend to be Gaussian. Since the assumption of Gaussian distributions is critical to the mathematical tractability of the model, the existence of non-Gaussian distributions appears to be problematic. The GLA model also does not distinguish effects relating to type frequency from effects relating to surface, or token, frequency. It also provides no way to downweight the grammatical impact of extremely frequent words, as Bybee (2001) and Bailey and Hahn (2001) show to be necessary.

In the presentation above, I have suggested that well-formedness judgments represent a decision with weighted inputs from two levels, the lexicon and a score established by the phonological grammar as the likelihood of the best parse. Well-formedness judgments come about differently in the GLA model. Boersma and Hayes (2001) propose that they arise from the likelihood that the given word would emerge as such under repeated runs of the grammar. The word is judged as poor if it would often be modified to some better form on repeated trials. This kind of virtual reality calculation is available in a closed form (e.g., without actually running the grammar many times) because of the simplifying assumptions of the model. The idea is applied with some success in an experiment on morphophonological alternation in Tagalog reported by Zuraw (2000).

This assumption is problematic for phonotactic judgments, which have made up most of the literature. Three experiments cited above show high accuracy rates for some sequences that are relatively infrequent and are

judged as poor. The transcription data reported by Hay, Pierrehumbert, and Beckman (in press) revealed that rates of correction of unusual clusters depended both on the cluster frequency and on the existence of an acoustically similar competitor. Rare clusters without a similar competitor were not corrected often, even though they were judged as poor. In an imitation experiment, Munson (2001) also found that error rates in adult productions were not significantly different for infrequent and frequent clusters, though frequency did affect wordlikeness judgments. This outcome also occurred in the adult baseline data for Zamuner, Gerken, and Hammond's (2001) acquisition study. Results such as these follow from the assumption that small phonetic effects can pile up over time in shaping the lexicon, which in turn shapes the grammer. With the lexicon standing between the phonetics and the grammar, it is possible for low rates of phonetic instability to coexist with strong lexical statistics. The view of the lexicon presented here also allows lexical neighborhood effects to affect well-formedness judgments. The observed combination of factors is not captured in the GLA model, in which well-formedness judgments are based on the grammar alone.

I have discussed the GLA in this much detail because it is by far the most coherent and comprehensive proposal to fold effects at different levels, from phonetic encoding up through morphophonological correspondences, into a single grammar. Its successes provide further evidence for the importance of stochastic grammars and robust learning algorithms to our understanding of phonology. Its specific weaknesses reveal the general weaknesses of responding to confluences across levels by conflating them.

6.6 Conclusion

In conclusion, entities at all levels of representation in phonetics and phonology display statistical variation. A wide assortment of behaviors reveals that speakers have implicit knowledge of this variation. It is relevant to speech processing, where it affects perceptual classification as well as speed and accuracy in perception and production. It is also reflected in long-term properties of the system, such as allophonic outcomes and the compositionality of complex patterns from subparts.

For any level in the system, we must consider not only its overall probability of occurrence, but also its probability distribution over the less abstract level that gave rise to it. Each category of phonetic encoding

has both a total rate of occurrence and a distribution over the parametric phonetic space. The availability of both is an automatic feature of exemplar theory, in which empirical distributions are built up incrementally from experienced tokens of speech. The range and likelihood of various phonetic realizations are revealed by the local density of memory traces on the parametric space, and the frequency of the category as a whole is revealed by the total quantity of traces. Analogous effects are found at higher levels, with word-forms also having probability distributions over phonetic outcomes, and phonological constraints having probability distributions over the space of word-forms.

Comparison of probabilities plays a crucial role in scientific inference and language learning. Both scientists and language learners posit a more complicated grammar only if systematic deviations from the output patterns of a simpler grammar are observable. The level of language exposure thus places bounds on the complexity of inferences that can be made. Children should be able to make gross inferences about the phonological system before they make subtler ones, and even for adults, the subtlety of inferences that are cognitively viable is limited by the size of the data set to which the generalization pertains. In particular, thanks to the tremendous volume of speech that people encounter, fine details of allophony can be learned as well as a large number of word-specific properties. Because of the much smaller size of the lexicon, general knowledge of words is more coarse grained. Thus, a probabilistic framework allows us to make inferences about the utilization of the phonetic space, and the possible constraint set in phonology, in a way that is not possible in a purely categorical approach.

The starting point of this discussion was the claim that cognitive entities have probabilities and probability distributions and that comparisons of probabilities are involved in comparing alternative models of the whole system. Comparisons of probabilities also play a role in processing, when there are two alternative analyses of the same form. In phonetic encoding of speech events, each event is categorized by finding the most probable of the competing labels. In parsing speech into words, a relevant comparison is the likelihood that a given sequence is word internal versus the likelihood that it bridges a word boundary. For example, in Hay, Pierrehumbert, and Beckman's (in press) study, the judged well-formedness of the nonsense forms was found to depend on the statistically most likely parse. For a form such as /strɪnpi/, containing a cluster that is impossible within words, the most likely parse includes a word boundary: /strɪn#pi/.

The score for such a form reflected the likelihood of the winning parse. In general, this means that in speech processing, the relationship of probabilities will be relevant exactly when there is more than one competing analysis in the cognitive system at the time when the speech processing takes place. The expected values that played a role in the original inferences about the form of the system are not necessarily relevant in the adult system; they may be associated with grammars that were supplanted long ago by a more mature conceptualization.

The phonological system as I have described it exhibits confluences across levels that permit bootstrapping of the system from surface regularities to more abstract ones, and that are implicated in the astonishing speed and accuracy of the adult system. One important case of confluence is the correlation of surface (token) statistics with type statistics, a correlation related to the special status of word-initial position as a locus for maximal contrasts. A second case is the privileged status of diphones with regard to acoustic distinctiveness, coarticulatory control, and complexity of the phonological grammar in relation to the size of the lexicon.

Note

I am very grateful to Jen Hay, Dan Jurafsky, and Norma Mendoza-Denton for useful discussions that contributed substantially to the structure of this chapter. Thanks, too, to Stef Jannedy for replotting Peterson and Barney 1952. I also much appreciate reactions from audiences at the Linguistic Society of America 2001 symposium "Probability Theory in Linguistics"; Journées d'Études sur l'Acquisition Phonologique Précoce, 6–8 October 2001, Carry-le-Rouet (Marseille); and the Laboratoire de Science Cognitive et Psycholinguistique.

Chapter 7

Probabilistic Approaches to Morphology

R. Harald Baayen

7.1 Introduction

In structuralist and generative theories of morphology, probability is a concept that, until recently, has not had any role to play. By contrast, research on language variation across space and time has a long history of using statistical models to gauge the probability of phenomena such as *t*-deletion as a function of age, gender, education, area, and morphological structure. In this chapter, I discuss four case studies that illustrate the crucial role of probability even in the absence of sociolinguistic variation. The first shows that by bringing probability into morphological theory, we can make the intuitive notion of morphological productivity more precise. The second considers a data set that defies analysis in terms of traditional syntagmatic rules, but that can be understood as being governed by probabilistic paradigmatics. The third illustrates how the use of item-specific underlying features can mask descriptive problems that can only be resolved in probabilistic morphology. Finally, the fourth focuses on the role that probability plays in understanding morphologically complex words. However, before we consider the different ways in which probability emerges in morphology, it is useful to ask why probability theory, until very recently, has failed to have an impact in linguistic morphology, in contrast to, for instance, biological morphology.

To answer this question, consider the developments in information technology since the 1950s, which have not been without consequences for the study of language. The computers of that era had comparatively reasonable computational capacity but very limited memory, on the order of 15 KB (it was common then to count memory capacities in bits). For a program to work efficiently, it had to minimize storage. Programming

languages such as Fortran, Cobol, Lisp, and Algol were being developed, the last being the first language (in 1958) with a formal grammar and the first language to allow recursive calling of functions.

With very few computers available for the academic community (by 1958, there were only some 2,500 computers in use in the United States), many researchers had to carry out statistical analyses by hand, a tedious and laborious process—even though methods for applied statistics were generally designed to minimize calculations by adopting various kinds of simplifying assumptions (such as independence, linearity, and normality). Linguistic data in electronic form did not exist. Not surprisingly, the linguistic theories of the time took formal languages as the model for language, emphasizing the generative capacity of language, denying any role of importance to probability and statistics, and elevating economy of storage in memory to a central theorem.

Currently, desktop computers have vastly increased processing power and virtually unlimited memory. Carrying out a multivariate analysis is no longer a week's work. Statisticians have developed new, often computationally intensive methods for analyzing data that cannot be modeled adequately by the traditional statistical techniques (e.g., bootstrap and permutation methods). In addition, artificial neural networks (ANNs) have become standard tools in statistics. Some ANN architectures have been found to be equivalent to existing statistical techniques. (See, for instance, Oja 1982 for principal components analysis and Lebart 1995 for correspondence analysis.) Other network architectures have made it possible to estimate probabilities that cannot be calculated efficiently by standard analytical means (see, e.g., Mehta and Patel 1986 and Clarkson, Fan, and Joe 1993, for Fisher's exact test of independence). Yet other network architectures such as feed-forward ANNs provide genuinely new statistical tools for the flexible generalization of linear regression functions (see, e.g., Venables and Ripley 1994, sec. 10.4).

Not only do we now have many more sophisticated statistical techniques, we also have an ever-increasing amount of data. The early corpora for English, such as the Brown corpus (Kučera and Francis 1967), comprised 1 million words; more recent corpora, such as the British National Corpus (⟨http://info.ox.ac.uk/bnc/⟩), contain 100 million words; and the World Wide Web is enjoying increasing use as a data resource with (for English) an estimated 47 billion words in February 2000 (Grefenstette 2000). Not surprisingly, these developments in technology and resources have left their mark on linguistics.

An area of linguistics on which these changes have had a very prominent impact is morphology. Early work on morphology in the generative framework focused on the properties of morphological rewrite rules (Aronoff 1976; Selkirk 1980), within a framework that, with the exception of Jackendoff's (1975) proposals, assumed a strict separation between the regular and the irregular. Although Bybee and Moder (1983) introduced morphological schemas to account for attraction phenomena among irregular forms in English, and although Bybee (1985) proposed to understand morphological phenomena in terms of similarities between stored representations in the lexicon, the study by Rumelhart and McClelland (1986) turned out to be the most effective in challenging the classic view of morphology as a symbolic system. They showed that a very simple ANN could, with a considerable degree of success, map English present tense forms onto past tense forms without making a distinction between regular and irregular forms, and without formulating any explicit symbolic rules.

The original biological motivation for ANNs stems from McCulloch and Pitts (1943). They published a seminal model of a neuron as a binary thresholding function in discrete time that has been very influential in the development of ANNs. A report from 1948 (in Ince 1992) shows that Alan Turing also developed the mathematics for networks of simple processing units (which he called "unorganized machines") in combination with genetic algorithms (what he called a "genetic search") as a model for understanding computation in the brain. This report did not attract attention at the time. Only recently has it become clear that Turing may have been the first connectionist (Copeland and Proudfoot 1999), and only now are his ideas being implemented and studied with computer simulations (Teuscher 2001). Real neurons are now known to be more complicated than the neurons of McCulloch and Pitts, Turing, or the ANNs used in statistics. The interest of McClelland, and Rumelhart, and the PDP Research Group's (1986) connectionist model for the creation of past tense forms, therefore, resides not in the precise form of its network architecture, which is biologically implausible. The value of their study is that, by showing that a network of very simple processing units can perform a linguistic mapping, they have provided a powerful scientific metaphor for how neurons in the brain might accomplish linguistic mappings.

The connectionist past tense model met with fierce opposition. Pinker and Prince (1988) argued that it was fundamentally flawed in just about

any conceivable way. Since then, the discussion has taken the shape of a stimulus-response series in which a paper in the symbolic tradition claiming that an ANN cannot model a certain fact is followed by a study showing how that fact naturally follows once one adopts the right kind of connectionist architecture and training regime (see, e.g., MacWhinney and Leinbach 1991; Plunkett and Juola 2000). Brain imaging data have been advanced as evidence against the connectionist account (Jaeger et al. 1996), without convincing the connectionist opposition (Seidenberg and Hoeffner 1998).

Nevertheless, the symbolic position seems to be in something of a retreat. For instance, Pinker and Prince (1988) and Pinker (1991) flatly reject the connectionist approach. More recently, however, Pinker (1997, 1999) allows for the possibility that irregular verbs are stored in some kind of associative memory, although he maintains the claim that language comprises a mental dictionary of memorized words, on the one hand, and a mental grammar of creative rules, on the other. Marcus (2001) seems to go a step further by accepting connectionist models as enlightening implementational variants of symbolic systems. But he too claims that ANNs are incapable of explaining those crucial data sets that would reveal the supposedly symbolic nature of human language processing. Another index of the retreat of the narrow symbolic position is the emergence of stochastic Optimality Theory (Boersma 1998; Zuraw 2000), an extension of Optimality Theory incorporating a mechanism accounting for nondeterministic data, and of interpretations of Optimality Theory in which similarity spaces and attractors play a crucial role (Burzio, in press b).

At least three issues play a role in the controversy between the connectionist and symbolic positions. The first issue is whether human cognition and language as a cognitive faculty are fundamentally symbolic in nature. This is an issue about which I will remain agnostic.

A second issue is whether ANNs are appropriate models for language. For instance, should ANNs be able to generalize outside the scope of their training space, as argued by Marcus (2001)? The answers to questions such as this depend on a host of assumptions about learning and generalizability in animals, primates, and humans, questions that go far beyond my competence and the scope of this chapter.

A third issue at stake here is whether language is, at its core, a deterministic phenomenon (one that can be handled by simple symbolic rules) or a probabilistic phenomenon (one for which such simple symbolic rules

are inadequate). This is the issue addressed in this chapter. I will argue that the role of probability in morphology is far more pervasive than standard textbooks on morphology would lead one to believe. However, this chapter will not be concerned with how ANNs deal with nondeterministic data, for two reasons. First, a good introduction to neural network theory requires a chapter of its own (see, e.g., McLeod, Plunkett, and Rolls 1998). Second, given that the neural networks used to model language are *artificial* neural networks providing abstract statistical models for linguistic mapping problems, it makes sense to consider a broader range of statistical tools available at present for understanding the quantitative structure of such problems. From this perspective, ANNs pair the advantage of maximal flexibility with the disadvantages of requiring considerable time with respect to training the model, on the one hand, and a loss of analytical user control, on the other: to understand how an ANN achieves a particular mapping itself requires application of conventional multivariate statistical techniques.

What I will therefore discuss in this chapter are some quantitative techniques for coming to grips with the probabilistic structure of morphological phenomena that, unlike ANNs, do not require extensive training time and that provide immediate insight into the quantitative structure of morphological data. I offer these techniques as useful analytical tools, without committing myself to any of them as "models of the mind."

However, I will also indulge in speculating how these techniques might be articulated in terms of the spreading activation metaphor, the current gold standard in psycholinguistics for modeling lexical processing. I indulge in these speculations in order to suggest how the mental lexicon might deal with probabilistic phenomena without invoking complex statistical calculations. Those committed to a connectionist approach will have no difficulty reformulating my symbolic spreading activation models at the subsymbolic level. Those committed to Optimality Theory will find that my models can be reformulated within stochastic Optimality Theory. Both kinds of reformulation, however, come with the cost of increased complexity in terms of the numbers of formal parameters required to fit the data.

The remainder of this chapter is structured as follows. Section 7.2 illustrates how probability theory can help us to understand a key issue in morphological theory, the enigmatic phenomenon of morphological productivity. Section 7.3 discusses two data sets illustrating the role of

probability in the production of complex words. Finally, section 7.4 considers the role of probability during the comprehension of morphologically complex words.

7.2 Probability and Productivity

Aronoff (1976) described productivity as one of the central mysteries of derivational morphology. What is so mysterious about productivity is not immediately evident from the definition given earlier by Schultink (1961)—namely, that productivity is the possibility available to language users to coin, unintentionally, an in principle uncountable number of formations. The empirical problem that makes productivity so mysterious is that some word formation rules give rise to few words, while other word formation rules give rise to many. In English, there are many words in *-ness* (*goodness*), fewer words in *-ee* (*employee*), and hardly any words in *-th* (*warmth*). In other words, productivity is graded or scalar in nature, with productive word formation at one extreme, semiproductive word formation in the middle, and unproductive word formation at the other extreme.

Productivity becomes an even more enigmatic notion once it is realized that unproductive word formation patterns can be fully regular (e.g., the Dutch suffix *-in* that creates female agent nouns, as in *boer* 'farmer', *boerin* 'female farmer'), while word formation need not be rule governed in the traditional sense to be productive (see the discussion of linking elements in Dutch in section 7.3).

Some researchers (Schultink 1961; Bauer 2001) have argued for a principled distinction between productive and unproductive word formation. An unproductive affix would be "dead"; it would not be part of the grammar. Productive affixes, on the other hand, would be "alive," and the question of degrees of productivity would only arise for such living affixes. The problem with this view is that in practice, it is very difficult to know whether an affix is truly unproductive. Consider, for instance, the following quotation from Bauer 2001, 206:

> Individual speakers may coin new words which are not congruent with currently predominating customs in the community as a whole. *The Oxford English Dictionary* credits Walpole with coining *gloomth* and *greenth* in the mid-eighteenth century, some 150 years after the end of the period of societal availability for *-th*; *greenth* appears to have survived into the late nineteenth century, but neither is now used, and neither can be taken to illustrate genuine productive use of *-th*.

Interestingly, a simple query of the World Wide Web reveals that words such as *coolth*, *greenth*, and even *gloomth* are used by speakers of English, even though *-th* is one of the well-worn examples of a supposedly completely unproductive suffix in English. Consider the following examples of the use of *coolth*:

(1) *Coolth*, once a nonce word made on analogy with *warmth*, is now tiresomely jocular: *The coolth of the water in the early morning is too much for me*. Kenneth G. Wilson (1923?). *The Columbia Guide to Standard American English*. 1993. ⟨http://www.bartleby.com/68/5/1505.htm⟩

(2) Increase the capacity of your house to store coolth. (Yes, it is a real word.) Using the mass in the house to store coolth in the summer and heat in the . . . ⟨http://www.tucsonmec.org/tour/tech/passcool.htm⟩

(3) The combination of high-altitude and low-latitude gives Harare high diurnal temperature swings (hot days and cool nights). The team developed a strategy to capture night-time coolth and store it for release during the following day. This is achieved by blowing night air over thermal mass stored below the verandah's . . . ⟨http://www.arup.com/insite/features/printpages/harare.htm⟩

(4) Do we see the whiteness of the snow, but only believe in its coolth. Perhaps this is sometimes so; but surely not always. Sometimes actual coolth is . . . ⟨http://www.ditext.com/sellars/ikte.htm⟩

(5) Early drafts of *Finnegans Wake*—HCE . . . behaved in an ungentlemanly manner opposite a pair of dainty maidservants in the greenth of the rushy hollow, whither, or so the two gown and pinners pleaded . . . ⟨http://www.robotwisdom.com/jaj/fwake/hce.html⟩

(6) Macom—Garden realization . . . realization 3. Delivery of carpet lawn—Fa Kotrba 4. Maintainance of greenth a) chemical treatment; weeding out; fertilization and plant nutrition; prevention of . . . ⟨http://www.macom.cz/english/service.htm⟩

(7) This year I discovered the Gothic novel. The first Gothic novel I read was "*Melmoth the Wanderer*." I read all 697 pages in about five days it was so good. In the Penguin Classics introduction to

> "*Melmoth*" it mentions other Gothic novels such as "*The Italian*," "*Vathek*" and "*The Castle of Otranto*." All of which I've since read and have discovered a new genre of fiction which I really enjoy. I've also had a new word added to my vocabulary: "*Gloomth*." ⟨http://www.geocities.com/prozacpark/gothnovel.htm⟩

Example (1) is an example of the prescriptive view, mirroring on the Web the quotation from Bauer 2001. Example (2) shows how *coolth* is used to fill the lexical gap in the series *hot*/*heat*, *warm*/*warmth*, *cool*/***coolth***, *cold*/*cold*. It is a technical term introduced to the reader as a real word of English. The writer of example (3) takes the use of *coolth* for granted, and the writer of example (4), in a discussion of Kant's theory of experience, seems to find the nontechnical use of *coolth* unproblematic. Examples (5) and (6) illustrate the use of *greenth*, and example (7) shows how modern speakers can even enjoy learning about *gloomth*. These examples show that forms such as *coolth*, *greenth*, and *gloomth* are occasionally used in current English, testifying to the residual degree of productivity of *-th* and the graded, scalar nature of productivity.

At first sight, it would seem that the degree of productivity of a word formation pattern might be captured by counting the number of distinct formations (henceforth word types). The problem with type frequency as a measure of productivity is that unproductive patterns may comprise more types than productive patterns. In Dutch, for instance, the suffix *-elijk* occurs more often than the prefix *her-* (see, e.g., Baayen 2001), but it is the latter and not the former that is generally judged to be productive. What we need, then, is a measure that captures the probability of new words, independently of the number of words that are already attested.

Two measures that formalize the notion of degree of productivity in terms of probability are available (Baayen and Renouf 1996; Baayen 2001). They are based on the probability theory of the number of different species (types) observed among a given number of observations (tokens). First consider the "productivity" of a fair die. There are six types: 1, 2, 3, 4, 5, and 6. Imagine how many different types we count as we throw a fair die 100 times. How many of these types may we expect to have seen after N throws? In other words, how does the expected vocabulary size $\mathrm{E}[V(N)]$ increase as a function of the sample size N? The growth curve of the vocabulary size for a fair die is shown in the upper left panel of figure 7.1 using a solid line. The vertical axis plots the expected count of types. The horizontal axis plots the number of throws,

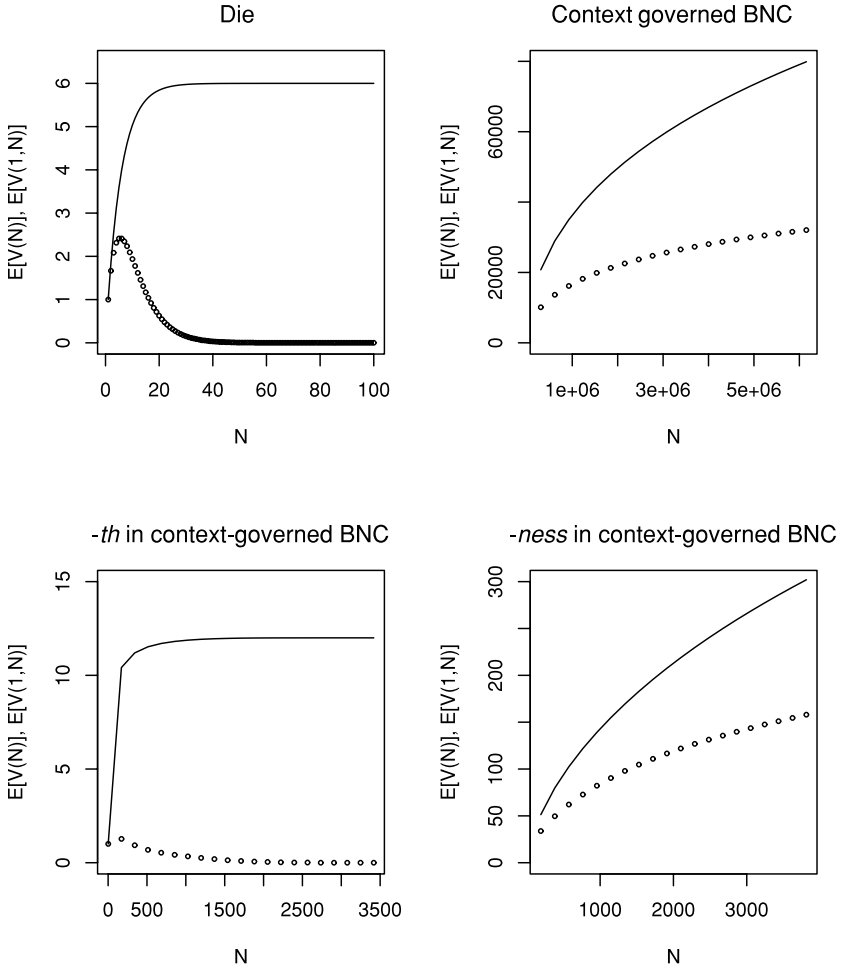

Figure 7.1
Type-token dynamics for 100 throws with a fair die (*upper left*), for the context-governed subcorpus of the British National Corpus (*upper right*), as well as for the nouns in *-th* (*lower left*) and the nouns in *-ness* (*lower right*) in this subcorpus. The horizontal axis plots the number of tokens (N), the vertical axis the number of types ($\mathrm{E}[V(N)]$) and number of hapax legomena ($\mathrm{E}[V(1,N)]$).

that is, the individual observations or tokens. The growth curve of the vocabulary shows that after 40 throws, we are almost certain to have seen each side of the die at least once. In fact, each type will probably have been counted more than once. This is clear from the dotted line in the graph, which represents the growth curve $\mathrm{E}[V(1,N)]$ of the hapax legomena, the types that occur exactly once in the sample. After 40 trials, it is very unlikely that there is a type left in the sample that has been observed only once.

Now consider the upper right panel of figure 7.1. Here we see the corresponding plot for some 6 million words from the British National Corpus (its context-governed subcorpus of spoken British English). Now imagine that instead of throwing a die, we are reading through this subcorpus, word token by word token, keeping track of the number of different word types, and also counting the hapax legomena. The upper right panel of figure 7.1 shows that both the growth curve of the vocabulary and the growth curve of the hapax legomena increase as we read through the corpus. There is no sign of the growth curve of the vocabulary reaching an asymptote, as in the case of the fair die. There is also no indication of the growth curve of the hapax legomena having an early maximum with the horizontal axis as asymptote as N is increased. This pattern is typical for word frequency distributions, irrespective of whether one is dealing with small texts of a few thousand words or with huge corpora of tens or even hundreds of millions of words. It is also typical for the word frequency distributions of productive affixes. For instance, the nouns in *-ness* in the context-governed subcorpus of the BNC are characterized by the growth curves of the vocabulary and the hapax legomena shown in the lower right panel. Conversely, the pattern shown for the fair die represents a frequency distribution prototypical for unproductive affixes. For instance, the nouns in *-th* in the context-governed subcorpus of the BNC are characterized by the growth curves shown in the lower left panel of figure 7.1.

It turns out that the rate $P(N)$ at which the vocabulary size $V(N)$ increases is a simple function of the number of hapax legomena $V(1,N)$ and the sample size N:

$$P(N) = \frac{\mathrm{E}[V(1,N)]}{N} \tag{1}$$

(see, e.g., Good 1953; Baayen 2001). Note that the growth rate of the vocabulary size is itself a function of the sample size. The rate at which

the vocabulary size increases decreases through sampling time. Initially, nearly every word is new, but as we read through the corpus, we see more and more words we have encountered before. Also note that the upper left panel of figure 7.1 clearly illustrates the relation between the growth rate $P(N)$ and the growth curve of the number of hapax legomena $V(1, N)$. After 50 throws, the growth curve of the vocabulary is, at least to the eye, completely flat. After 50 throws, therefore, the growth rate of the vocabulary should be very close to zero. Since after 50 throws the number of hapax legomena has become practically zero, $P(N)$ must also be zero, as required.

The growth rate $P(N)$ has a simple geometric interpretation: it is the slope of the tangent to the growth curve of the vocabulary size at sample size N. The growth rate $P(N)$ is also a probability—namely, the probability that, after having sampled N tokens, the next token to be sampled will represent a type that has not been observed among the previous N tokens. To see this, consider an urn containing a fixed number of marbles. Each marble has one color, there are V different colors, there are N marbles, and there are $V(1)$ marbles with a unique color (i.e., with a color no other marble has). The probability that the first marble drawn from the urn will have a color that will not be sampled again is $V(1)/N$. Since there is no reason to suppose that sampling the last marble will be different from sampling the first marble, the probability that the very last marble taken from the urn represents a color that has not been seen before must also be $V(1)/N$. This probability approximates the probability that, if marbles from the same population are added to the urn, the first added marble drawn from the urn will have a new color. The expectation operator in (1) makes this approximation precise.

In order to derive productivity measures from the growth rate of the vocabulary size, consider the case where we sample a new token after having read through N tokens. Let $\{A\}$ denote the event that this token represents a new type. We furthermore regard the vocabulary as a whole to be a mixture of C different kinds of words: various kinds of monomorphemic words (simplex nouns, adjectives, pronouns, etc.), and many different kinds of complex words (compounds, nouns in *-ness*, verbs in *-ize*, adverbs in *-ly*, etc.). Let $\{B\}$ denote the event that the $N + 1$-th token belongs to the i-th mixture component of the vocabulary. The hapax-conditioned degree of productivity $\mathscr{P}^*(N, i)$ of the i-th mixture component is the conditional probability that the $N + 1$-th token belongs to the i-th mixture component, given that it represents a type that has not been

observed before. Let $V(1, N, i)$ denote the number of hapax legomena belonging to the i-th mixture component, observed after N tokens have been sampled:

$$\begin{aligned}\mathscr{P}^*(N,i) &= P(\{B\}|\{A\}) \\ &= \frac{P(\{B\} \cap P(\{A\})}{P(\{A\})} \\ &= \frac{\frac{\mathrm{E}[V(1,N,i)]}{N}}{\frac{\sum_{j=1}^{C} \mathrm{E}[V(1,N,j)]}{N}} \\ &= \frac{\mathrm{E}[V(1,N,i)]}{\mathrm{E}[V(1,N)]}. \end{aligned} \tag{2}$$

For *-th* and *-ness*, the values of $\mathscr{P}^*$ are 1.0e–13/32042 = 3.1e–18 and 158/32042 = .0049, respectively. Note that for spoken British English, the probability of observing new formations in *-th* is vanishingly small. It is probably only for written English that an extremely large corpus such as the World Wide Web (for which in the year 2000 $N > 47{,}000{,}000{,}000$; Grefenstette 2000) succeeds in showing that there are unobserved formations—in other words, that there is some very small residual productivity for *-th*.

The category-conditioned degree of productivity $\mathscr{P}(N, i)$ of mixture component i is the conditional probability that the $N + 1$-th token represents a new type, given that it belongs to mixture component (or morphological category) i. With N_i the number of tokens counted for the i-th mixture component, we have

$$\begin{aligned}\mathscr{P}(N,i) &= P(\{A\}|\{B\}) \\ &= \frac{P(\{A\} \cap P(\{B\})}{P(\{B\})} \\ &= \frac{\frac{\mathrm{E}[V(1,N,i)]}{N}}{\frac{N_i}{N}} \\ &= \frac{\mathrm{E}[V(1,N,i)]}{N_i}. \end{aligned} \tag{3}$$

Applying (3) to *-th* and *-ness*, we obtain 1.0e–13/3512 = 2.8e–17 and 158/3813 = .04 as estimates of $\mathscr{P}$ for *-th* and *-ness*, respectively.

To understand the difference between the interpretations of $\mathscr{P}$ and $\mathscr{P}^*$, it is important to realize that productivity is determined by a great many factors, ranging from structural and processing constraints to register and modality (Bauer 2001; Plag, Dalton-Puffer, and Baayen 1999). Since $\mathscr{P}^*$ estimates the contribution of an affix to the growth rate of the vocabulary as a whole, it is a measure that is very sensitive to the different ways in which nonsystemic factors may affect productivity. For instance, the suffix *-ster* attaches productively to Dutch verbs to form female agent nouns (*zwem-er* 'swimmer', *zwemster* 'female swimmer'). However, even though *-ster* is productive, speakers of Dutch are somewhat hesitant to use it. Consequently, its contribution to the overall growth rate of the vocabulary is quite small.

The category-conditioned degree of productivity of a given affix does not take counts of other affixes into consideration. This measure is strictly based on the morphological category of the affix itself. It estimates its productivity, independently of the nonsystemic factors. Hence, it provides a better window on the potentiality of the affix. Measured in terms of $\mathscr{P}$, *-ster* emerges with a high degree of productivity (see Baayen 1994 for experimental validation).

The prominent role of the hapax legomena in both productivity measures makes sense from a processing point of view. The more frequent a complex word is, the more likely it is that it is stored in memory and the less likely it is that its constituents play a role during production and comprehension (Hasher and Zacks 1984; Scarborough, Cortese, and Scarborough 1977; Bertram, Schreuder, and Baayen 2000). Conversely, the more infrequent words there are with a given affix, the more likely it is that its structure will be relevant during comprehension and production. The number of hapax legomena, the lowest-frequency words in the corpus, therefore provides a first approximation of the extent to which the words with a given affix are produced or accessed through their constituents.

Hay (2000) provides a more precise processing interpretation for the category-conditioned degree of productivity. This study brings together the insight that phonological transparency codetermines productivity and the insight that relative frequency is likewise an important factor. First, consider phonological transparency. The more the phonological form of

the derived word masks its morphological structure, the less such a form will contribute to the productivity of the morphological category to which it belongs (see, e.g., Cutler 1981; Dressler 1985). Hay operationalizes the role of phonological transparency in terms of offset-onset probabilities and then shows that the resulting juncture probabilities predict other properties of the words in which they occur, such as prefixedness ratings, semantic transparency ratings, and number of meanings.

Next, consider relative frequency. The idea here is that the frequency relation between a derived word and its base should codetermine the parsability of that word. If the frequency of the derived word is substantially greater than that of its base, it is unlikely that the base will effectively contribute to the processes of production and comprehension. If, on the other hand, the frequency of the derived word is much lower than the frequencies of its constituents, it is much more likely that these constituents do have a role to play. Hay (2000) shows that relative frequency predicts pitch accent placement: prefixes in words for which the derived frequency is greater than the frequency of the base are less likely to attract pitch accent than prefixes in words for which the derived frequency is less than the base frequency. She also shows that *t*-deletion is more likely to occur in case the derived word is more frequent than its base. Finally, she shows that complexity ratings for such words tend to be lower than for words for which base frequency exceeds derived frequency.

Interestingly, Hay demonstrates that for a sample of 12 English derivational affixes, the category-conditioned degree of productivity is a linear function of mean relative frequency and mean juncture probability of the formations in the corresponding morphological categories. In other words, the probability that a morphological category will give rise to new formations emerges as being demonstrably codetermined by the juncture probabilities of its members and the frequency relations between these members and their base words. Hay and Baayen (2002) provide a more detailed analysis of the correlation between relative frequency and the two productivity measures $\mathscr{P}$ and $\mathscr{P}^*$ for 80 English derivational affixes. Their results suggest that the degree of productivity of an affix correlates surprisingly well with the likelihood that it will be parsed in comprehension.

Having looked at how probability theory can help us come to grips with the elusive notion of degrees of productivity, let us now consider the possibility that morphological regularity itself is probabilistic in nature.

7.3 Probability in Morphological Production

In this section, I introduce two data sets illustrating the role of probability in the production of morphologically complex words. The first data set concerns the production of linking elements in Dutch nominal compounds. The second addresses the selection of voice specification of syllable-final obstruents in Dutch.

7.3.1 Linking Elements in Dutch

The immediate constituents of nominal compounds in Dutch are often separated by what I will refer to as a *linking element*. Whether a linking element should be inserted, and if so, which linking element, is difficult to predict in Dutch. To see this, consider the compounds in (8):

(8) a. schaap-herder
sheep-herder
'shepherd'
b. schaap-S-kooi
sheep-s-fold
'sheepfold'
c. schaap-EN-vlees
sheep-EN-meat
'mutton'

The same left constituent appears without a linking element, with the linking element *-s-*, and with the linking element *-en-*. Intensive study of this phenomenon has failed to come up with a set of rules that adequately describe the distribution of linking elements in Dutch compounds (see Krott, Baayen, and Schreuder 2001, for discussion and further references). This suggests that the appearance of linking elements is fairly random and that the use of linkers is unproductive. This is not the case, however. Linking elements are used productively in novel compounds, and there is substantial agreement among speakers about which linking element is most appropriate for a given pair of immediate constituents. The challenge that the linking elements of Dutch pose for linguistic theory is how to account for the paradox of an apparently random morphological phenomenon that nevertheless is fully productive.

The key to solving this paradox is to exchange a syntagmatic approach for a paradigmatic approach, and to exchange greedy learning for lazy

learning. A syntagmatic approach assumes that it is possible to formulate a generalization describing the properties that the context should have for a given linking element to appear. A paradigmatic approach assumes that the set of compounds similar to the target compound requiring the possible insertion of a linking element, its compound paradigm, forms the analogical basis from which the probabilities of the different linking elements are derived. The syntagmatic approach is most often coupled with greedy learning, in the sense that once the generalization has been abstracted from a set of examples, these examples are discarded. In fact, researchers working in the tradition of generative grammar tend to believe that learning a rule and forgetting about the examples that allowed the rule to be deduced go hand in hand. Pinker (1991, 1997, 1999), for instance, has argued extensively for a strict division of labor between rules accounting for what is productive and regular, on the one hand, and storage in memory for what is unproductive and irregular, on the other hand. Conversely, the paradigmatic, analogical approach is based on the insight that learning may involve a continuous process driven by an ever-increasing instance base of exemplars. In this approach, it may even be harmful to forget individual instances. This kind of learning, then, is lazy in the sense that it does not attempt to formulate a rule that allows the data to be discarded. Greedy learning, once completed, requires little memory. By contrast, lazy learning assumes a vast storage capacity.

Two mathematically rigorously defined approaches to the modeling of paradigmatic analogy are available: Analogical Modeling of Language (AML; Skousen 1989, 1992), and the many machine learning algorithms implemented in the TIMBL program of Daelemans et al. (2000). Both AML and TIMBL determine the choice of the linking element for a given target compound on the basis of the existing compounds that are most similar to this target compound. The two methods differ with respect to what counts as a similar compound. AML makes use of a similarity metric that also plays a role in quantum mechanics (Skousen 2000). I will return to AML below. In this section, I describe the IB1-IG metric (Aha, Kibler, and Albert 1991; Daelemans, van den Bosch, and Weijters 1997) available in TIMBL.

In formal analogical approaches, the question of which linking element to choose for a compound amounts to a classification problem: does this compound belong to the class of compounds selecting *-en-*, to the class of compounds selecting *-s-*, or to the class of compounds with no overt

Table 7.1
Features and their values for a hypothetical instance base of Dutch compounds. *L* denotes the linking element. The numbers in parentheses refer to the first and second constituents.

Modifier (1)	Head (2)	Nucleus (1)	Onset (2)	Coda (2)	L	Translation
schaap	bout	aa	b	t	-en-	'leg of mutton'
schaap	herder	aa	h	r	-∅-	'shepherd'
schaap	kooi	aa	k	i	-s-	'sheepfold'
schaap	vlees	aa	v	s	-en-	'mutton'
lam	bout	a	b	t	-s-	'leg of lamb'
lam	vlees	a	v	s	-s-	'lamb'
lam	gehakt	a	g	t	-s-	'minced lamb'
paard	oog	aa	—	g	-en-	'horse's eye'
koe	oog	oe	—	g	-en-	'cow's eye'
varken	oog	e	—	g	-s-	'pig's eye'

linking element (for notational convenience, henceforth the compounds with the linking element *-0-*)? In order to establish which class a compound for which we have to determine the linking element belongs to, we need to define the properties of compounds on which class assignment has to be based. In other words, we need to know which features are relevant and what values these features might have. Now consider table 7.1, which lists a hypothetical instance base with 10 compounds. In this example, there are five features: the modifier, the head, the nucleus of the modifier, the onset of the head, and the coda of the head. The values of the feature Nucleus are the vowels *aa*, *a*, *oe*, and *e*. The values of the feature Modifier are the left constituents *schaap*, *lam*, *paard*, *koe*, and *varken*. What we want to know is what the most probable linking element is for the novel compound *schaap-?-oog* 'sheep's eye'.

To answer this question, we need to know which exemplars in the instance base are most relevant. We want to discard exemplars that are very different from the target compound, and we want to pay special attention to those compounds that are very similar. In other words, we need a similarity metric, or, alternatively, a distance metric. The IB1-IG distance metric is based on a simple distance metric (known as the "simple matching coefficient" or "Hamming distance") that tracks the number of features that have different values. Let X denote the target compound, and let Y denote an exemplar in the instance base. The

Hamming distance between these two compounds, $\Delta(X, Y)$, is defined as the number of features with mismatching values:

$$\Delta(X, Y) = \sum_{i=1}^{n} \mathrm{I}_{[x_i \neq y_i]}. \tag{4}$$

In the present example, the number of features n equals 5. The value of the i-th feature of X is denoted by x_i. The operator $\mathrm{I}_{[z]}$ evaluates to 1 if the expression z is true, and to 0 otherwise. This metric allows us to group the compounds in the instance base according to their distance from the target compound. For instance, *schaap-en-bout* is at distance 3 from the target *schaap-?-oog*, because these two compounds mismatch with respect to the features Head, Onset(2), and Coda(2). We can now determine the set $\mathscr{S}$ of compounds that, given the features and their values, are most similar to the target compound. The distribution of linking elements in this set of nearest neighbors determines the probabilities of selection. Denoting the cardinality of $\mathscr{S}$ by S, the probability of the linking element *-en-* is given by the proportion of compounds in $\mathscr{S}$ that select *-en-*:

$$P(L = \text{-}en\text{-}) = \sum_{i=1}^{S} \frac{I_{[L_i = \text{-}en\text{-}]}}{S}. \tag{5}$$

The IB1-IG distance measure improves considerably upon (4) by weighting the features for their relevance, using the information-theoretic notion of entropy. The entropy $H(L)$ of a distribution of linking elements L, with J different linking elements,

$$H(L) = -\sum_{j=1}^{J} p_j \log_2 p_j, \tag{6}$$

is a measure of uncertainty about which linking element to choose in the situation where no information is available about the values of the features of a given word. For the data in table 7.1,

$$\begin{aligned} H(L) &= -[P(L = en) \log_2(P(L = en)) + P(L = s) \log_2(P(L = s)) \\ &\qquad + P(L = \varnothing) \log_2(P(L = \varnothing))] \\ &= -[0.4 \log_2(0.4) + 0.5 \log_2(0.5) + 0.1 \log_2(0.1)] \\ &= 1.36. \end{aligned}$$

The degree of uncertainty changes when extra information is provided, for instance, the information that the value of the feature Modifier is *schaap*:

$$
\begin{aligned}
&H(L|\text{Modifier} = \textit{schaap}) \\
&\quad = -[P(L = \textit{en}|\text{Modifier} = \textit{schaap}) \log_2(P(L = \textit{en}|\text{Modifier} = \textit{schaap})) \\
&\qquad + P(L = s|\text{Modifier} = \textit{schaap}) \log_2(P(L = s|\text{Modifier} = \textit{schaap})) \\
&\qquad + P(L = \varnothing|\text{Modifier} = \textit{schaap}) \\
&\qquad \times \log_2(P(L = \varnothing|\text{Modifier} = \textit{schaap}))] \\
&\quad = -[0.5 * \log_2(0.5) + 0.25 * \log_2(0.25) + 0.25 * \log_2(0.25)] \\
&\quad = 1.5.
\end{aligned}
$$

We can gauge the usefulness or weight w_i of a feature F_i for predicting the linking element by calculating the probability-weighted extent to which knowledge of the value v of F_i decreases our uncertainty:

$$w_i = H(L) - \sum_{v \in F_i} P(v)H(L|v). \tag{7}$$

For the feature Modifier, v ranges over the values *schaap*, *lam*, *paard*, *koe*, and *varken*. Note that when v has the value *schaap*, the probability $P(v)$ equals 4/10. Because $H(L|v) = 0$ when $v \neq \textit{schaap}$ (the other modifiers all occur with just one linking element, so there is absolute certainty about the appropriate linking element in these cases), the information gain weight for the feature Modifier is

$$
\begin{aligned}
w_{\text{Modifier}} &= H(L) - P(\text{Modifier} = \textit{schaap})H(L|\text{Modifier} = \textit{schaap}) \\
_{\text{Modifier}} &= 1.36 - 0.4 * 1.5 \\
_{\text{Modifier}} &= 0.76.
\end{aligned}
$$

The feature with the lowest information gain weight is Coda(2), which is not surprising as there is no obvious phonological reason to suppose the coda of the second constituent to codetermine the choice of the linking element. Crucially, when this technique is applied to a realistic instance base, the information gain weights are a powerful means for establishing which features are important for understanding the quantitative structure of a data set.

When this technique is applied to an instance base of Dutch compounds as available in the CELEX lexical database (Baayen, Piepenbrock, and

Gulikers 1995), it turns out that, of a great many features, the Modifier and Head features have the highest information gain values (1.11 and 0.41, respectively) and that other features, such as whether the first constituent bears main stress, have a very low information gain weight (0.07). When we modify our distance metric by weighting for information gain,

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i I_{[x_i \neq y_i]}, \tag{8}$$

and when we choose the linking element on the basis of the distribution of linking elements in the set of compounds with the smallest distance Δ, some 92% of the linking elements in Dutch compounds are predicted correctly (using 10-fold cross-validation). Experimental studies (Krott, Baayen, and Schreuder 2001; Krott, Schreuder, and Baayen 2002) have confirmed the crucial importance of the Modifier and Head features. Apparently, the compounds sharing the modifier constituent (the left constituent compound family), and to a lesser extent the compounds sharing the head constituent (the right constituent compound family), form the analogical exemplars on which the choice of the linking element in Dutch is based. What we have here is paradigmatically determined selection instead of syntagmatically determined selection. Instead of trying to predict the linking element on the basis of specific feature values of the surrounding constituents (its syntagmatic context), it turns out to be crucial to zoom in on the constituents themselves and the distributional properties of their positional compound paradigms.

To see how such paradigmatic effects might be accounted for in a psycholinguistic model of the mental lexicon, consider figure 7.2. This figure outlines the functional architecture of a spreading activation model for paradigm-driven analogy using the small instance base of table 7.1. At the left-hand side of the graph, the modifier (labeled LEFT) and the head constituent (labeled RIGHT) are shown. These two lexemes are connected to their positional compound paradigms, listed in the center. The weights on the connections to the compounds in the instance base are identical to the information gain weights of the left and right constituents. The different linking elements are displayed at the right-hand side of the graph. Each linking element is connected with the compounds in which it appears. Activation spreads from the left and right constituents to the compounds in the paradigmatic sets, and from there to the linking element. The linking element that receives the most activation is the

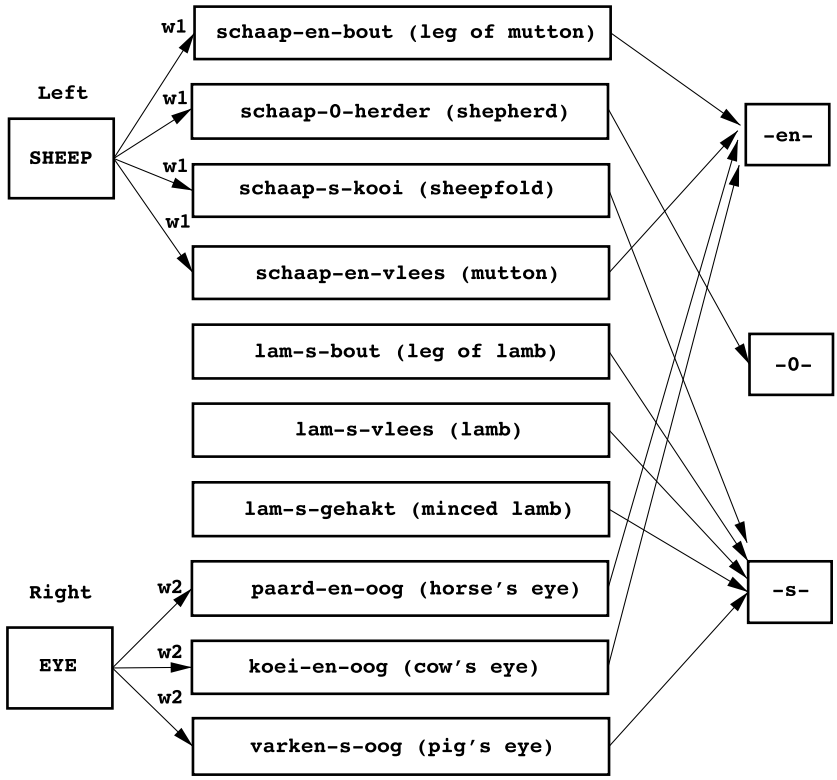

Figure 7.2
A spreading activation model for the selection of linking elements in Dutch. (After Krott, Schreuder, and Baayen 2002.)

one selected for insertion in the novel compound *schaap-?-oog*. Krott, Schreuder, and Baayen (2002) show that an implemented computational simulation model along these lines provides excellent fits to both the choices and the times required to make these choices in experiments with novel Dutch compounds.

This example shows that morphological network models along the lines proposed by Bybee (1985, 1995a, 2001) can be made precise and that, once formalized, they have excellent predictive power. It should be kept in mind, however, that the present model presupposes considerable structure in the mental lexicon, both in terms of the specific connectivity required and in terms of the specific information gain weights on these connections. In this light, it makes more sense to speak of an analogical

or paradigmatic rule for the selection of the linking element rather than of a network model, because we are dealing not with undifferentiated connectivity in an encompassing network for the whole mental lexicon, but with highly structured connectivity in a subnetwork dedicated to the specific task of selecting the appropriate linking element for Dutch compounds.

The next section provides a second example of a phenomenon that turns out to be analogical in nature, the morphophonology of final obstruents in Dutch.

7.3.2 Syllable-Final Obstruents in Dutch

The feature [voice] is distinctive in Dutch. This is illustrated in (9) for the Dutch nouns /rat-en/ and /rad-en/. When the alveolar stop is voiceless, the noun means 'honeycombs'; when the stop is voiced, the noun means 'councils'. When the stop is in syllable-final position, as in isolated singular forms, the distinction between the voiced and voiceless obstruent is neutralized, and in phrase-final position both obstruents are realized as voiceless.

(9)

Form	*Translation*	*Voicing*
/rat-en/	'honeycomb'-PLURAL	voiceless
/rat/	'honeycomb'	voiceless
/rad-en/	'council'-PLURAL	voiced
/rat/	'council'	voiceless

Traditionally, this phenomenon is accounted for by assuming that the obstruents in /rat-en/ and /rad-en/ are specified as being underlyingly voiceless and voiced, respectively, with a rule of syllable-final devoicing accounting for the neutralization of the voice distinction in the singular (e.g., Booij 1995). Whether the final obstruent in a given word alternates between voiced and voiceless is taken to be an idiosyncratic property of that word that has to be specified lexically, although it has been noted that fricatives following long vowels tend to be underlyingly voiced and that bilabial stops tend to be voiceless following long vowels (Booij 1999).

Ernestus and Baayen (2003) report that there is far more structure to the distribution of the voice specification of final obstruents in the lexicon of Dutch than expected on the basis of this standard analysis. Figure 7.3 summarizes some of the main patterns in the data for three major rime patterns by means of the barplots at the left-hand side. The data on which

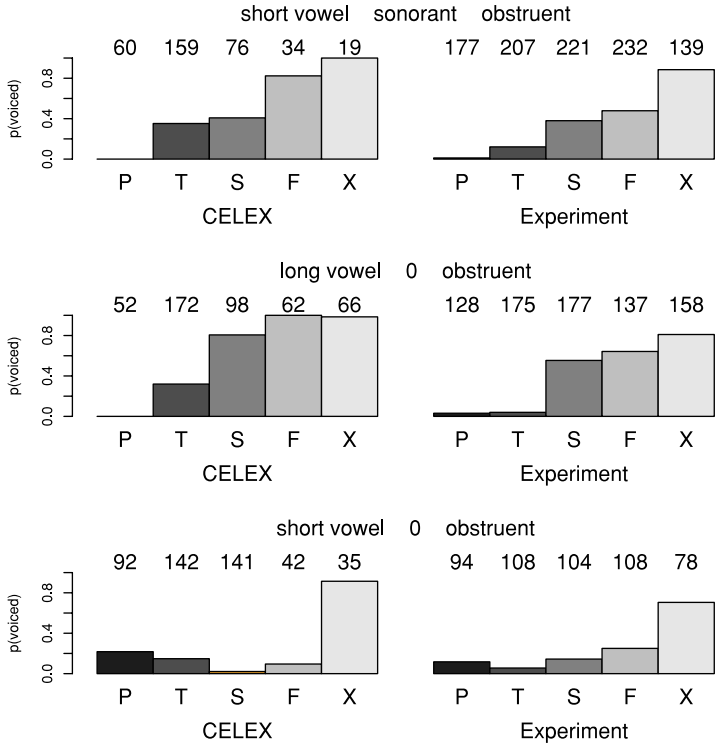

Figure 7.3
Proportions of voiced morpheme-final obstruents in Dutch as a function of type of obstruent, broken down by three rime structures. P: /p,b/; T: /t,d/; S: /s,z/; F: /f,v/; X: /x,g/. The numbers above the bars specify the total counts of monomorphemic words (*left*) and the total number of responses for pseudoverbs (*right*) for the class represented by that bar.

these graphs are based are some 1,700 monomorphemic words attested in the Dutch part of the CELEX lexical database. These words are nouns, verbs, or adjectives that end in an obstruent that has both voiced and voiceless counterparts in Dutch, and that are attested with a following schwa-initial suffix. For these words, we therefore know whether they have an alternating or a nonalternating final obstruent. The top left panel of figure 7.3 plots the proportion of words exhibiting voice alternation (p(voiced)) as a function of the kind of obstruent (bilabial (P) and alveolar (T) stops; labiodental (F), alveolar (S), and velar (X) fricatives) for words with a rime consisting of a short vowel followed by a sonorant

consonant, followed by the final obstruent. Note that as we proceed from left to right, the percentage of words with underlying voiced obstruents increases. The center left panel shows a fairly similar pattern for words ending in a long vowel that is directly followed by the final obstruent without any intervening consonant. The bottom left panel shows the distribution for words ending in a short vowel immediately followed by the final obstruent. For these words, we observe a U-shaped pattern. (A very similar pattern characterizes the subset of verbs in this database of monomorphemic words.) Ernestus and Baayen show that the quality of the vowel, the structure of the coda (does a consonant precede the final obstruent, and if so, is it a sonorant?), and the type of final obstruent are all significant predictors of the distribution of the percentage of voicing in Dutch.

The right panels of figure 7.3 present the corresponding barplots for the data obtained in a production experiment in which participants were asked to produce the past tense form for some 200 artificially created but phonotactically legal pseudoverbs. The past tense suffix was selected because it has two allomorphs the selection of which depends on whether the final obstruent alternates. If the final obstruent alternates, the past tense suffix has the form *-de* and the final obstruent is realized as voiced. If the final obstruent does not alternate, the appropriate past tense suffix is *-te*, and the final obstruent is realized as voiceless. When participants are asked to produce the past tense form, the status they assign to the final obstruent of the pseudoverb, alternating or not alternating, can be determined simply on the basis of the form of the past tense suffix.

Interestingly, the percentages of verbs for which the participants used the past tense suffix *-de* reflect to a considerable degree the percentages of the words with alternating final obstruents in the lexicon shown in the left panels. Note that even the U-shaped pattern in the lower left panel is present to some extent in the lower right panel. This pattern of results is incompatible with theories that hold that the selection of the past tense suffix crucially depends on the availability of a lexically specified feature marking the verb as underlyingly voiced. After all, the participants in the experiment were asked to produce the past tense for pseudoverbs, forms that are not available in the lexicon and for which no such lexically specified feature is available. We are therefore faced with the question of how the participants might have arrived at their choice of the allomorph of the past tense suffix for these pseudoverbs.

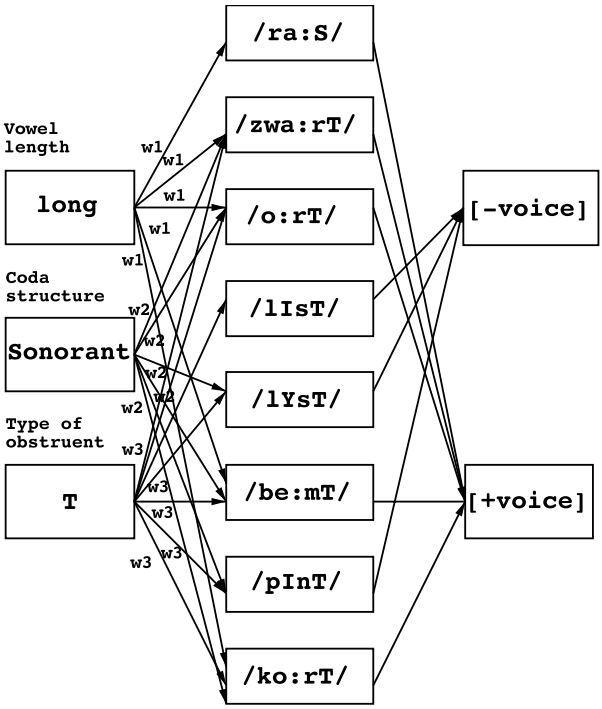

Figure 7.4
A spreading activation model for the selection of the voice specification of final obstruents in Dutch

Figure 7.4 illustrates the kind of lexical connectivity required for a spreading activation network to predict the choice of the past tense allomorph. Completely analogous to the spreading activation model illustrated in figure 7.2 for modeling the choice of the linking element in Dutch compounds, activation spreads from the phonological feature values (left) to the words sharing these feature values (center) and from there to the voicing specification of the final obstruent and of the past tense allomorph (right). As before, this model embodies an analogical, paradigmatic rule. It presupposes that vowel length, coda structure, and type of obstruent can be identified for any given input form and that the speaker has learned that it is these features that are primarily relevant for the voicing alternation. Once the values of these features are activated, the paradigms of words in the lexicon sharing these feature values are

coactivated, proportionally to the weights $w_1, w_2, \ldots$. The support from the weighted paradigmatic cohorts determines the probability of selecting [−voice] or [+voice]. Note that this probability is *not* determined by the proportion of words with exactly the same values as the target for the features vowel length, coda structure, and type of obstruent. Words that share only two feature values, and even words that share only one feature value, also codetermine the probabilities of a voiced or voiceless realization.

A formal definition of these probabilities proceeds as follows. Let F denote the number of features, and let $\bar{v}$ denote the vector

$$\bar{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_F \end{pmatrix}$$

specifying for the target word the value v_i of each feature F_i. We define n_{ijk} as the number of lexemes with value j for feature i that support exponent k, the exponents in this example being the voicing outcomes voiced and voiceless. The support s_k that exponent k receives given target $\bar{v}$ and weights w equals

$$s_k = \sum_{i=1}^{F} w_i n_{iv_i k}, \tag{9}$$

and the probability p_k that it will be selected, given K different exponents, is

$$p_k = \frac{s_k}{\sum_{m=1}^{K} s_m}. \tag{10}$$

The maximum likelihood choice of this spreading activation model is the exponent for which p_k is maximal. This maximum likelihood choice is the choice that the model predicts that the majority of the participants should opt for.

When we set the weights w to the information gain weights (7), the maximum likelihood prediction of the model coincides with the majority choice of the participants in 87.5% of the experimental pseudoverbs. When we optimize the weights using the simplex algorithm of Nelder and Mead (1965), this accuracy score improves to 91.7%. The by-word probabilities for voicelessness in this model correlate well with the proportions

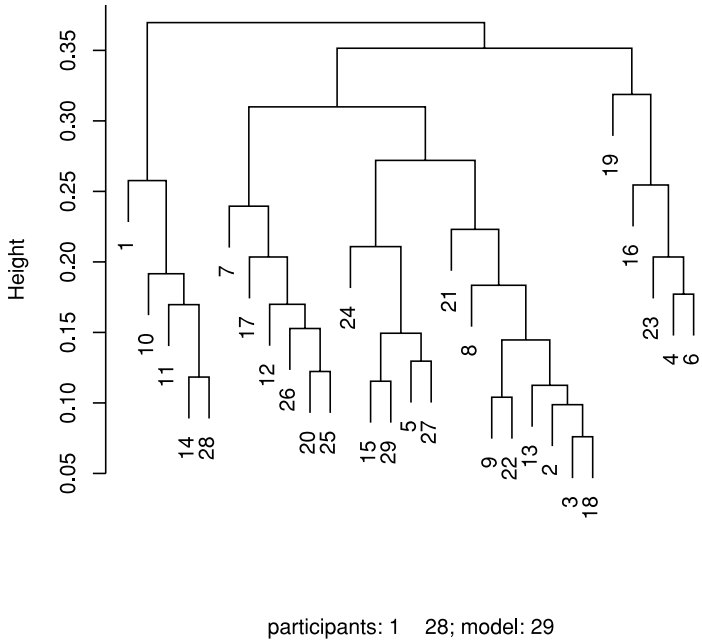

Figure 7.5
Hierarchical clustering of participants (1–28) and the spreading activation model (29) on the basis of pairwise proportional matching coefficients

of participants selecting the voiceless allomorph of the past tense suffix, *-te* ($r = .85$, $t(190) = 22.1$, $p < .0001$). That the maximum likelihood choices of the model are similar to those made by the participants is illustrated in figure 7.5, the dendrogram for the hierarchical clustering of the participants and the model. (The clustering is based on a distance matrix of by-participant pairwise proportions of pseudoverbs that differ with respect to the assignment of voice.) The participants are labeled 1–28, and the model, which can be found in the lower center of the tree diagram, is labeled 29. In a dendrogram such as this, participants who are very similar will be in a very similar position in the tree. For instance, participants 20 and 25 are very similar, and the group formed by participants 20 and 25 is in turn very similar to participant 12. These participants are very different from participants 4 and 6. One has to traverse the tree almost to its root to go from participant 25 to participant 4. In other words, vertical traversal distance, labeled Height in figure 7.5, tells us how dissimilar two participants are. The position of the model, 29,

near to participants 15, 5, and 27, shows that the model's behavior is quite similar to that of a number of actual participants. If the model had occupied a separate position in the dendrogram, this would have been an indication that its voicing predictions might be in some way fundamentally different from those of the participants, which would have been a source of worry about the validity of the model as a model of how speakers of Dutch arrive at their choice of the past tense suffix.

Summing up, the distribution of voice alternation for final obstruents in the lexicon of Dutch is far from random. Speakers of Dutch make use of this information when confronted with novel (pseudo)verbs (see Ernestus and Baayen 2001, 2002, for evidence that even the voice specification of existing words is likewise affected by the distributional properties of voicing in the lexicon). The probability that a speaker of Dutch will select the voiced or voiceless allomorph of the Dutch past tense suffix can be approximated with a reasonable degree of accuracy on the basis of only three parameters, one for each relevant feature.

There are many other formal quantitative models that provide good fits to the present data (see Ernestus and Baayen 2003, for detailed discussion), two of which are of special interest. Boersma (1998) proposes a stochastic version of Optimality Theory (SOT) in which constraints are assigned a position on a hierarchy scale. The exact position of a constraint on this scale is stochastic; that is, it varies slightly from instance to instance, according to a normal distribution. The positions of the constraints are determined by Boersma's Gradual Learning Algorithm (see also Boersma and Hayes 2001). This algorithm goes through the list of forms in the model's input, adjusting the constraints at each step. If the model predicts an outcome that is at odds with the actually attested outcome, the positions of the constraints are adjusted. Those constraints that are violated by the actual input form are moved down. At the same time, those constraints that are violated by the words that the model thought were correct instead of the actual input form are moved up in the hierarchy. The simplest model that provides a reasonable fit to the data is summarized in table 7.2. It has 10 constraints, and hence 10 parameters. The maximum likelihood predictions of this model correspond for 87% of the experimental pseudoverbs with the majority choice of the participants, a success rate that does not differ significantly from the success rate (91%) of the spreading activation model ($p > .25$, proportions test).

Given that SOT and the spreading activation model have the same observational adequacy, the question arises which model is to be pre-

Table 7.2
Constraints and their position in Stochastic Optimality Theory for the voice specification of final obstruents in Dutch

Constraint	Gloss	Position
*P[+voice]	no underlyingly voiced bilabial stops	−173.5
*T[+voice]	no underlyingly voiced alveolar stops	−217.5
*S[+voice]	no underlyingly voiced aveolar fricatives	−512.2
*F[−voice]	no underlyingly voiceless labiodental fricatives	−515.4
*X[−voice]	no underlyingly voiceless velar fricatives	−515.6
*V:O[−voice]	no underlyingly voiceless obstruents following long vowels	−516.4
*iuyO[−voice]	no underlyingly voiceless obstruents following [i, u, y]	−516.7
*VO[+voice]	no underlyingly voiced obstruents following short vowels	−517.0
*SonO[−voice]	no underlyingly voiceless obstruents following sonorants	−1300.1
*OO[+voice]	no underlyingly voiced obstruents following obstruents	−1302.1

ferred for this specific data set. SOT implements a greedy learning strategy, in that the individual examples to which the model is exposed are discarded. It is a memoryless system that presupposes that it is known beforehand which constraints might be relevant. The spreading activation model, by contrast, crucially depends on having in memory the phonological representations of Dutch monomorphemic lexemes. Interestingly, this requirement does not add to the complexity of the grammar, as the phonological form of monomorphemic words must be stored in the lexicon anyway. Since there is no intrinsic advantage to greedy learning for these data, Occam's razor applies in favor of the spreading activation model as the more parsimonious theory, at least for this data set.

The interactive activation model is in turn challenged by a lazy learning model with no parameters at all, Analogical Modeling of Language (AML; Skousen 1989, 1992). In what follows, I give a procedural introduction to AML. For a rigorous mathematical analysis of the statistical properties of AML, see Skousen 1992; and for the use of this natural statistic in quantum physics, see Skousen 2000.

Like the interactive activation model, AML requires an input lexicon that specifies for each lexeme the values of a series of features describing

Table 7.3
Feature specifications in the lexicon for Analogical Modeling of Language

Lexeme	Onset	Vowel type	Vowel	Coda	Obstruent	Voicing
aap	empty	long	a	none	P	voiceless
aard	empty	long	a	sonorant	T	voiced
aars	empty	long	a	sonorant	S	voiced
aas	empty	long	a	none	S	voiced
abrikoos	k	long	o	none	S	voiced
abt	empty	short	A	obstruent	T	voiceless
accent	s	short	E	sonorant	T	voiceless
accijns	s	short	K	sonorant	S	voiced
accuraat	r	short	a	none	T	voiceless
acht	empty	short	A	obstruent	T	voiceless
...	...	...	...	...	...	...
zwijg	zw	long	K	none	X	voiced
zwoeg	zw	iuy	u	none	X	voiced
zwoerd	zw	iuy	u	sonorant	T	voiced

its final syllable together with the underlying voice specification of the final obstruent. Table 7.3 lists some examples of an instance base with features Onset, Vowel Type, Vowel (using the DISC computer phonetic alphabet), Coda structure (whether a prefinal consonant is present and, if so, whether it is a sonorant or a stop), and (final) Obstruent.

When AML has to predict the voice specification for a pseudoverb such as *puig*, it considers the exemplars in its lexicon for all possible supracontexts of the target. A supracontext of the target is the set of exemplars (possibly empty) that share a minimum number (possibly even zero) of feature values with the target. Table 7.4 lists all the supracontexts of *puig*. The first supracontext has distance 0: the values of all its features are fixed. That is, in order for a word to belong to this supracontext, it must share *puig*'s values for all five of its features. Because *puig* is not a real word of Dutch, this fully specified supracontext is empty. The next 5 supracontexts have distance 1. They contain the exemplars that share four feature values and that differ from the target at (at most) one position. This position is indicated by a hyphen in table 7.4. The next 10 supracontexts cover the sets of exemplars that have two variable positions. The final supracontext has five variable positions. As we move down table 7.4, the supracontexts become less specific in their similarity

Table 7.4
Supracontexts for the pseudoverb *puig*

Supracontext					Distance	Voiced	Voiceless	Homogeneity
p	long	L	none	x	0	0	0	empty
–	long	L	none	x	1	7	1	**homogeneous**
p	–	L	none	x	1	0	0	empty
p	long	–	none	x	1	1	0	**homogeneous**
p	long	L	–	x	1	0	0	empty
p	long	L	none	–	1	0	1	**homogeneous**
–	–	L	none	x	2	7	1	**homogeneous**
–	long	–	none	x	2	65	1	heterogeneous
–	long	L	–	x	2	7	1	**homogeneous**
–	long	L	none	–	2	38	34	heterogeneous
p	–	–	none	x	2	2	1	heterogeneous
p	–	L	–	x	2	0	0	empty
p	–	L	none	–	2	0	1	**homogeneous**
p	long	–	–	x	2	1	0	**homogeneous**
p	long	–	none	–	2	4	8	heterogeneous
p	long	L	–	–	2	0	2	**homogeneous**
–	–	–	none	x	3	107	4	heterogeneous
–	–	L	–	x	3	7	1	**homogeneous**
–	–	L	none	–	3	38	34	heterogeneous
–	long	–	–	x	3	65	1	heterogeneous
–	long	–	none	–	3	261	188	heterogeneous
–	long	L	–	–	3	38	38	heterogeneous
p	–	–	–	x	3	2	1	heterogeneous
p	–	–	none	–	3	7	34	heterogeneous
p	–	L	–	–	3	0	2	**homogeneous**
p	long	–	–	–	3	7	11	heterogeneous
–	–	–	–	x	4	126	5	heterogeneous
–	–	–	none	–	4	409	636	heterogeneous
–	–	L	–	–	4	38	38	heterogeneous
–	long	–	–	–	4	300	231	heterogeneous
p	–	–	–	–	4	13	59	heterogeneous
–	–	–	–	–	5	583	1,101	heterogeneous

requirements and contain nondecreasing numbers of exemplars. The columns labeled *Voiced* and *Voiceless* tabulate the number of exemplars in a given context that have the corresponding voice specification. Thus, there are 8 exemplars in the second supracontext, 7 of which are underlyingly voiced. The last, most general supracontext at the bottom of the table covers all 1,684 words in the lexicon, of which 583 are voiced.

Supracontexts can be *deterministic* or *nondeterministic*. A supracontext is deterministic when all its exemplars support the same voice specification (e.g., *p long L – –*); otherwise, it is nondeterministic (e.g., *– long L none –*). When predicting the voice specification for a new word that is not yet in the model's instance base, AML inspects only those supracontexts that are *homogeneous*. All deterministic supracontexts are homogeneous. A nondeterministic supracontext is homogeneous only when all more specific supracontexts that it contains have exactly the same distribution for the voice specification. Consider, for example, the nondeterministic supracontext *– long L – x*, which has the outcome distribution (7, 1). It contains the more specific supracontext *– long L none x*. This supracontext is more specific because the fourth feature has the specific value none. This supracontext is also nondeterministic, and it has the same outcome distribution (7, 1). The supracontext *– long L – x* has one other more specific supracontext, *p long L – x*. This supracontext is the empty set and does not count against the homogeneity of the more general supracontexts of which it is a subset. Therefore, the supracontext *– long L – x* is homogeneous. It is easy to see that the nondeterministic supracontext *– long L None –* is heterogeneous, as there is no other more specific supracontext with the distribution (38, 34). The homogeneous contexts jointly constitute the analogical set on which AML bases its prediction. Intuitively, one can conceptualize the homogeneity of a supracontext as indicating that there is no more specific information (in the form of a more fully specified supracontext) with contradicting distributional evidence. In other words, distributional evidence tied to more specific supracontexts blocks contradicting distributional evidence from less specific supracontexts from having an analogical contribution.

Table 7.5 lists the exemplars that appear in the analogical set. AML offers two ways for calculating the probabilities of the outcomes (voiced or voiceless), depending on whether the size of the supracontexts is taken into account. In occurrence-weighted selection, the contribution of an exemplar (its similarity score in the model) is proportional to the count of different supracontexts in which it occurs. The third column of table 7.5

Table 7.5
The exemplars predicting the voice specification for the pseudoverb *puig*. The column labeled *Composition* specifies the sizes of the supracontexts to which an exemplar belongs.

		Occurrence weighted		Size weighted		
Exemplar	Voicing	Count	Contribution	Count	Composition	Contribution
buig	voiced	4	0.10	32	8-8-8-8	0.12
duig	voiced	4	0.10	32	8-8-8-8	0.12
huig	voiced	4	0.10	32	8-8-8-8	0.12
juich	voiceless	4	0.10	32	8-8-8-8	0.12
poog	voiced	2	0.05	2	1-1	0.01
puist	voiceless	2	0.05	4	2-2	0.02
puit	voiceless	4	0.10	6	2-2-1-1	0.02
ruig	voiced	4	0.10	32	8-8-8-8	0.12
luig	voiced	4	0.10	32	8-8-8-8	0.12
vuig	voiced	4	0.10	32	8-8-8-8	0.12
zuig	voiced	4	0.10	32	8-8-8-8	0.12
P(voiced)			0.75			0.84

lists these counts, and the fourth column the proportional contributions. The probability of a voiced realization using occurrence-weighted selection is .75, as the summed count for the voiced exemplars equals 25 out of a total score of 100. Applied to all experimental pseudoverbs, the maximum likelihood choice of AML with occurrence-weighted selection agrees with the majority choice of the participants in 166 out of 192 cases, an accuracy score of 86% that does not differ significantly from the accuracy score of 91% obtained with the spreading activation model ($X^2(1) = 1.6761$, $p = .1954$).

When we use size-weighted selection, the contribution of an exemplar is proportional to the sum of the sizes of the supracontexts in the analogical set in which it appears. This size-weighted selection amounts to using a squaring function for measuring agreement, similar to the quadratic measure of agreement found in Schrödinger's wave equation (Skousen 2000). The exemplar *buig*, for instance, occurs in four homogeneous supracontexts, the homogeneous supracontexts in table 7.4 with the (7, 1) distribution. The size of each of these four supracontexts is 8; hence, *buig*

now contributes a count of 32 instead of 4. In the case of the exemplar *poog*, the two homogeneous supracontexts in which it appears both have a size of 1. Hence, the contribution of *poog* remains proportional to a count of 2. The probability of a voiced realization using size-weighted selection is .84. The accuracy score of AML with respect to the complete set of experimental pseudoverbs is again 86%. Although AML seems to be slightly less accurate than the spreading activation model, the fact that AML is a parameter-free model (i.e., a model with no parameters that the analyst can tweak to get the model to better fit the data) makes it a very attractive alternative.

It is important to realize that AML bases its predictions on the local similarity structure in the lexicon given a target input. There are no global computations establishing general weights that can subsequently be applied to any new input. It is not necessary to survey the instance base and calculate the information gain weight for, say, the onset. Likewise, it is not necessary to establish a priori whether constraints pertaining to the onset should or should not be brought into a stochastic optimality grammar. (The only requirement is the a priori specification of a set of features and their values, but this minimal requirement is a prerequisite for any current theory.) What I find interesting is that the microstructure of local similarities as captured by the analogical set of AML is by itself sufficient to capture the support in the language for the voice specification for a given target word. The absence of a role for the onset follows without further specification from the fact that the supracontexts containing the onset (the supracontexts with an initial *p* in table 7.4) are either very sparsely populated or heterogeneous. Although generalizations in the form of abstract rules may provide a good first approximation of morphological regularities, for more precise prediction it is both necessary and, surprisingly, sufficient to take into account the microstructure of the similarity space around individual words. Global similarity structure is grounded in the local similarity structure around individual words.

7.3.3 Discussion

The case studies surveyed in this section have a number of interesting consequences for linguistic theory. A first observation concerns the notion of productivity. Regularity, of the kind that can be captured by symbolic rules, is often seen as a necessary condition for productivity. The Dutch linking elements, however, are productive without being regular in this sense. Similarly, the morphophonological voicing alternation of obstru-

ents in Dutch also enjoys a fair degree of productivity, even though standard analyses have treated it as idiosyncratic and lexically specified. To understand the basis of productivity, the paradigmatic, probabilistic dimension of morphological structure is crucial.

A second observation is that rejecting syntagmatic symbolic rules as the appropriate framework for the analysis of a given morphological phenomenon does not imply embracing subsymbolic connectionism. The present examples show that a symbolic approach in which paradigmatic structure provides a similarity space over which probabilities are defined can provide an excellent level of granularity for understanding the role of probability in language production. This is not to say that the present data sets cannot be modeled by means of subsymbolic ANNs. On the contrary, ANNs are powerful nonlinear classifiers, whereas the classification problems discussed in this section are trivial compared to the classification problems that arise in, for instance, face recognition. ANNs have the disadvantage that they require large numbers of parameters (the weights on the connections) that themselves reveal little about the linguistic structure of the data, unlike the information gain weights in the spreading activation model. The hidden layers in an ANN often provide a compressed re-representation of the structure of the data, but the cost in terms of the number of parameters and the complexity of the training procedure are high. And it is not always clear what one has learned when a three-layer network successfully maps one type of representation onto another (see Forster 1994). For those who take the task of morphological theory as part of linguistics to be to provide the simplest possible account for (probabilistic) phenomena in word formation, ANNs are probably not the analytically most insightful tool to use. However, those who view morphology as part of cognitive science may gladly pay the price of greater analytical complexity, especially when more biologically realistic neural network models become available.

A third observation concerns the notion of learning. Traditional symbolic approaches such as the one advocated by Pinker (1991) are based on the a priori assumption that greedy learning is at the core of the language faculty. The Gradual Learning Algorithm of Boersma (1998) is in line with this tradition: occurrences leave a trace in the positions of the constraints; they themselves need not be stored in memory. TIMBL and AML, by contrast, are based on lazy learning, with extensive storage of exemplars in memory and similarity-based reasoning taking the place of abstract rules.

A question that arises here is to what extent these models provide a reasonable window on language acquisition. It should be noted that all models discussed here share, in their simplest form, the assumption that it is known at the outset which features are relevant and what values these features can assume, and that this knowledge is constant and not subject to development over time. This is a strong and unrealistic assumption. Granted this assumption, SOT, TIMBL, and AML can all approximate acquisition as a function of the input over time. Given which constraints are relevant for a given linguistic mapping, SOT can chart how the positioning of constraints develops over time. What TIMBL requires for modeling classificatory development is a continuous reevaluation of the information gain weights as the instance base is increased with new exemplars. AML, by contrast, predicts changing classificatory behavior as a function of a changing lexicon without further assumptions, a property it shares with connectionist models of language acquisition.

A related final question concerns whether there are differences in the extent to which different models depend on a priori assumptions. All models reviewed here, including SOT and AML, do not differ with respect to the minimal levels of representation they require. The differences between these models concern how they make use of these representations and what happens with the examples from which a given mapping has to be learned. Both TIMBL and AML instantiate lazy learning algorithms that do not require any further a priori knowledge. The same holds for connectionist models. SOT instantiates a greedy learning algorithm, an algorithm that does require the a priori knowledge of which constraints are potentially relevant, or, at the very least, that does require considerable hand-crafting in practice. For instance, there is no constraint *iuyF[+voice] in the SOT model summarized in table 7.2, simply because it turned out to be unnecessary given the other constraints that had already been formulated. Theoretically, it might be argued that every combination of feature values (e.g., T, or SonO) and outcome values (e.g., [−voice]) is linked automatically with a (supposedly innate, universal) constraint, with irrelevant constraints dropping to the bottom of the grammar during learning. Under this view, SOT would not depend on a priori knowledge either, although such an SOT grammar is encumbered with an enormous pile of useless constraints lying inactive at the bottom of the ranking. Note, however, that TIMBL and AML are encumbered in a different way, namely, with useless features that are themselves harmless but that render the on-line calculation of the simi-

larity space more complex. Similarly in connectionist networks, such useless features add noise to the system, delaying learning and slowing down convergence.

Summing up, from a statistical point of view, SOT, AML, and TIMBL, and connectionist models as well, all have their own advantages and disadvantages as explanatory frameworks for the data sets discussed. One's choice of model will in practice be determined by one's view of the explanatory value of these models as instantiations of broader research traditions (Optimality Theory, machine learning, and cognitive science) across a much wider range of data sets.

7.4 Probability in Morphological Comprehension

In the previous section, we saw how probabilities based on counts of word *types* falling into different similarity classes play a role in the production of morphologically complex words. In this section, we will see that probabilities based on *token* frequencies play a crucial role in solving the ambiguity problem in morphological segmentation. Consider the examples in (10):

(10)

acute+ness	32	a+cuteness	39
expert+ness	25	ex+pert+ness	68
intent+ness	29	in+tent+ness	57
perverse+ness	31	per+verse+ness	66
sacred+ness	27	sac+redness	56
tender+ness	28	tend+er+ness	55
prepared+ness	19	prep+a+red+ness	58

The first column lists correct segmentations for a number of words with the suffix *-ness*; the third column lists some incorrect or implausible segmentations. I will explain the interpretation of the numbers in the second and fourth columns below. How do we know, upon reading a string such as *preparedness*, which of the segmentations in (11) is the one to choose?

(11)

preparedness	19
prepared+ness	27
pre+pared+ness	56
prep+a+red+ness	58
prep+a+redness	58
pre+par+ed+ness	71
pre+pa+red+ness	78

pre+pa+redness 78
pre+pare+d+ness 78
prep+are+d+ness 78
prepare+d+ness 78

Some of these segmentations can be ruled out on the basis of combinatorial restrictions. For instance, the form *are* (2 sg., 1/2/3 pl. present of *be*, or singular of the noun denoting an area of 100 m^2) in *prep+are+d+ness* does not combine with the suffix *-d*. Other segmentations in (11), however, are possible albeit implausible. For example, *((pre((pare)d))ness)* has the same structure as *((pre((determine)d))ness)*, but it is unlikely to be the intended reading of *preparedness*. Why are such forms implausible? In other words, why are their probabilities so low? In various computational approaches (e.g., probabilistic context-free grammars, probabilistic head-lexicalized grammars, and Data-Oriented Parsing, as described in chapter 2), the probability of the whole, *P*(preparedness), is obtained from the probabilities of combinations of its constituents, some of which are listed in (12):

(12) *P*(prepared, ness)
P(pre, pared)
P(pared, ness)
P(pre, pared, ness)
...
P(prepare, d)
P(d, ness)
P(prepare, d, ness)

Crucially, these probabilities are estimated on the basis of the relative token frequencies of these bigrams and trigrams in large corpora, with the various approaches using different subsets of probabilities and combining them according to different grammars (for an application to Dutch morphology, see Heemskerk 1993; for a memory-based segmentation system using TIMBL, see van den Bosch, Daelemans, and Weijters 1996). In this section, I consider how human morphological segmentation might be sensitive to probabilities of combinations of constituents.

It is important to realize that if the brain does indeed make use of probabilities, then it must somehow keep track of (relative) frequency information for both irregular and completely regular complex words. The issue of whether fully regular complex words are stored in the mental lexicon, however, is hotly debated. Pinker (1991, 1997), Marcus et al.

Table 7.6
Frequencies in a corpus of 42 million tokens as available in the CELEX lexical database of the singular, plural, and diminutive forms of the Dutch nouns *tong* 'tongue', *gast* 'guest', and the corresponding probabilities

Word	Inflection	Frequency	Probability
tong	singular	2,085	4.96e-05
tongen	plural	179	0.43e-05
tongetje	diminutive	24	0.06e-05
		2,288	5.45e-05
gast	singular	792	1.89e-05
gasten	plural	1,599	3.80e-05
gastje	diminutive	0	0
		2,391	5.69e-05

(1995), Clahsen, Eisenbeiss, and Sonnenstuhl-Henning (1997), and Clahsen (1999) have argued that frequency information is stored in the brain only for irregular complex words, and not at all for regular complex words. They argue that regular morphology is subserved by symbolic rules that are not sensitive to the frequencies of the symbols on which they operate, and that irregular morphology is subserved by a frequency-sensitive associative storage mechanism.

The only way in which probabilities might play a role for regular complex words in this dual-route model is at the level of the rules themselves; that is, rules might differ with respect to their probability of being applied. Consider table 7.6, which lists the frequencies of the singular, plural, and diminutive forms of the Dutch nouns *tong* 'tongue' and *gast* 'guest', as listed in the CELEX lexical database. By assigning different probabilities to the rules for diminutivization and pluralization, this approach can account for the lower probabilities of diminutives compared to plurals. However, it cannot account for the differences in the probabilities of the two plural forms. Even though the lexemes *tong* and *gast* have very similar probabilities, the former occurs predominantly in the singular, and the latter predominantly in the plural. I will refer to *tong* as a *singular-dominant noun* and to *gast* as a *plural-dominant noun*. What the dual-route model predicts is that such differences in frequency dominance are not registered by the brain, and hence that such differences do not affect lexical processing. Note that this amounts to the claim that the

brain has no knowledge of the probability that a particular noun co-occurs with the plural suffix. The only frequency count that should be relevant in the dual-route approach is the stem frequency, the summed frequency of the singular and the plural forms.

There is massive evidence, however, that the claim that frequency information is retained by the brain for irregular words only is incorrect (see, e.g., Taft 1979; Sereno and Jongman 1997; Bertram et al. 1999; Bertram, Schreuder, and Baayen 2000; Baayen et al. 2002). These studies show that probabilistic information in the form of knowledge of co-occurrence frequencies of constituents forming complex words must be available to the brain. But how then might the brain make use of this information? Although I think that formal probabilistic models provide excellent mathematically tractable characterizations of what kinds of knowledge are involved in morphological segmentation, I do not believe that, for example, the algorithms for estimating the parameters of hidden markov models can be mapped straightforwardly onto the mechanisms used by the brain. The way the brain solves the ambiguity problem may well be more similar to a dynamic system in which a great many inter-dependent morphological units compete to provide a segmentation that spans the target word in the input. In what follows, I will outline a model implementing a simple dynamic system, and I will show that it provides a reasonable framework for understanding some interesting aspects of the processing of Dutch and German plural nouns.

7.4.1 A Dynamic System for Morphological Segmentation

Whereas computational parsing models in linguistics have successfully used token-count-based probabilities of occurrence, psycholinguistic research on the segmentation problem has focused on the role of form similarity. In the Shortlist model of the segmentation of the auditory speech stream (Norris 1994), for instance, the lexical representations that are most similar to the target input are wired into a connectionist network implementing a similarity-based competition process. The resulting model is very sensitive to differences in form between lexical competitors, and captures important aspects of auditory lexical processing. However, Shortlist does not address how to account for the word frequency effect in auditory word recognition (see, e.g., Rubenstein and Pollack 1963).

MATCHECK (Baayen, Schreuder, and Sproat 2000; Baayen and Schreuder 1999, 2000) implements an approach to the segmentation problem in which form similarity and token-frequency-based probability

simultaneously play a role. This model is a dynamic system articulated within the interactive activation framework. It shares with dynamic systems in general the properties that its behavior depends crucially on the initial condition of the model, that it is deterministic, and that there is order in what seems to be chaotic behavior. The components of the model are an input lexicon, a mechanism for ascertaining whether a lexical representation should be taken into account as a candidate for a segmentation, and a competition mechanism. In what follows, I outline the architecture of MATCHECK for the visual modality.

The input lexicon contains form representations for stems, affixes, and full forms, irrespective of their regularity. Each representation w has an initial activation level $a(w, 0)$ equal to its frequency in a corpus. The initial probability of a lexical representation $p_{w,0}$ is its relative frequency in the lexicon.

The mechanism for determining whether a lexical representation should be taken into account as a possible constituent in a segmentation makes use of an activation probability threshold $0 < \theta < 1$. Only those lexical representations with a probability $p_{w,t} \geq \theta$ at timestep t are candidates for inclusion in a segmentation.

The competition mechanism consists of a probability measure imposed on the activation levels of the lexical representations, combined with a similarity-based function that determines whether the activation of a given lexical representation should increase or decrease. The activation probability of representation w at timestep t is

$$p_{w,t} = \frac{a(w,t)}{\sum_{i=1}^{V} a(w_i,t)}, \tag{11}$$

with V the number of representations in the lexicon. The details of the criteria for whether the activation of a lexical representation increases or decreases need not concern us here. What is crucial is that a lexical representation that is aligned with one of the boundaries of the target word is allowed to increase its activation until its activation probability has reached the threshold θ. Once this threshold has been reached, activation decreases. Given a decay rate δ_w $(0 < \delta_w < 1)$ for representation w, the change in activation from one timestep to the next is defined as

$$a(w,t) = a(w,0) + \delta_w\{a(w,t-1) - a(w,0)\}. \tag{12}$$

Because $\delta_w < 1$, the activation at each successive timestep becomes smaller than it was at the preceding timestep. Asymptotically, it will

decrease to its original resting activation level. Activation increase, which occurs at the timesteps before w has reached threshold, is also defined in terms of δ_w,

$$a(w,t) = \frac{a(w,t-1)}{\delta_w}, \tag{13}$$

but now we divide by δ_w instead of multiplying by δ_w. Consequently, the activation at timestep t becomes greater than the activation at timestep $t-1$. Because activation increase and activation decrease are both defined in terms of δ_w, words with a decay rate close to one are slow to decay and slow to become activated. Conversely, words that decay quickly (δ_t close to zero) also become activated quickly. Note that if the activation level of a lexical representation increases while the activation levels of the other representations remain more or less unchanged, its probability increases as well.

The key to the accuracy of MATCHECK as a segmentation model lies in the definition of the activation and decay parameter δ_w, which differs from representation to representation. It is defined in terms of the frequency f_w and the length L_w of a lexical representation w, in combination with three general parameters (α, δ, ζ), as follows:

$$\delta_w = f(g(\delta, \alpha, w), \zeta, \delta, w), \tag{14}$$

with

$$g(\delta, \alpha, w) = \delta \frac{1}{1 + \alpha \dfrac{\log(L_w + 1)}{\log(f_w)}}, \tag{15}$$

and, with T denoting the target word, and using d as shorthand for $g(\delta, \alpha, w)$,

$$f(d, \zeta, \delta, w) = \begin{cases} d + (1-d)\left(\dfrac{|L_w - L_T|}{\max(L_w, L_T)}\right)^{\zeta} & \text{iff } \zeta > 0 \\ \delta & \text{otherwise.} \end{cases} \tag{16}$$

The parameter δ $(0 < \delta < 1)$ denotes a basic activation and decay rate that is adjustable for each individual word. Lexical representations with higher frequencies receive higher values for δ_w. They have a reduced information load and become activated less quickly than lower-frequency representations. The parameter α $(\alpha \geq 0)$ specifies how the frequency and length of the representation should be weighted, independently of its

similarity to the target word. The parameter ζ ($\zeta \geq 0$) determines to what extent the relative difference in length of a lexical competitor and the target word should affect the activation and decay rate of w. Increasing α or ζ leads to a decrease of δ_w and hence to more rapid activation and decay. Finally, constituents that are more similar in length to the target word have a smaller δ_w than shorter constituents (for experimental evidence that longer affixes are recognized faster than shorter affixes, see Laudanna, Burani, and Cermele 1994).

Figure 7.6 illustrates the activation dynamics of MATCHECK. On the horizontal axis, it plots the timesteps in the model. On the vertical axis, it plots the activation probability. The solid horizontal line represents the activation probability threshold $\theta = 0.3$. Consider the left panel, which plots the activation curves for the singular-dominant plural *tongen*. The curved solid line represents the plural suffix *-en*, which, because of its high frequency, has a high initial probability. During the first timesteps, the many words with partial similarity to the target, such as *long* 'lung', become activated along with the constituents themselves. Hence, although the activation of *-en* is actually increasing, its activation probability decreases. The base *tong* starts out with a very low initial probability, but because of its greater similarity to the plural form, it reaches threshold long before *-en*. Upon reaching threshold, its activation begins to decay. The subsequent erratic decay pattern for *tong* is due to the interference of a lexical competitor, not shown in figure 7.6, the orthographic neighbor of the plural form, *tonnen* (see, e.g., Andrews 1992, Grainger 1990, and Grainger and Jacobs 1996, for experimental evidence for orthographic neighborhood effects). Thanks to the probability measure imposed on the activations of the lexical representations, the decay of the activation of *tong* makes activation probability available for the lower-frequency plural form, which can now reach threshold as well. This point in model time (timestep 19) is represented by a vertical solid line. This is the first timestep in the model at which a full spanning of the input is available. In other words, the first "parse" to become available for *tongen* is the plural form. Once both the singular and plural forms have entered activation decay, the plural affix finally reaches threshold. It is only now that stem and suffix provide a full spanning of the input, so the second parse to become available arrives at timestep 41. The following erratic activation bumps for the base noun arise because of the difference in the speed with which the singular and plural forms decay, and they have no theoretical significance.

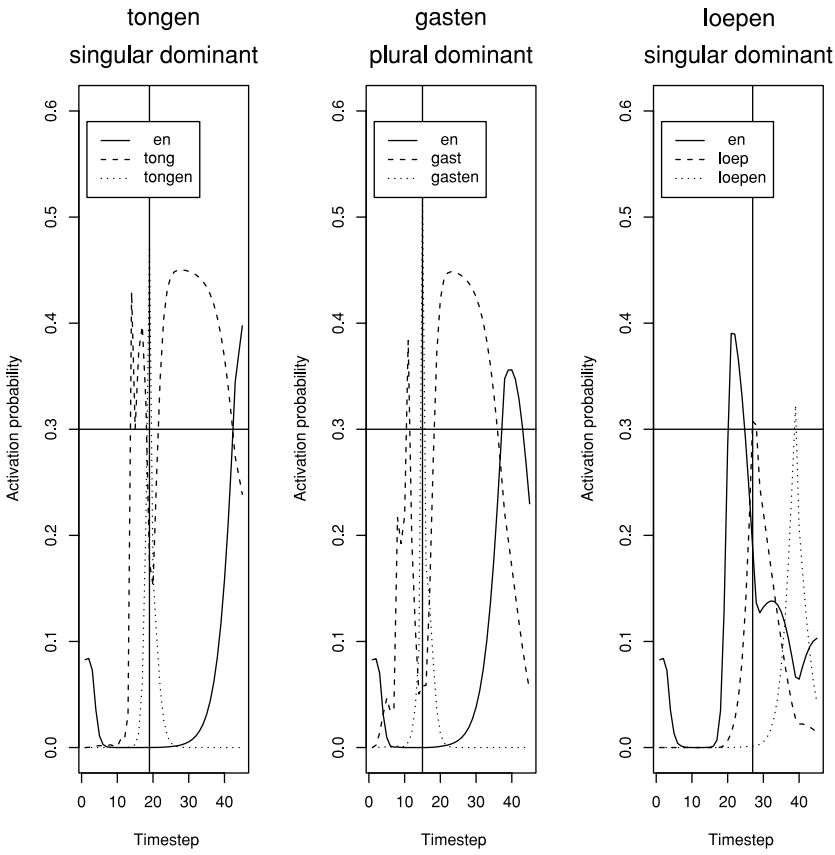

Figure 7.6
Patterns of activation probabilities over time in MATCHECK for the Dutch singular-dominant plural *tongen* 'tongues' (*left*), the matched plural-dominant plural *gasten* 'guests' (*center*), and the low-frequency singular-dominant plural *loepen* 'magnification glasses' (*right*), with threshold at .3. The first full spannings of the input are *tongen* (t = 19), *gasten* (t = 15), and *loep*+*en* (t = 27).

The center panel in figure 7.6 reveals a similar pattern for the plural-dominant plural *gasten*. However, its representations reach threshold at an earlier timestep. The singular *tong* reaches threshold at timestep 13 and its plural *tongen* becomes available at timestep 18, whereas the corresponding model times for *gast* and *gasten* are 10 and 14, respectively.

Finally, the right panel of figure 7.6 shows the timecourse development for a very low-frequency singular-dominant noun plural, *loepen* 'magnification glasses'. Note that the first constituent to reach threshold is the plural suffix, followed by the base, which jointly provide a full spanning of the target long before the plural form reaches threshold.

Two questions arise at this point. First, how well does MATCHECK solve the ambiguity problem? Second, how good is MATCHECK at modeling actual processing times in psycholinguistic experiments? The first question is addressed by Baayen and Schreuder (2000). They showed, for instance, that for 200 randomly selected words of lengths 5–8 (in letters) with on average three incorrect segmentations, the first segmentation to be produced by MATCHECK was correct in 194 cases, of which 92 were due to the full form being the first to become available, and 102 to a correct segmentation becoming available first. The model times listed in examples (10) and (11) illustrate how MATCHECK tends to rank correct segmentations before incorrect segmentations, using the parameter settings of Baayen and Schreuder (2000). For instance, the model times listed for *preparedness* in (11) show that the full form *preparedness* and the correct parse *prepared-ness* become available at timesteps 19 and 27, respectively, long before the first incorrect parse *pre-pared-ness* (timestep 56) or the complete correct segmentation *prepare-d-ness* (timestep 78). Because the model gives priority to longer constituents, a parse such as *prepared-ness*, which is based on the regular participle *prepared*, becomes available before the many incorrect or implausible parses containing shorter constituents. The presence of the regular participle *prepared* in the lexicon protects the model against having to decide between alternative segmentations such as *prepare-d-ness* and *pre-pared-ness*. In other words, paradoxically, storage enhances parsing. In MATCHECK, storage in memory does not just occur for its own sake; its functionality is to reduce the ambiguity problem.

We are left with the second question, namely, whether the model times produced by MATCHECK have any bearing on actual human processing latencies. The next two sections address this issue by means of data sets from Dutch and German.

7.4.2 Regular Noun Plurals in Dutch

Baayen, Dijkstra, and Schreuder (1997) studied regular Dutch plurals in *-en* and their corresponding singulars using visual lexical decision. The visual lexical decision task requires participants to decide as quickly and accurately as possible whether a word presented on a computer screen is a real word of the language. Response latencies in visual lexical decision generally reveal strong correlations with the frequencies of occurrence of the words. This particular study made use of a factorial experimental design contrasting three factors: Stem Frequency (high vs. low), Number (singular vs. plural), and Dominance (singular-dominant vs. plural-dominant). Within a stem frequency class, the average stem frequency was held constant in the mean, as illustrated in table 7.6 for the nouns *tong* and *gast*. The right panel of figure 7.7 provides a graphical summary

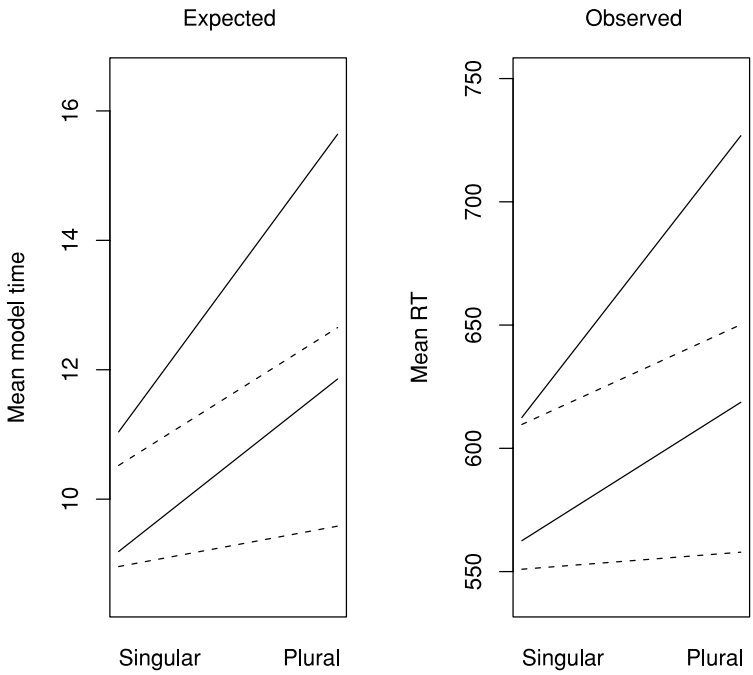

Figure 7.7
Predicted mean model times (*left*) and observed mean reaction times (*right*) for Dutch regular noun plurals in *-en* cross-classified by Number (singular vs. plural), Stem Frequency (high vs. low), and Frequency Dominance (singular-dominant (solid lines) vs. plural-dominant (dashed lines))

of the pattern of results. The horizontal axis contrasts singulars (left) with plurals (right). The dashed lines represent plural-dominant singulars and plurals. The solid lines represent singular-dominant singulars and plurals. The lower two lines belong to the nouns with a high stem frequency, and the upper two lines to the nouns with a low stem frequency.

What this graph shows is that high-frequency nouns, irrespective of number, are processed faster than low-frequency nouns. It also reveals a frequency effect for the plural forms. For each of the two stem frequency conditions, singular-dominant plurals are responded to more slowly than plural-dominant plurals. Speakers of Dutch are clearly sensitive to how often the plural suffix *-en* co-occurs with particular noun singulars to form a plural. Especially the fast response latencies for plural-dominant plurals bear witness to the markedness reversal studied by Tiersma (1982): while normally the singular is the unmarked form both with respect to its phonological form and with respect to its meaning, many plural-dominant plurals are semantically unmarked compared to their corresponding singulars (see Baayen, Dijkstra, and Schreuder 1997, for further discussion). This plural frequency effect, however, is completely at odds with the dual-route model advanced by Pinker and Clahsen and their coworkers. Finally, figure 7.7 shows that for singular nouns it is the frequency of the lexeme (i.e., the summed frequencies of the singular and the plural forms) that predicts response latencies, and not so much the frequencies of the singular forms themselves. If the frequencies of the singulars as such had predicted the response latencies, one would have expected to see a difference in the response latencies between singular-dominant and plural-dominant singulars. The observed pattern of results has been replicated for Dutch in the auditory modality (Baayen et al., in press) and for Italian in the visual modality (Baayen, Burani, and Schreuder 1997).

The left panel of figure 7.7 summarizes the results obtained with MATCHECK for the parameter settings $\delta = 0.3$, $\theta = 0.25$, $\alpha = 0.4$, $\zeta = 0.3$, and $\rho = 3$. This last parameter determines the granularity of the model, that is, the precision with which the timestep at which the threshold is reached is ascertained (see Baayen, Schreuder, and Sproat 2000). The lexicon of the model contained the 186 experimental words and in addition some 1,650 randomly selected nouns as well as various inflectional and derivational affixes. The frequencies of the affixes were set to the summed frequencies of all the words in the CELEX lexical database in which they occur. Singular nouns received the summed frequencies of

the singular and plural forms as frequency count, the assumption being that in a parallel dual-route model the processing of the (globally marked) plural leaves a memory trace on the (globally unmarked) singular. Plural nouns were assigned their own plural frequency. The lexicon also contained a dummy lexeme consisting of a series of *x* characters, which received as frequency count the summed frequencies of all words in the CELEX database that were not among the 1,650 randomly selected nouns. This ensured that all words in the lexicon had initial probabilities identical to their relative frequencies in the corpus on which the CELEX counts are based.

Interestingly, it is impossible to obtain the pattern shown in the left panel of figure 7.7 with just these settings. The reason for this is that the singular form, thanks to its cumulated frequency, is a strong competitor of the plural form. Although plural-dominant plurals reach threshold before singular-dominant plurals with the same stem frequency, as desired, they also reach threshold well after the corresponding singulars. Indistinguishable processing times for singulars and plurals, as observed for plural-dominant singulars and plurals in the high stem frequency condition in the experiment, cannot be simulated.

The adaptation of MATCHECK that leads to the pattern of results actually shown in the left panel of figure 7.7 is to enrich the model with a layer of lexemes in the sense of Aronoff (1994) or lemmas in the sense of Levelt (1989). The representations of the singular and plural forms have pointers to their lexeme, which in turn provides pointers to its associated semantic and syntactic representations. The lexemes serve a dual function in MATCHECK. Their first function is to accumulate in their own activation levels the summed activation levels of their inflectional variants. Once the lexeme has reached threshold activation level, a response can be initiated. This allows the model to take into account the combined evidence in the system supporting a given lexeme. The second function of the lexemes is to pass on part of the activation of the (marked) plural form to the (unmarked) singular form. I assume that it is this process that leaves a memory trace with the singular that results over time in the summed frequency of the singular and the plural form being the predictor of response latencies for singulars. To implement this modification, we have to revise the definition of increasing activation given in equation (13) for the representation of the singular as follows. Let w_s denote the representation of a singular, and let i range over its n inflectional variants. For

Table 7.7
F-values for 1 and 178 degrees of freedom and the corresponding *p*-values for the model times (expected) and the reaction times (observed)

	Expected		Observed	
	F	*p*	*F*	*p*
Number	64.1	.000	46.4	.000
Dominance	27.3	.000	27.1	.000
StemFreq	65.7	.000	90.4	.000
Number : Dominance	13.9	.000	15.8	.000
Number : StemFreq	7.3	.007	7.9	.005
Dominance : StemFreq	0.6	.434	0.1	.828
Number : Dominance : StemFreq	0.1	.459	0.6	.451

the present data, $n = 1$, as there is only one inflectional variant other than the singular itself, namely, the plural. We define $a(w_s, t)$ as follows:

$$a(w_s, t) = \frac{a(w_s, t-1)}{\delta_{w_s}} + \log(a(w_s, t-1)) \sum_{i=1}^{n} a(w_i, t)^{\lambda}, \tag{17}$$

with λ the parameter determining how much activation flows from the plural to the singular. In the simulation leading to the left panel of figure 7.7, λ was set to 1.1.

How well does this model approximate the observed patterns in the experimental data? A comparison of the left and right panels of figure 7.7 suggests that the model provides a good fit, an impression that is confirmed by table 7.7: the same main effects and interactions that are significant in the experiment are also significant according to the model. The model also captures that high-frequency plural-dominant singulars and plurals are processed equally fast ($t(45.99) = -0.82$, $p = .416$ for the response latencies, $t(35.44) = -0.86$, $p = .398$ for the model times, Welch two-sample t-tests), in contrast to the low-frequency plural-dominant singulars and plurals ($t(41.65) = -2.38$, $p = .022$ for the response latencies, $t(32.33) = -4.394$, $p = .0001$ for the model times).

MATCHECK provides this good fit to the experimental data with parameter settings that make it make the most of the available full-form representations. The only word for which a parse into base and affix reached threshold before the full form was the plural form *loepen*, for which the activation probabilities over time were shown in the right panel of figure 7.6. In other words, if the model has to parse, it can do so

without any problem; otherwise, the stem will tend to contribute to the recognition process only through its contribution to the activation of the lemma.

The conclusion that parsing as such plays a minor role for this data set also emerged from the mathematical modeling study of Baayen, Dijkstra, and Schreuder (1997). Their model, however, was based on the assumption that there is no interaction between the representations feeding the direct access route and those feeding the parsing route. The present model, by contrast, implements the idea that all lexical representations are in competition and that the direct route and the parsing route are not independent. In fact, by allowing activation to spread from the plural to the base, and by allowing the evidence for the word status of a target in lexical decision to accumulate at the lexeme layer, there is much more synergy in this competition model than in the earlier mathematical model.

7.4.3 Noun Plurals in German

Clahsen, Eisenbeiss, and Sonnenstuhl-Henning (1997) and Sonnenstuhl-Henning and Huth (2002) report that in German, high-frequency plurals in *-er* are responded to faster in visual lexical decision than low-frequency plurals in *-er*, while matched high- and low-frequency plurals in *-s* are responded to equally fast. They attribute this difference to the linguistic status of these two plurals. The *-s* plural is argued to be the default suffix of German (Marcus et al. 1995), the only truly regular plural formative. According to these authors, plurals in *-s* are therefore not stored, which would explain why high- and low-frequency plurals in *-s* require the same processing time, a processing time that is determined only by the speed of the parsing route and the resting activation levels of the representations on which it operates. Plurals in *-er*, by contrast, are described as irregular. These forms must be stored, and the observed frequency effects for high- and low-frequency plurals reflect this.

In the light of their theory, it is not surprising that Clahsen, Eisenbeiss, and Sonnenstuhl-Henning (1997) suggest that Dutch plurals in *-en* might be irregular rather than regular (Clahsen, Eisenbeiss, and Sonnenstuhl-Henning 1997). This hypothesis, which does not make sense from a linguistic point of view (see Baayen et al. 2002), is inescapable if one accepts the dual-route model. However, this model faces many problems.

Burzio (in press a), for instance, points out that many morphologically regular past tense forms are phonologically irregular, while many morphologically irregular past tense forms are phonologically regular, an

inverse correlation between morphological and phonological regularity. In a dual-route model, morphological rules should be able to coexist with phonological rules, leading to a positive correlation between morphological and phonological regularities, contrary to fact.

Behrens (2001) presents a detailed examination of the German plural system showing that the *-s* plural does not fulfill the criteria for instantiating a symbolic rule, a conclusion supported by her acquisition data based on a very large developmental corpus.

Other problems for the dual-route model are pointed out by Ramscar (2002) and Hare, Ford, and Marslen-Wilson (2001). Hare, Ford, and Marslen-Wilson report frequency effects for regular past tense forms in English that have unrelated homophones (e.g., *allowed* and *aloud*). Particularly interesting is the study by Ramscar, who shows that past tense inflection in English is driven by semantic and phonological similarity, instead of by encapsulated symbolic rules that would be sensitive only to the phonological properties of the stem.

If the separation of rule and rote in the dual-route model is incorrect (see also Bybee 2001), the question arises why there seems to be no frequency effect for the German *-s* plural. In what follows, I will show that a simulation study with MATCHECK sheds new light on this issue.

Serving as our point of departure is the data set from experiment 1 of the study by Sonnenstuhl-Henning and Huth (2002). Table 7.8 summarizes the design of this experiment: six affix classes (plurals in *-s*, plurals in *-er*, and four sets of plurals in *-n*/*-en*, grouped by gender and the presence or absence of a final schwa), each of which is crossed with plural frequency (high vs. low) while matching for stem frequency. Table 7.9 summarizes the pattern of results obtained by means of a series of t-tests on the item means, and the right panel of figure 7.8 displays the results graphically. The high- and low-frequency plurals in *-s* elicited response latencies that did not differ significantly in the mean. The same holds for the plurals in *-en* of nonfeminine schwa-final nouns (labeled (4) in the tables and figure). The authors present these results as evidence that plurals in *-s* are not stored in the German mental lexicon. A question that remains unanswered in this study is why the supposed default plural suffix, the rule-governed formative par excellence, is the one to be processed most slowly of all plural classes in the study.

To study this data set with MATCHECK, a lexicon was constructed with the 120 singulars and their corresponding plurals that were used in this experiment. Some 2,100 randomly selected words from the CELEX

Table 7.8
German plurals cross-tabulated by class and frequency. Response latencies (RT) and frequencies of stem and plural as tabulated in Sonnenstuhl-Henning and Huth 2002.

Suffix	Class	Fre-quency	*F*(stem)	*F*(pl)	RT	Model time
-er	(1)	low	112	13	556	11.3
-er	(1)	high	112	43	510	9.3
-s	(2)	low	105	11	591	15.5
-s	(2)	high	97	50	601	15.2
-en	(3) [+fem, +e]	low	107	22	568	12.1
-en	(3) [+fem, +e]	high	111	59	518	10.6
-en	(4) [−fem, +e]	low	178	100	571	14.2
-en	(4) [−fem, +e]	high	209	170	544	13.0
-en	(5) [+fem, −e]	low	115	11	588	15.7
-en	(5) [+fem, −e]	high	116	66	548	12.8
-en	(6) [−fem, −e]	low	110	17	594	15.1
-en	(6) [−fem, −e]	high	112	84	535	12.7

Table 7.9
T-tests by affix. The t-tests for the observed data are as reported in Sonnenstuhl-Henning and Huth 2002; the t-tests for MATCHECK are Welch two-sample t-tests based on the model times.

		Observed			Expected		
Suffix	Class	*t*	df	*p*	*t*	df	*p*
-s	(1)	0.31	9	.760	0.53	16.59	.601
-er	(2)	4.92	9	.001	2.35	17.93	.030
-en	(3)	6.44	9	$<.001$	2.95	16.70	.009
-en	(4)	1.51	9	.166	1.19	17.00	.248
-en	(5)	4.39	9	.002	2.80	17.06	.012
-en	(6)	4.19	9	.001	2.82	12.32	.015

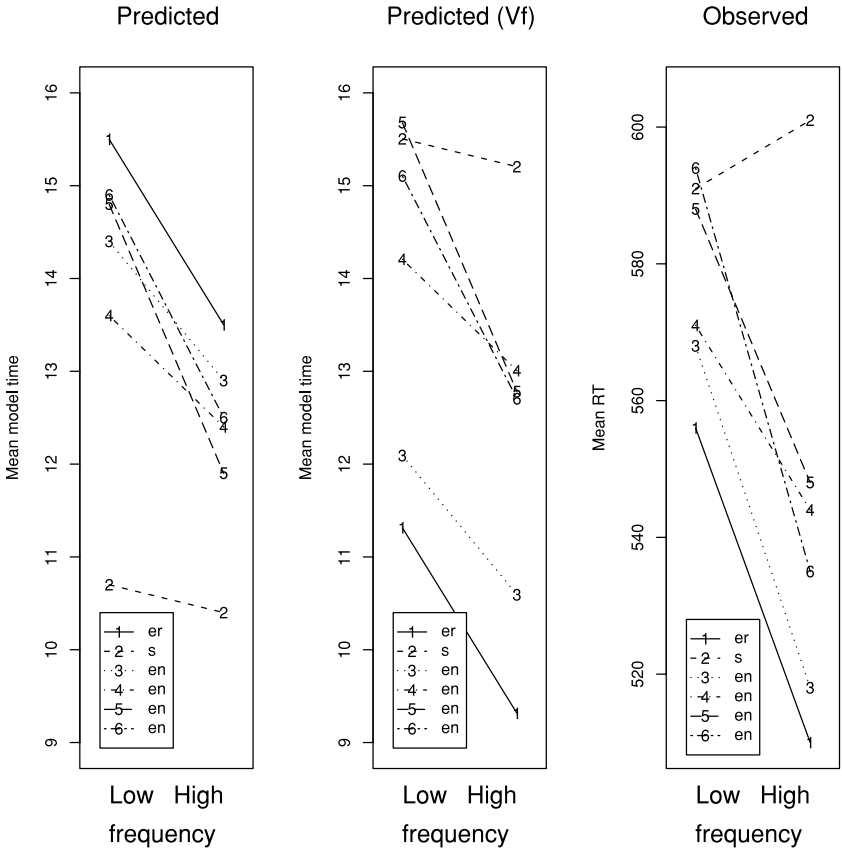

Figure 7.8
Predicted mean model times (*left*), predicted model times adjusted for the effect of Family Size (*center*), and observed mean reaction times (*right*) for German noun plurals in *-en* cross-classified by Frequency (low vs. high) and Affix Class (1: *-er*; 2: *-s*; 3: *-n* for feminine nouns ending in *e*; 4: *-n* for nonfeminine nouns ending in *e*; 5: *-en* for feminine nouns; 6: *-en* for nonfeminine nouns).

lexical database were added to the model's lexicon, including representations for the plural suffixes *-en*, *-n*, *-er*, and *-s* as well as for other inflectional suffixes such as *-d* and *-e*. All lexical entries received a frequency based on the frequency counts in the German section of the CELEX lexical database. Because these counts are based on a small corpus of only 6 million words, they were multiplied by 7 to make them similar in scale to the Dutch frequency counts, which are derived from a corpus of 42 million words. Suffixes received frequencies equal to the summed frequency of all the words in which they occur as a constituent. The singulars of the target plural forms were assigned the summed frequency of their inflectional variants, the lemma frequency of CELEX on which the high- and low-frequency sets of plurals in Sonnenstuhl-Henning and Huth's (2002) study were matched, just as in the simulation study of Dutch plurals in the preceding section. Again, a dummy lexeme was added with a frequency equal to the summed frequencies of the words not explicitly represented in the model's lexicon. Parameter values that as a first step yield a good fit to the experimental data are $\delta = 0.28$, $\theta = 0.20$, $\rho = 3$, $\alpha = 0.7$, $\zeta = 0.9$, and $\lambda = 0.5$. This fit is shown in the left panel of figure 7.8. As shown by the Welch two-sample t-tests listed in table 7.9, those frequency contrasts that were found to be statistically significant in the experiment are also significant in the model, and those contrasts that are not significant in the experiment are likewise not significant in the model.

What is disturbing about these results is that the ordering in model time seems to be wrong. Compare the left and right panels of figure 7.8. According to the model, the plurals in *-s* should be processed more quickly than any other kind of plural, but in fact they elicit the longest response latencies. Conversely, the model predicts the longest response latencies for the plurals in *-er*, even though in fact they were responded to very quickly.

Why does MATCHECK produce this reversed pattern? A possible answer can be found by considering the average family size of the six plural classes. The family size of an inflected simplex noun is the type count of the derived words and compounds in which that noun occurs as a constituent. Various studies have shown that, other things being equal, words with a large morphological family are responded to more quickly than words with a small morphological family (Schreuder and Baayen 1997; Bertram, Baayen, and Schreuder 2000; de Jong, Schreuder, and Baayen 2000; de Jong et al. 2002). The family size effect is semantic in nature and probably arises because of activation spreading in the mental

Table 7.10
Mean family size of the six German plural classes in Sonnenstuhl-Henning and Huth's (2002) study

Plural	Class	Family size
-er	(1)	12.5
-s	(2)	1.7
-en	(3)	8.8
-en	(4)	3.1
-en	(5)	2.6
-en	(6)	3.9

lexicon to morphologically related words. Interestingly, table 7.10 shows that the nouns with the *-er* plural have the largest family size in this data set, while the nouns with the *-s* plural have the smallest family size. Thus, plurals in *-s* might elicit long response latencies because of their small morphological families. Conversely, the plurals in *-er* might be responded to quickly thanks to their large families.

It is possible to test this hypothesis by asking ourselves whether we can systematically reorder the lines in the left panel of figure 7.8 on the basis of the family counts listed in table 7.10 such that a pattern approaching the one in the right panel is obtained. It turns out that this is not possible using the counts of family members for the individual words, probably because these counts (based on a corpus of only 6 million words) introduce too much noise compared to the model times of MATCHECK. A mapping is possible, however, on the basis of the class means, using the transformation $h(t_{ij})$ for the j-th plural in the i-th plural class with family size V_i and model time t_{ij}. Let $x_i = \log(V_i) - 1.2$, the distance of log family size from a baseline log family size of 1.2. The further the family size of a plural class is from this baseline, the greater the distance y_i that it will shift:

$$y_i = 1.7 * e^{|x_i|} * s(x_i), \tag{18}$$

with $s(x) = 1$ if $x > 0$ and $s(x) = -1$ if $x < 0$. Adjustment with the mean of the shifts y_i leads to the transformation

$$h(t_{ij}) = t_{ij} - y_i - \frac{\sum_k^n y_k}{n}. \tag{19}$$

Application of (19) results in the pattern shown in the center panel of figure 7.8, a reasonable approximation of the actually observed pattern

represented in the right panel. Although a more principled way of integrating MATCHECK with subsequent semantic processes is clearly called for, the present result suggests that differences in morphological family size may indeed be responsible for the difference in ordering between the left and right panels of figure 7.8. The family size effect observed here underlines that the plurals in *-er* are tightly integrated into the network of morphological relations of German and that, by contrast, the plurals in *-s* are rather superficially integrated and relatively marginal in the language. This is not what one would expect for a default suffix in a dual-route model, especially if default status is meant to imply prototypical rule-based processing rather than marginal rule-based processing applicable mainly to recent loans, recent abbreviations, interjections, and other normally uninflected words.

These modeling results receive further support from a simulation of experiment 4 of the study by Clahsen, Eisenbeiss, and Sonnenstuhl-Henning (1997), who contrasted plurals in *-s* with plurals in *-er*. As in Sonnenstuhl-Henning and Huth's (2002) study, a significant frequency effect was observed only for the plurals in *-er*. Interestingly, application of MATCHECK to this data set with exactly the same parameter values leads to the same pattern of results, with a significant difference in model time for the plurals in *-er* ($t(15.45) = 4.13$, $p = .0008$) but not for the plurals in *-s* ($t(15.52) = 1.60$, $p = .1290$).

We are left with one question. Why is it that MATCHECK does not produce a frequency effect for plurals in *-s*? Consider figure 7.9, which plots the timecourse of activation for the German nouns *Hypotheken* 'mortgages' (left panel) and *Taxis* 'taxis' (right panel). The pattern in the left panel is typical for the plurals in *-en* and for those plurals in *-er* for which no vowel alternation occurs. Crucially, the base and the plural form become active one after the other, indicating that there is little competition between them. In the case of plurals in *-s*, however, the base and the plural become active at more or less the same time—they are in strong competition, masking the effect of the frequency of the plural. The reason for this strong competition is the small difference in length between the singular forms and their corresponding plurals in *-s*. Recall that the activation/decay rate of a word as defined in equations (14)–(16) depends on its length in relation to the length of the target word. The more similar a word is in length to the target, the faster it will become active. This property contributes to the segmentation accuracy of MATCHECK as reported by Baayen and Schreuder (2000), and it is

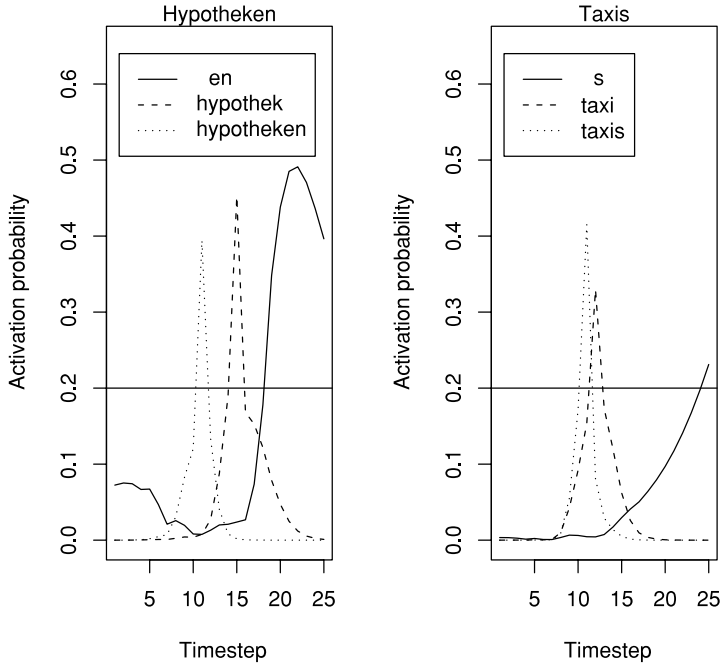

Figure 7.9
Patterns of activation probabilities over time in MATCHECK for the German plurals *Hypotheken* 'mortgages' (*left*) and *Taxis* 'taxis' (*right*), with a threshold at .2

responsible for the absence of a frequency effect for *-s* plurals in the present simulation study. Note that this property of MATCHECK is of the kind typically found in dynamic systems in general, in that it allows a tiny difference in the initial conditions (here the length of the base) to lead to quite different outcomes (here the presence or absence of a frequency effect).

From a methodological point of view, it is important to realize that null effects, in this case the absence of a frequency effect for German plurals in *-s*, should be interpreted with caution. Clahsen and his coworkers infer from this null effect that German *-s* plurals do not develop representations of their own. We have seen that this null effect may also arise as a consequence of the differences between the forms of the *-s* and *-en* suffixes and the kinds of words that they attach to (see Laudanna and Burani 1995 and Bertram, Schreuder, and Baayen 2000, for the processing consequences of affix-specific differences). Another

argument based on a null effect is made by Frost, Forster, and Deutsch (1997), who claim that connectionist models would not be able to model the presence and absence of semantic transparency effects in priming experiments in English and Hebrew, respectively. As shown by Plaut and Gonnerman (2000), the null effect of transparency in Hebrew emerges in connectionist simulation studies as a consequence of the difference in morphological structure between English and Hebrew.

7.4.4 Discussion

This section addressed the role of probability in morphological comprehension. Techniques developed by computational linguists make profitable use of co-occurrence probabilities to select the most probable segmentation from the set of possible segmentations. It is clear from the computational approach that knowledge of co-occurrence probabilities is indispensable for accurate and sensible parsing.

From this perspective, the hypothesis defended by Pinker and Clahsen and their coworkers that co-occurrence knowledge is restricted to irregular complex words is counterproductive (see also Bybee 2001). Bloomfieldian economy of description cannot be mapped so simply onto human language processing. Much more productive are, by contrast, those psycholinguistic studies that have addressed in great detail how form-based similarity affects the segmentation process. These studies, however, seem to implicitly assume that the segmentation problem can be solved without taking co-occurrence probabilities into account.

MATCHECK is a model in which probabilities develop dynamically over time on the basis of both frequency of (co-)occurrence and form similarity. I have shown that this model provides reasonable fits to experimental data sets and that it provides an explanation of why frequency effects for regular plurals may be absent even though these plurals are stored in the mental lexicon. The mechanisms used by MATCHECK to weight the role of frequency and similarity are crude and in need of considerable refinement. Nevertheless, I think that the model provides a useful heuristic for understanding how the human brain might use probability in morphological comprehension to its advantage.

7.5 Concluding Remarks

In this chapter, I have presented some examples of how concepts from probability theory and statistics can be used to come to grips with the

graded nature of many morphological data. I first showed that the graded phenomenon of morphological productivity can be formalized as a probability—one that is itself grounded, at least in part, in junctural phonotactic probabilities and parsing probabilities. Following this, I discussed examples of morphological regularities that are intrinsically probabilistic in nature, outlining how simple spreading activation architectures (symbolic connectionist networks) might capture the role of probability while avoiding complex statistical calculations. Finally, I have shown that part of the functionality of the storage of regular complex forms in the mental lexicon may reside in contributing to resolving parsing ambiguities in comprehension.

Not surprisingly, I agree with researchers such as Seidenberg and Gonnerman (2000) and Plaut and Gonnerman (2000) that traditional, nonprobabilistic theories of morphology are inadequate in the sense that they cannot handle graded phenomena in an insightful way. These researchers seem to suggest, however, that the graded nature of morphology shows that the subsymbolic connectionist approach to language is the *only* way to go. This is where they and I part company. In this chapter, I have shown that computational models of analogy provide excellent analytical tools for understanding the role of probability in morphology.

Note

I am indebted to Kie Zuraw and Janet Pierrehumbert for their careful and constructive criticism; to Wolfgang Dressler, Mirjam Ernestus, Robert Schreuder, and Royal Skousen for their comments and suggestions for improvements; and to Stefanie Jannedy, Jennifer Hay, and Rens Bod for their wonderful editorial work. Errors in content and omissions remain my responsibility.

This page intentionally left blank

Chapter 8

Probabilistic Syntax

Christopher D. Manning

8.1 The Tradition of Categoricity and Prospects for Stochasticity

"Everyone knows that language is variable." This is the bald sentence with which Sapir (1921, 147) begins his chapter on language as a historical product. He goes on to emphasize how two speakers' usage is bound to differ "in choice of words, in sentence structure, in the relative frequency with which particular forms or combinations of words are used." I should add that much sociolinguistic and historical linguistic research has shown that the *same* speaker's usage is also variable (Labov 1966; Kroch 2001, 722). However, the tradition of most syntacticians has been to ignore this thing that everyone knows.

Human languages are the prototypical example of a symbolic system. From very early on, logics and logical reasoning were invented for handling natural language understanding. Logics and formal languages have a languagelike form that draws from and meshes well with natural languages. It is not immediately obvious where the continuous and quantitative aspects of syntax are. The dominant answer in syntactic theory has been "nowhere" (Chomsky 1969, 57; also 1956, 1957, etc.): "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term."

In the 1950s there were prospects for probabilistic methods taking hold in linguistics, in part owing to the influence of the new field of Information Theory (Shannon 1948).[1] Chomsky's influential remarks had the effect of killing off interest in probabilistic methods for syntax, just as for a long time McCarthy and Hayes (1969) discouraged exploration of probabilistic methods in artificial intelligence. Among his arguments were that (1) probabilistic models wrongly mix in world knowledge (*New York*

occurs more in text than *Dayton, Ohio*, but for no linguistic reason); (2) probabilistic models don't model grammaticality (neither *Colorless green ideas sleep furiously* nor *Furiously sleep ideas green colorless* has previously been uttered—and hence must be estimated to have probability zero, Chomsky wrongly assumes—but the former is grammatical while the latter is not); and (3) use of probabilities does not meet the goal of describing the mind-internal I-language as opposed to the observed-in-the-world E-language. This chapter is not meant to be a detailed critique of Chomsky's arguments—Abney (1996) provides a survey and a rebuttal, and Pereira (2000) has further useful discussion—but some of these concerns are still important to discuss. I argue in section 8.3.2 that in retrospect none of Chomsky's objections actually damn the probabilistic syntax enterprise.

Chambers (1995, 25–33) offers one of the few clear discussions of this "tradition of categoricity" in linguistics of which I am aware. He makes a distinction between standardly used linguistic units that are *discrete* and *qualitative* versus an alternative that allows units that are *continuous* and *quantitative*. He discusses how the idea that linguistics should keep to a categorical base precedes modern generative grammar. It was standard among American structuralists and is stated most firmly by Joos (1950, 701–702):

> Ordinary mathematical techniques fall mostly into two classes, the continuous (e.g., the infinitesimal calculus) and the discrete or discontinuous (e.g., finite group theory). Now it will turn out that the mathematics called "linguistics" belongs to the second class. It does not even make any compromise with continuity as statistics does, or infinite-group theory. Linguistics is a quantum mechanics in the most extreme sense. All continuities, all possibilities of infinitesimal gradation, are shoved outside of linguistics in one direction or the other.

Modern linguistics is often viewed as a cognitive science. Within cognitive science in general, it is increasingly understood that there is a central role for the use of probabilistic methods in modeling and understanding human cognition (for vision, concept learning, etc.—see, e.g., Kersten 1999; Tenenbaum 1999). Indeed, in many areas, probabilistic modeling has become so prevalent that Mumford (1999) has seen fit to declare the Dawning of an Age of Stochasticity. Human cognition has a probabilistic nature: we continually have to reason from incomplete and uncertain information about the world, and probabilities give us a well-founded tool for doing this.

Language understanding is a subcase of this. When someone says, "It's cold in here," in some circumstances I'm understanding correctly if I interpret that utterance as a request to close the window. Ambiguity and underspecification are ubiquitous in human language utterances, at all levels (lexical, syntactic, semantic, etc.), and how to resolve these ambiguities is a key communicative task for both human and computer natural language understanding. At the highest level, the probabilistic approach to natural language understanding is to view the task as trying to learn the probability distribution $P(meaning|utterance, context)$—a mapping from form to meaning conditioned by context. In recent years, such probabilistic approaches have become dominant within computational linguistics (Manning and Schütze 1999), and they are becoming increasingly used within psycholinguistics (Jurafsky, this volume; Baayen, this volume). Probabilistic methods have largely replaced earlier computational approaches because of the more powerful evidence combination and reasoning facilities they provide. This greater power comes from the fact that knowing how likely a module thinks a certain sense, structure, or interpretation is is much more useful and powerful information than just knowing whether it is deemed possible or impossible. Language acquisition is also a likely place for probabilistic reasoning: children are necessarily doing uncertain analytical reasoning over uncertain input. But is the same true for the core task of describing the syntax—the grammar—of a human language?

I will advocate that the answer is yes. However, so far, syntax has not seen the dawn. Probabilistic syntax (or, equivalently, stochastic syntax) has been a little-studied area, and there is not yet a large, coherent body of work applying sophisticated probabilistic models in syntax. Given the great promise of probabilistic techniques, the reader should see this as an invitation to get involved on the ground floor of an exciting new approach. I will attempt to explain why probabilistic models have great value and promise for syntactic theory, and to give a few small examples and connections to relevant literature. Motivation for noncategorical models in syntax comes from language acquisition, historical change, and typological and sociolinguistic variation, and I touch on those motivations briefly, but only briefly since they are dealt with further in other chapters. Throughout, the aim is to examine probabilistic models for explaining language structure, as opposed to simply using techniques like significance tests to buttress empirical results.

8.2 The Joys and Perils of Corpus Linguistics

8.2.1 On the Trail of a Neat Fact

Halfway through a long overseas plane flight, I settled down to read a novel (Russo 2001). However, I only made it to the third page before my linguist brain reacted to this sentence:

(1) By the time their son was born, though, Honus Whiting was beginning to understand and privately share his wife's opinion, as least as it pertained to Empire Falls.

Did you notice anything abnormal? It was the construction ***as*** *least as* that attracted my attention: for my categorical syntactician self, this construction is simply ungrammatical. I should add the rider "in my idiolect"; but in just this manner, despite what is said in introductory linguistics classes, modern linguistics has become highly prescriptive, or at least has come to use "normative edited texts"—a term Chambers (1995, 27) attributes to Labov.

For my corpus linguist self, this sentence is a piece of incontrovertible primary data. This status for data is importantly different from Chomsky's (1965, 4) stance that "observed use of language ... may provide evidence ... but surely cannot constitute the actual subject matter of linguistics, if this is to be a serious discipline." But there remains the question of how to interpret this datum, and here Chomsky's (1986) distinction between externally visible E-language and the internalized I-language of the mind still seems important. One possibility, which seemed the most probable one to a linguist friend sitting in the next seat, is that *as least as* is just a typo, or some similar result of editing. Or it might have been a "speech error" that somehow survived into print, but that the author would judge as ungrammatical if asked. Or it could be a matter of dialectal, sociolectal, or just idiolectal variation. In general one cannot make a highly confident decision between these and other choices. In this case one could conceivably have telephoned; but in general that option is not available, and at any rate sociolinguists have long observed that people's actual language use deviates from their reports of how they use language. This unclarity of attribution means that corpus linguistics necessarily has a statistical basis: one has to reason about the likelihood of different explanations based on both the frequency with which different forms occur and prior probabilistic models of language and other aspects of cognition and the world.

Back on land, the obvious way to address this question, within a scientific empirical linguistics, is to collect more evidence. A search of 1994 *New York Times* newswire yields no examples—is this just because they have better copy editors?—but then I find four examples including (2a–b) in 1995 *New York Times* newswire, and two more examples from 1996. It already appears less likely that this was just a typo. A search on the Web (which with every passing year becomes a better medium for empirical linguistic research) yields hundreds of further examples. There are ones (apparently) from U.S. East Coast speakers (2c), from Midwestern speakers (2d), and from West Coast speakers (2e). There are ones from college professors (2c) and from fishing boat owners (2f). I can't even regard this as a U.S. aberration (I'm Australian): I find examples from Australia (2g) and South Africa (2h). Finally, I find examples with *at least as* in a neighboring sentence (2i–j), showing intraspeaker variation. While much less common than *at least as* (perhaps about 175 times less common, based on this small study), *as least as* seems to have robust support and perhaps indicates a development within the language (there were several examples in discussions of the 2000 U.S. election, but there was also one citation from 1972—more research would be needed to establish this). Unlike the initial example from Russo, many—but not all—of these additional examples use *as least as* in conjunction with an *as* Adj *as* construction, which perhaps provides a pathway for the development of this apparently new form.[2]

(2) a. Indeed, the will and the means to follow through are *as least as* important *as* the initial commitment to deficit reduction.
b. Steven P. Jobs has reemerged as a high-technology captain of industry, *as least as* far *as* the stock market is concerned.
c. Alternatively, *y* is preferred to *x* if, in state *x*, it is not possible to carry out hypothetical lump-sum redistribution so that everyone could be made *as least as* well off *as* in *y*.
d. There is *as least as* much investment capital available in the Midwest *as* there is on either Coast.
e. The strip shall be of a material that is *as least as* slip-resistant *as* the other treads of the stair.
f. As many of you know he had his boat built at the same time as mine and it's *as least as* well maintained and equipped.
g. There is a history of using numbers to write music that dates *as least as* far back to Pythagoras.

h. He added: "The legislation is *as least as* strict *as* the world's strictest, if not the strictest."
i. Black voters also turned out *at least as* well *as* they did in 1996, if not better in some regions, including the South, according to exit polls. Gore was doing *as least as* well among black voters *as* President Clinton did that year.
j. Second, if the required disclosures are made by on-screen notice, the disclosure of the vendor's legal name and address must appear on one of several specified screens on the vendor's electronic site and must be *at least as* legible and set in a font *as least as* large *as* the text of the offer itself.

Thus, we have discovered a neat fact about English lexicogrammar—previously unremarked on as far as I am aware—and we have at least some suspicions about its origin and range of use, which we could hope to confirm with further corpus-based analysis. And this has been quite fun to do: there was the chance discovery of a "smoking gun" followed by "text mining" to discover further instances. However, a few observations are immediately in order.[3]

8.2.2 The Important Lessons

First, this example was easy to investigate because the phenomenon is rooted in particular lexical items. It's easy to search text corpora for *as least as* constructions; it is far harder to search corpora for something like participial relative clauses or locative inversion constructions. Such technological limitations have meant that corpus linguistic research has been largely limited to phenomena that can be accessed via searches on particular words. The average corpus linguist's main research tool remains the word-concordancing program, which shows a searched-for keyword in context (perhaps with morphological stemming, sorting options, etc.).[4] However, a (theoretical) syntactician is usually interested in more abstract structural properties that cannot be investigated easily in this way. Fortunately for such research, recent intensive work in statistical natural language processing (statistical NLP; Manning and Schütze 1999) has led to the availability of both many more richly annotated text corpora that can support such deeper investigations, and many more tools capable of being used with good accuracy over unannotated text in order to recover deep syntactic relationships. The prospects for applying these corpora and tools to probabilistic syntactic analysis are bright, but it remains fair

to say that these tools have not yet made the transition to the Ordinary Working Linguist without considerable computer skills.[5]

Second, one needs a large corpus to be able to do interesting exploration of most syntactic questions. Many syntactic phenomena, especially those commonly of interest to theoretical syntacticians, are incredibly rare in text corpora. For instance, in the *New York Times* newswire corpus mentioned above, I searched through about 230 million words of text to find a paltry six examples of *as least as*. This is one aspect of the problem that Sinclair (1997, 27) sums up by saying, "The linguistics of the twentieth century has been the linguistics of scarcity of evidence." All approaches to linguistics have dealt with this problem in one way or another. Generative approaches resort to inventing the primary data, on the basis of intuitions. I will not discuss the possible limitations of such an approach in detail here. Suffice it to say that even Chomsky (1986, 63) admits that "the facts are often quite obscure."[6] Traditional field linguists have been constrained by quite small collections of texts and conversations. While the major syntactic features will be clear, there is generally quite insufficient data for the kind of subtle issues that are the mainstay of theoretical syntax. Similarly for sociolinguists: the data they collect by traditional in-person techniques are rich in social and contextual information, but poor in quantity. It is probably for this reason that the vast majority of sociolinguistic work has dealt with phonetic realization (for which the data are dense), and the small amount of syntactic work has been mainly on phenomena such as copula deletion and use of modals where again the density of the use of function words makes analysis from a small corpus possible. In contrast, corpus linguists have tended to work with large amounts of data in broad pseudogenres like "newspaper text," collapsing the writing of many people from many places. There is no easy answer to the problem of getting sufficient data of just the right type: language changes across time, space, social class, method of elicitation, and so on. There is no way that we can collect a huge quantity of data (or at least a collection dense in the phenomenon of current interest) unless we are temporarily prepared to ride roughshod over at least one of these dimensions of variation. To deal with rare syntactic phenomena, we must either work with intuitions or else be prepared to aggregate over speakers and time periods. Even then, the *sparsity* of linguistic data is nearly always a major technical challenge in probabilistic approaches to linguistics or NLP.

Finally—and most seriously—whatever their interest, the above observations on *as least as* unquestionably fit into the category of activities that Chomsky (1979, 57) long ago derided as butterfly collecting: "You can also collect butterflies and make many observations. If you like butterflies, that's fine; but such work must not be confounded with research, which is concerned to discover explanatory principles of some depth and fails if it does not do so." A central question is whether probabilistic models can be used for linguistic explanation and insight in this manner. I think this is a serious concern. To go out on a limb for a moment, let me state my view: generative grammar has produced many explanatory hypotheses of considerable depth, but is increasingly failing because its hypotheses are disconnected from verifiable linguistic data. Issues of frequency of usage are by design made external to matters of syntax, and as a result categorical judgments are overused where not appropriate, while a lack of concern for observational adequacy has meant that successive versions have tended to treat a shrinking subset of data increasingly removed from real usage. On the other side, corpus linguistics (McEnery and Wilson 2001)—or "usage-based models of grammar" (Barlow and Kemmer 2000)—has all the right rhetoric about being an objective, falsifiable, empirical science interested in the totality of language use, but is failing by largely restricting itself to surface facts of language, rather than utilizing sophisticated formal models of grammar, which make extensive use of *hidden structure* (things like phrase structure trees and other abstract representational levels). This reflects an old dichotomy: one sees it clearly in the 1960s *Handbook of Mathematical Psychology* (Luce, Bush, and Galanter 1963a,b), where in some chapters probabilistic finite state models (markov chains) are being actively used to record surface-visible stimulus-response behavior, whereas in another chapter Chomsky is arguing for richer tools (then, transformational grammars) for attacking deeper analytical problems. The aim of the current chapter is to indicate a path beyond this impasse: to show how probabilistic models can be combined with sophisticated linguistic theories for the purpose of syntactic explanation.

Formal linguistics has traditionally equated structuredness with homogeneity (Weinreich, Labov, and Herzog 1968, 101), and it has tried too hard to maintain categoricity by such devices as appeal to an idealized speaker/hearer. I would join Weinreich, Labov, and Herzog (1968, 99) in hoping that "a model of language which accommodates the facts of vari-

able usage ... leads to more adequate descriptions of linguistic competence.'' The motivation for probabilistic models in syntax comes from two sides:

- Categorical linguistic theories claim too much. They place a hard categorical boundary of grammaticality where really there is a fuzzy edge, determined by many conflicting constraints and issues of conventionality versus human creativity. This is illustrated with an example in section 8.3.1.
- Categorical linguistic theories explain too little. They say nothing at all about the soft constraints that explain how people choose to say things (or how they choose to understand them). The virtues of probabilistic models for this task are developed in section 8.5.

The first problem is a foundational worry, but the second is more interesting in showing the possibility for probabilistic models to afford increased linguistic explanation.

8.3 Probabilistic Syntactic Models

As an example of the noncategorical nature of syntax, and the idealization that occurs in most of the syntactic literature, I will look here at verbal clausal subcategorization frames. I will first look at problems with categorical accounts, then explore how one might build a probabilistic model of subcategorization, and finally consider some general issues in the use of probabilistic approaches within syntax.

8.3.1 Verbal Subcategorization

It is well known that different verbs occur with different patterns of arguments that are conventionally described in syntax by subcategorization frames (Chomsky 1965; Radford 1988). For example, some verbs must take objects while others do not:[7]

(3) a. Kim devoured the meal.
    b. *Kim devoured.
    c. *Dana's fist quivered Kim's lip.
    d. Kim's lip quivered.

Other verbs take various forms of sentential complements.

A central notion of all current formal theories of grammar (including Government-Binding (Chomsky 1981a), the Minimalist Program

(Chomsky 1995), Lexical-Functional Grammar (Bresnan 2001), Head-Driven Phrase Structure Grammar (Pollard and Sag 1994), Categorial Grammar (Morrill 1994), Tree-Adjoining Grammar (Joshi and Schabes 1997), and other frameworks) is a distinction between arguments and adjuncts combining with a head. Arguments are taken to be syntactically specified and required by the head, via some kind of subcategorization frame or argument list, whereas adjuncts (of time, place, etc.) can freely modify a head, subject only to semantic compatibility constraints.

This conception of the argument/adjunct distinction is the best one can do in the categorical 0/1 world of traditional formal grammars: things have to be either selected (as arguments) or not. If they are not, they are freely licensed as adjuncts, which in theory should be able to appear with any head and to be iterated any number of times, subject only to semantic compatibility. And there is certainly some basis for the distinction: a rich literature (reviewed in Radford 1988; Pollard and Sag 1987; Schütze 1995) has demonstrated argument/adjunct asymmetries in many syntactic domains, such as phrasal ordering and extraction.

However, categorical models of selection have always been problematic. The general problem with this kind of model was noticed early on by Sapir (1921, 38), who noted that "all grammars leak." In context, language is used more flexibly than such a model suggests. Even Chomsky in early work (1955, 131) notes that "an adequate linguistic theory will have to recognize degrees of grammaticalness." Many subcategorization distinctions presented in the linguistics literature as categorical are actually counterexemplified in studies of large corpora of written language use. For example, going against (3c), Atkins and Levin (1995) find examples of *quiver* used transitively, such as *The bird sat, quivering its wings*. In this section, I will look at some examples of clausal complements, focusing particularly on what realizations of arguments are licensed, and then touch on the argument/adjunct distinction again at the end.

Some verbs take various kinds of sentential complements, and this can be described via subcategorization frames:

(4) *regard*: ____ NP[acc] *as* {NP, AdjP}
consider: ____ NP[acc] {AdjP, NP, VP[inf]}
think: {____ CP[that], ____ NP[acc] NP}

In (5), I reproduce the grammaticality judgments for subcategorizations given by Pollard and Sag (1994, 105–108) within the context of an

argument that verbs must be able to select the category of their complements.[8] I begin with *consider*, which Pollard and Sag say appears with a noun phrase object followed by various kinds of predicative complements (nouns, adjectives, clauses, etc.), but not with *as* complements:

(5) a. We consider Kim to be an acceptable candidate.
b. We consider Kim an acceptable candidate.
c. We consider Kim quite acceptable.
d. We consider Kim among the most acceptable candidates.
e. *We consider Kim as an acceptable candidate.
f. *We consider Kim as quite acceptable.
g. *We consider Kim as among the most acceptable candidates.
h. ?*We consider Kim as being among the most acceptable candidates.

However, this lack of *as* complements is counterexemplified by various examples from the *New York Times*:[9]

(6) a. The boys consider her as family and she participates in everything we do.
b. Greenspan said, ''I don't consider it as something that gives me great concern.''
c. ''We consider that as part of the job,'' Keep said.
d. Although the Raiders missed the playoffs for the second time in the past three seasons, he said he considers them as having championship potential.
e. Culturally, the Croats consider themselves as belonging to the ''civilized'' West, . . .

If this were an isolated incident (counterexemplifying (5e) and (5h) in particular), then we would merely have collected one more butterfly, and we would conclude that Pollard and Sag got that one particular fact wrong. But the problem is much more endemic than that. The subcategorization facts that they provide can in general be counterexemplified. Since this is an important point in the argument for probability distributions in syntax, let me give a few more examples.

According to Pollard and Sag—and generally accepted linguistic wisdom—*regard* is the opposite of *consider* in disallowing infinitival VP complements, but allowing *as* complements:

(7) a. *We regard Kim to be an acceptable candidate.
b. We regard Kim as an acceptable candidate.

But again we find examples in the *New York Times* where this time *regard* appears with an infinitival VP complement:

(8) a. As 70 to 80 percent of the cost of blood tests, like prescriptions, is paid for by the state, neither physicians nor patients regard expense to be a consideration.
b. Conservatives argue that the Bible regards homosexuality to be a sin.

Pollard and Sag describe *turn out* as allowing an adjectival phrase complement but not a present participle VP complement:

(9) a. Kim turned out political.
b. *Kim turned out doing all the work.

But again we find counterexamples in the *New York Times*:

(10) But it turned out having a greater impact than any of us dreamed.

Similarly, *end up* is predicted to disallow a perfect participle VP complement,

(11) a. Kim ended up political.
b. *Kim ended up sent more and more leaflets.

but an example of the sort predicted to be ungrammatical again appears in the *New York Times*:

(12) On the big night, Horatio ended up flattened on the ground like a fried egg with the yolk broken.

What is going on here? Pollard and Sag's judgments seem reasonable when looking at the somewhat stilted "linguists' sentences." But with richer content and context, the *New York Times* examples sound (in my opinion) in the range between quite good and perfect. None of them would make me choke on my morning granola. They in all likelihood made it past a copy editor. While one must consider the possibility of errors or regional or social variation, I think it is fair to say that such explanations are not terribly plausible here.

The important question is how we should solve this problem. Within a categorical linguistics, there is no choice but to say that the previous model was overly restrictive and that these other subcategorization frames should also be admitted for the verbs in question. But if we do that, we lose a lot. For we are totally failing to capture the fact that the subcategorization frames that Pollard and Sag do not recognize are

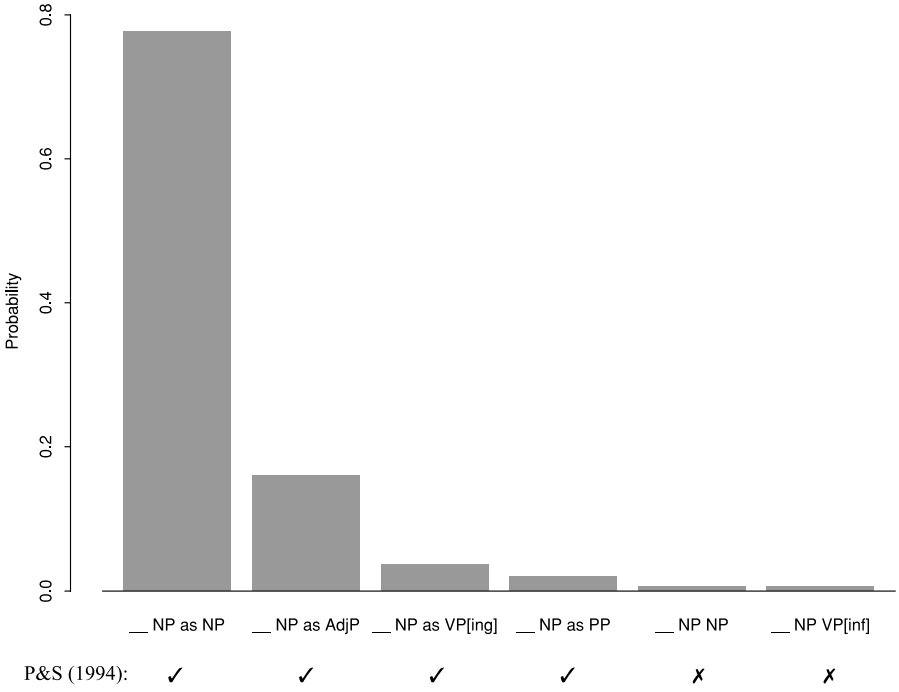

Figure 8.1
Estimated probability mass function (pmf) for subcategorizations of *regard* based on a 300-token *New York Times* corpus. The graph is annotated with grammaticality judgments of Pollard and Sag (1994).

extremely rare, whereas the ones they give encompass the common subcategorization frames of the verbs in question. We can get a much better picture of what is going on by estimating a probability mass function (pmf) over the subcategorization patterns for the verbs in question. A pmf over a discrete random variable (a set of disjoint choices) gives the probability of each one. Here, we will estimate a pmf for the verb *regard*. The subcategorization pmf in figure 8.1 was estimated from 300 tokens of *regard* from the *New York Times*, by simply counting how often each subcategorization occurred in this sample and then graphing these frequencies divided by the number of tokens—so as to give the maximum likelihood point estimate for each frame.[10]

The graph is annotated with Pollard and Sag's grammaticality judgments. We can note two things. First, the line of grammaticality has

apparently been drawn at a fairly arbitrary point of low frequency. Roughly, things that occur at least 1 time in 100 are regarded as grammatical, while things that occur less commonly are regarded as ungrammatical. If the cutoff had been set as 1 time in 10, then only the *as* NP and *as* AdjP complementation patterns would be "grammatical"; if the cutoff had been set at 1 time in 1,000, then they would all be grammatical.[11] Second, note how much important information is lost by simply providing categorical information. In a categorical description, there is no record whatsoever of the fact that over three-quarters of all predicative complements of *regard* are in the form of an *as* NP complement. A probability distribution captures such information in a straightforward manner.

Looking more broadly, careful consideration of the assumed categorical distinction between arguments and adjuncts further argues for new foundations for syntax. There are some very clear arguments (normally, subjects and objects), and some very clear adjuncts (of time and "outer" location), but also a lot of stuff in the middle. Things in this middle ground are often classified back and forth as arguments or adjuncts depending on the theoretical needs and convenience of the author (Babby 1980; Fowler 1987; Maling 1989, 1993; Li 1990; Wechsler and Lee 1996; Przepiórkowski 1999a,b). Additionally, various in-between categories are postulated, such as argument-adjuncts (Grimshaw 1990) and pseudo-complements (Verspoor 1997).

In a probabilistic approach, in contrast, it is not necessary to categorically divide verbal dependents into subcategorized arguments and freely occurring adjuncts. Rather, we can put a probability function over the kinds of dependents to expect with a verb or class, conditioned on various features. This is especially useful in difficult in-between cases. For example, in the *Oxford Advanced Learners Dictionary* (Hornby 1989), the verb *retire* is subcategorized as a simple intransitive and transitive verb, and as an intransitive taking a PP[*from*] or PP[*to*] argument.[12] While prepositional phrases headed by *to* or *from* are common with *retire* (13a–b)—and are arguably arguments by traditional criteria—this does not exhaust the list of putative arguments of *retire*. While *in* most often occurs with *retire* to specify a time point (a canonical adjunct PP), it sometimes expresses a destination (13c), and it seems that these examples demand the same treatment as examples with PP[*to*]. Similarly, uses of a PP[*as*] to express the position that one is retiring from (13d) seem at least as

selected by the verb as those expressed by a PP[*to*] or PP[*from*]. What about the usage of *retire on* (13e), where the PP[*on*] expresses the source of monetary support?[13]

(13) a. Mr. Riley plans to retire to the $1.5 million ranch he is building in Cody, Wyo.
b. Mr. Frey, 64 years old, remains chairman but plans to retire from that post in May.
c. To all those wishing to retire in Mexico (international section, March 10 and 11), let me offer three suggestions: ...
d. Donald W. Tanselle, 62 years old, will retire as vice chairman of this banking concern, effective Jan. 31.
e. A worker contributing 10% of his earnings to an investment fund for 40 years will be able to retire on a pension equal to two thirds of his salary.

Rather than maintaining a categorical argument/adjunct distinction and having to make in/out decisions about such cases, we might instead try to represent subcategorization information as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability. For instance (sticking for the moment to active uses, and assuming that controlled and imperative subjects are present, etc.), we might estimate that[14]

$$\begin{aligned} P(\text{NP[SUBJ]} \mid V = \textit{retire}) &= 1.0 \\ P(\text{NP[OBJ]} \mid V = \textit{retire}) &= .52 \\ P(\text{PP[}\textit{from}\text{]} \mid V = \textit{retire}) &= .05 \\ P(\text{PP[}\textit{as}\text{]} \mid V = \textit{retire}) &= .06 \\ &\vdots \end{aligned}$$

By writing things like this, I am implicitly assuming independence between arguments (see Bod, this volume): the chance of getting a PP[*as*] is independent of whether a PP[*from*] is present or not. This is presumably not true: intuition suggests that having either a PP[*as*] or a PP[*from*] makes having the other less likely. We could choose instead to provide joint estimates of a complete subcategorization frame:

$$\begin{aligned} P(\text{NP[SUBJ]} \_\_\_\_ \mid V = \textit{retire}) &= .25 \\ P(\text{NP[SUBJ]} \_\_\_\_ \text{ NP[OBJ]} \mid V = \textit{retire}) &= .5 \\ P(\text{NP[SUBJ]} \_\_\_\_ \text{ PP[}\textit{from}\text{]} \mid V = \textit{retire}) &= .04 \\ P(\text{NP[SUBJ]} \_\_\_\_ \text{ PP[}\textit{from}\text{] PP[}\textit{after}\text{]} \mid V = \textit{retire}) &= .003 \\ &\vdots \end{aligned}$$

In this latter case, the sums of the probabilities of all frames would add to one (in the former, it is the probabilities of whether a verb will have a certain argument or not that add to one).

Regardless of details of the particular modeling choices, such a change of approach reshapes what questions can be asked and what results are achievable in terms of language description, learnability, production, and comprehension. Moreover, the probabilistic approach opens up possibilities for renewed productive interplay between formal syntax and various areas of applied linguistics, including computational linguistics. A language teacher is likely to be interested in how a certain semantic type of argument is most commonly expressed, or which frames a verb is most commonly used with. While they incorporate a number of further notions, modern statistical NLP parsers (Collins 1997; Charniak 1997b) incorporate generative probabilistic models of surface subcategorization, much like the above kind of lexicalized dependency information. Rather than using simple probabilistic context-free grammar rules like VP → V NP PP of the sort discussed by Bod (this volume), these generative models use lexicalized argument frames by modeling the probability that a VP is headed by a certain verb, and then the probability of certain arguments surrounding that verb:[15]

$$P(\text{VP} \rightarrow \text{V}[\textit{retire}]\ \text{PP}[\textit{from}]) = P(\text{VP}[\textit{retire}]|\text{VP}) \times P(\text{VP}[\textit{retire}] \rightarrow \text{V NP PP}|\text{VP}[\textit{retire}]) \times P(\text{PP}[\textit{from}]|\text{PP}, \text{VP}[\textit{retire}]).$$

Perhaps most importantly, such models combine formal linguistic theories and quantitative data about language use, in a scientifically precise way.

However, dealing simply with surface subcategorization is well known to be problematic. A problem that I glossed over in the earlier discussion is that *retire* also occurs in the passive, and it is sometimes followed by an NP that isn't its object but a temporal NP:

(14) The SEC's chief accountant, Clarence Sampson, will retire next year and may join the FASB, which regulates the accounting profession. (*WSJ* newswire 1987/09/25)

Similarly, one might feel that use of a PP[*in*] rather than PP[*to*] (as in (13a, c)) is simply a choice from several alternative ways of expressing one thing (a goal of motion), the choice being mainly determined by the NP that occurs within the PP rather than the verb. If we accept, following much linguistic work (Grimshaw 1979, 1990; Bresnan and Moshi 1990),

that selection is better described at a level of argument structure, with a subsequent mapping to surface subcategorization, we might instead adopt a model that uses two distributions to account for the subcategorization of a verb *V*. We would first propose that the verb occurs with a certain probability with a certain argument structure frame, conditioned on the verb, and presumably the context in a rich model.[16] Then we could assume that there is a *mapping* or *linking* of argument structure roles onto surface syntactic realizations. The actual surface subcategorization *Subcat* would be deterministically recoverable from knowing the argument structure *ArgStr* and the mapping *Mapping*. Then, assuming that the mapping depends only on the *ArgStr* and *Context*, and not on the particular verb, we might propose the model

$$P(Subcat|V, Context) = \sum_{\{Mapping, ArgStr : subcat(Mapping, ArgStr)=Subcat\}} P(ArgStr|V, Context) \cdot P(Mapping|ArgStr, Context).$$

For these probability distributions, we might hope to usefully condition via the class of the verb (Levin 1993) rather than on particular verbs, or to do the linking in terms of a mapping from semantic roles to grammatical functions (Bresnan and Zaenen 1990), the space of which is quite constrained.

However, such a frequency-based account is not satisfactory, because frequency of occurrence needs to be distinguished from argument status (Grimshaw and Vikner 1993; Przepiórkowski 1999b). Many arguments are optional, while some adjuncts are (almost) compulsory (e.g., a *how* adjunct for *worded* as in *He worded the proposal* ***very well***). Returning to our example of *retire* and using the same 1987 *WSJ* data, we find that the canonical temporal adjunct use of PP[*in*] (*in December*, *in 1994*) occurs with *retire* about 7 times as often as a PP[*to*] expressing a destination (and about 30 times as often as a PP[*in*] expressing a destination). If frequency is our only guide, a temporal PP[*in*] would have to be regarded as an argument. Linguists tend to regard the temporal PP as an adjunct and the destination as an argument on the basis of other criteria, such as phrasal ordering (arguments normally appear closer to the head than adjuncts) as in (15), iterability (adjuncts can be repeated) as in (16), or a feeling of semantic selection by the verb (any verb can have a temporal modifier):[17]

(15) a. Mr. Riley plans to retire to the Cayman Islands in December.
b. ?Mr. Riley plans to retire in December to the Cayman Islands.

(16) a. Mr. Riley plans to retire within the next five years on his birthday.
b. ?Mr. Riley plans to retire to the Cayman Islands to his ranch.

Beginning at least with Tesnière 1959, various criteria (morphological, syntactic, semantic, and functional) have been discussed for distinguishing arguments from adjuncts. Evidence for degree of selection (or argumenthood) could thus be derived from data in more sophisticated ways, by examining such phenomena as ordering, iteration, and semantic selectivity in examples that give relevant evidence. For example, Merlo and Leybold (2001) suggest also measuring head dependence (i.e., what range of heads a particular PP appears with), which is operationalized by counting the number of different verbs that occur with a given PP (where matching is broadened to include not just exact ⟨preposition, head noun⟩ matches, but also a match over a semantic classification of the head noun). A low number indicates argument status, while a high number indicates modifier status. We would then judge gradient argumenthood via a more complex evaluation of a variety of morphological, syntactic, semantic, and functional factors, following the spirit of diagram (22) below.

However, it remains an open question whether all these phenomena are reflections of a single unified scale of argumenthood or manifestations of various underlying semantic and contextual distinctions. It has long been noted that the traditional criteria do not always converge (Vater 1978; Przepiórkowski 1999b). Such lacks of convergence and categoricity have also been found in many other places in syntax; witness, for instance, problems with the categorical unaccusative/unergative division of intransitive verbs assumed by most linguistic theories (Napoli 1981; Zaenen 1993; Sorace 2000), or the quite gradient rather than categorical ability to extract, or extract from, adjuncts (Rizzi 1990; Hukari and Levine 1995).[18] There is clearly a trade-off between simplicity of the theory and factual accuracy (in either the categorical or the probabilistic case), but the presence of probabilities can make dealing with the true complexity of human language more palatable, because dominant tendencies can still be captured within the model.

8.3.2 On the Nature of Probabilistic Models of Syntax

In this section, I briefly address some of the questions and concerns that commonly come up about the use of probabilistic models in syntax.

8.3.2.1 Probabilities Are Not Just about World Knowledge What people actually say has two parts. One part is contingent facts about the world. For example, at the time I write this, people in the San Francisco Bay Area are talking a lot about electricity, housing prices, and stocks. Knowledge of such facts is very useful for disambiguation in NLP, but Chomsky was right to feel that they should be excluded from syntax. But there is another part to what people say, which is the way speakers choose to express ideas using the resources of their language. For example, in English, people don't often put *that* clauses before the verb, even though it's "grammatical" to do so:

(17) a. It is unlikely that the company will be able to meet this year's revenue forecasts.
b. #That the company will be able to meet this year's revenue forecasts is unlikely.

The latter statistical fact is properly to be explained as part of people's knowledge of language—that is, as part of syntax.[19] This is also the kind of knowledge that people outside the core community of theoretical syntacticians tend to be more interested in: sociolinguists, historical linguists, language educators, people building natural language generation and speech-understanding systems, and others dealing with real language. To account for such facts, as we have seen in this section, one can put probability mass functions over linguistic structures. Working out an estimate of an assumed underlying distribution is referred to as *density estimation*—normally regardless of whether a discrete probability mass function or a continuous probability density function is going to be used. Probability functions can be conditionalized in various ways, or expressed as joint distributions, but for the question of how people choose to express things, our eventual goal is *density estimation* for $P(\mathit{form}|\mathit{meaning}, \mathit{context})$.[20] In section 8.5, we will consider in more detail some methods for doing this.

One should notice—and could possibly object to the fact—that the domain of grammar has thus been expanded to include grammatical preferences, which will often reflect factors of pragmatics and discourse. Such preferences are traditionally seen as part of performance. However, I think that this move is positive. In recent decades, the scope of grammar has been expanded in various ways: people now routinely put into grammar semantic and discourse facts that would have been excluded in earlier decades. And it is not that there is nothing left in

performance or contingent facts about the world: whether the speaker has short-term memory limits or is tired or drunk will influence performance, but is not part of grammar. But principles that have a linguistic basis, and that will commonly be found to be categorical in some other language, are treated uniformly as part of grammar. This corresponds with the position taken by Prince and Smolensky (1993, 198): ''When the scalar and the gradient are recognized and brought within the purview of theory, Universal Grammar can supply the very substance from which grammars are built: a set of highly general constraints, which, through ranking, interact to produce the elaborate particularity of individual languages.'' Keller (2000, 29) cites a study by Wolfgang Sternefeld that concludes that the bulk of instances of gradience in grammar appear to be matters of competence grammar, rather than attributable to performance, contentfully construed (i.e., things that can reasonably be attributed to the nature of on-line human sentence processing), or extralinguistic factors.

8.3.2.2 One Can Put Probabilities over Complex Hidden Structure People often have the impression that one can put probability distributions only over things one can count (surface visible things like the words of a sentence and their ordering), and not over what probabilists call ''hidden structure'' (the things linguists routinely use to build theories, such as phrase structure trees, features with values, and semantic representations). However, this is certainly no longer the case: There is now a range of techniques for putting probability distributions over hidden structure, widely used for sophisticated probabilistic modeling—we saw some simple examples above for probabilities of argument structure frames. Even if the nature or values of assumed hidden structure have never been observed, one can still build probabilistic models. In such situations, a particularly well known and widely used technique is the Expectation Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1996), an iterative algorithm that attempts to estimate the values of hidden variables so as to make the observed data as likely as possible. This is not to say that there are not sometimes still formidable problems in successfully estimating distributions over largely hidden structure; but statisticians want to work on problems of putting probabilities over good model structures, not on denying that these structures exist. In general, it is now possible to *add* probabilities to any existing linguistic model. The most studied case of

this is for context-free grammars (Booth and Thomson 1973; Manning and Schütze 1999; Bod, this volume), but distributions have also been placed over other more complex grammars such as Lexical-Functional Grammars (Bod and Kaplan 1998; Johnson et al. 1999; Riezler et al. 2000).

This is not to say that one should not rethink linguistic theories when adding probabilities, as I have tried to indicate in the beginning part of this section. Nevertheless, especially in the psycholinguistic literature, the dominant terms of debate have been that one either has to stick to good old-fashioned rule/constraint systems from linguistic theory or move to connectionist modeling (Seidenberg and MacDonald 1999; Pinker 1999). Without in any way wishing to dismiss connectionist work, I think it is important to emphasize that this is a false dichotomy: "soft" probabilistic modeling can be done over rule systems and/or symbolic structures. The numerate connectionist community has increasingly morphed into a community of sophisticated users of a variety of probabilistic models.

8.3.2.3 Can Probabilities Capture the Notion of Grammaticality? Several recent pieces of work have argued with varying degrees of strength for the impossibility of collapsing notions of probability and grammaticality (Abney 1996; Culy 1998; Keller 2000). This is a difficult question, but I feel that at the moment this discussion has been clouded rather than clarified by people not asking the right probabilistic questions. It is certainly true that in general a sentence can have arbitrarily low probability in a model and yet still be perfectly good—in general this will be true of all very long sentences, as Abney (1996) notes. It is also the case that real language use (or a statistical NLP model aimed at disambiguating between meanings) builds in facts about the world. In these cases, the probability of different sentences reflects the nature of the world, and this clearly needs to be filtered out to determine a notion of grammatical acceptability independent of context. A profitable way to connect grammaticality and probability is perhaps to begin with the joint distribution $P(\mathit{form}, \mathit{meaning}, \mathit{context})$. We could then consider ways of marginalizing out the context and the meaning to be expressed.[21] We would not want to do this via the actual empirical distributions of contexts and ideas that people express, which reflect contingent facts about the world (see section 8.3.2.1), but by imposing some flat prior distribution, which is uninformative about the contingent facts of the world. For instance, in this distribution, people would be equally likely to travel to

North Dakota for holidays as to Paris. Using such an uninformative prior is a rough probabilistic equivalent of considering *possible worlds* in semantics. Under this analysis, forms might be considered gradiently grammatical to the extent that they had probability mass in the resulting marginal distribution.

Alternatively, one might think that the above suggestion of marginalizing out context is wrong because it effectively averages over possible contexts (see note 21), and a closer model to human judgments would be that humans judge the grammatical acceptability of sentences by assuming a most favorable real-world context. For this model, we might calculate probabilities of forms based on the probability of their most likely meaning in their most favorable context, by instead looking at[22]

$$P(f) = \frac{1}{Z} \max_m \max_c P(f, m, c).$$

However, while humans clearly have some ability to consider sentences in an imagined favorable context when judging (syntactic) "grammaticality," sociolinguistic and psycholinguistic research has shown that judgments of grammaticality are strongly codetermined by context and that people don't automatically find the best context (Labov 1972, 193–198; Carroll, Bever, and Pollack 1981). Nevertheless, Keller (2000) argues for continuing to judge grammaticality via acceptability judgments (essentially still intuitive judgments, though from a more controlled experimental setting) and modeling soft constraints by looking at relative acceptability. The issues here deserve further exploration, and there may be valuable alternative approaches, but, as exemplified at the beginning of this section, I tend to feel that there are good reasons for getting at synchronic human grammars of gradient phenomena from large data sets of actual language use rather than from human judgments. In part this is due to the unclarity of what syntactic acceptability judgments are actually measuring, as just discussed.

8.3.2.4 The Formal Learnability of Stochastic Grammars One might think that adding probability functions to what are already very complex symbol systems could only make formal learnability problems worse, but this is not true. An important thing to know is that adding probability distributions to a space of grammars can actually *improve* learnability. There is not space here for a detailed discussion of language acquisition data, but I wish to address briefly the formal learnability of syntactic systems.

If linguists know anything about formal learnability, it is that Gold (1967) proved that formal languages (even finite state ones) are unlearnable from positive evidence alone. This result, together with the accepted wisdom that children do not have much access to negative evidence and do not pay much attention to it when it is given, has been used as a formal foundation for the Chomskyan assumption that an articulated and constrained Universal Grammar must be innate.

There are many directions from which this foundation can be undermined. Some recent work has emphasized the amount of implicit negative evidence (through lack of comprehension, etc.) that children do receive (Sokolov and Snow 1994; Pullum 1996). One can question the strictness of the criterion of identifiability in the limit used by Gold (namely, that for any language in the hypothesis space, and for any order of sentence presentation, there must be a finite length of time after which the inference device always returns the correct language). Under a weaker criterion of approachability in the limit (each incorrect grammar is rejected after a finite period of time), context-free and context-sensitive grammars are learnable from positive evidence alone (Feldman et al. 1969).

But most importantly in this context, assuming probabilistic grammars actually makes languages easier to learn. The essential idea is that such a grammar puts probability distributions over sentences. If the corpus was produced from such a grammar, it is highly probable that any phenomenon given a high probability by the grammar will show up in the corpus. In contrast, Gold's result depends on the fact that the learner may observe an unrepresentative sample of the language for any period of time. Using probabilistic grammars, absence of a sentence from the corpus thus gives implicit negative evidence: if a grammar would frequently generate things that never appear in the corpus, that grammar becomes increasingly unlikely as the amount of observed corpus data grows. In particular, Horning (1969) showed that providing one assumes a denumerable class of possible grammars with a prior distribution over their likelihood, then stochastic grammars *are* learnable from positive evidence alone.

Such formal learnability results give guidance, but they assume conditions (stationarity, independence, etc.) that mean they are never going to exactly model "real-world conditions." But more generally it is important to realize that whereas almost nothing was known about learning around 1960 when the linguistic orthodoxy of "poverty of the stimulus" developed, a lot has changed since then. The fields of statistics and machine learning have made enormous advances, stemming from

both new theoretical techniques and the opportunities for computer modeling. One relevant result is that it is often easier to learn over continuous spaces than discontinuous categorical spaces, essentially because the presence of gradients can direct learning.[23]

8.3.2.5 Explanatory Value and Qualms about Numbers There are two final worries. First, one could doubt whether it will ever be possible to determine probabilities correctly for abstract but important linguistic features. What is the probability of a parasitic gap (Engdahl 1983) in a sentence given that it contains an adjunct extraction? Obviously, we can never hope to estimate all these probabilities exactly from a finite corpus, and we would be defeated by the nonstationarity of language even if we tried. But this is unimportant. Practically, all we need are reasonable estimates of probabilities, which are sufficient to give a good model of the linguistic phenomenon of interest. The difficulty of producing perfect probability estimates does not mean we are better off with no probability estimates. Theoretically, our assumption is that such probabilities do exist and have exact values: there is some probability that a sentence an English speaker utters will contain a parasitic gap. There is also a probability that the next sentence I utter will contain a parasitic gap. This probability may differ from the average for the population, and it may change when one conditions on context, the meaning to be expressed, and so on; but it exists. This probability simply captures an aspect of the behavior of the wetware in the human brain (and is, at this level, uncommitted even about whether some, much, or none of syntax is innate). People also sometimes just object to the use of numbers whatsoever in grammatical theory, independent of whether they can be determined. This strikes me as an invalid objection. Kroch (2001, 722) argues that "there is no doubt, however, that human beings, like other animals, track the frequencies of events in their environment, including the frequency of linguistic events."[24]

Second, one could think that there are so many numbers flying around that we can just fit (or "learn") anything and that there is thus no explanatory value with regard to the goal of describing possible human languages. This is a genuine worry: if the model is too powerful or general, it can essentially just memorize the data, and by considering varying parameter settings, one might find that the model provides no constraint on what languages are possible.[25] This is one possible objection to the Data-Oriented Parsing models developed by Bod (this volume): they do essentially just memorize all the data (to provide constraints on possible

linguistic systems, one needs to add substantive constraints on the representations over which the model learns and on the fragment extraction rules—and even then, there are no restrictions on the space of possible soft constraints). But this just means that we should do a good job at the traditional syntactic tasks of looking for constraints that apply across syntactic constructions, and looking for model constraints that delimit possible human linguistic systems. Some starting ideas about how to do this appear in section 8.5. While there are approaches to "softness" other than probabilities (such as using prototypes or fuzzy logic), the sound theoretical foundation of probability theory, and the powerful methods for evidence combination that it provides, makes it the method of choice for dealing with variable phenomena.

8.4 Continuous Categories

Earlier, I followed Chambers (1995) in emphasizing an alternative approach to linguistics that allows units that are *continuous* and *quantitative*. Any probabilistic method is quantitative, and the actual probabilities are continuous quantities, but most work using probabilities in syntax, and in particular the large amount of related work in statistical NLP, has put probabilities over discrete structures and values, of the kind familiar from traditional linguistics. However, while language mostly acts like a discrete symbol system (making this assumption workable in much of linguistics), there is considerable evidence that the hidden structure of syntax defies discrete classification. People continually stretch the "rules" of grammar to meet new communicative needs, to better align grammar and meaning, and so on. Such leakage in grammars leads to gradual change. As a recent example, the term *e-mail* started as a mass noun like *mail* (*I get too much junk e-mail*). However, it is becoming a count noun (filling the role of the nonexistent **e-letter*): *I just got an interesting e-mail about that*. This change happened in the last decade: I still remember when this last sentence sounded completely wrong (and ignorant (!)). It then became commonplace, but still didn't sound quite right to me. Then I started noticing myself using it.

As a more detailed example of the blurring of syntactic categories during linguistic change, consider the group of English words sometimes called "marginal prepositions" (Quirk et al. 1985, 666).[26] These words have the form of present participles of verbs, and they are used as verbs, but they also appear to function as prepositions. Examples include *concerning*, *supposing*, *excepting*, *according*, *considering*, *regarding*, and

following. Historically and morphologically, these words began as verbs, but they are moving in at least some of their uses from being participles to being prepositions. Some still clearly maintain a verbal existence (e.g., *following*, *concerning*, *considering*); for others, their verbal character is marginal (e.g., *according*, *excepting*); for still others, it is completely lost (e.g., *during* (previously a verb; cf. *endure*),[27] *pending*, and *notwithstanding*).

In the remainder of this section, I will illustrate with *following*, which is one of the more recent participles to make the move and which has been the subject of an earlier corpus study (Olofsson 1990), to which I am indebted for various references cited below. While some prepositional uses of *following* are recorded quite early, as in (18),

(18) Following his ordination, the Reverend Mr Henry Edward intends to go to Rome. (1851)

such uses seem to have become considerably more frequent in the second half of the twentieth century. One manifestation of this, as so often in cases of language change, is that prescriptive commentators on language began to condemn the innovation. The first edition of Fowler's well-known (British) dictionary of usage makes no mention of *following* but elsewhere shows considerable liberality about this process of change: "there is a continual change going on by which certain participles or adjectives acquire the character of prepositions or adverbs, no longer needing the prop of a noun to cling to. . . . [We see] a development caught in the act" (Fowler 1926). However, Gowers (1948, 56) declares, "*Following* is not a preposition. It is the participle of the verb *follow* and must have a noun to agree with." With its continued spread, the 1954 revised edition of Fowler, edited by Gowers, still generally condemns the temporal usage, but softens the dictum by saying that it can be justified in certain circumstances.

One test for a participial use is that participial clauses require their subject to be controlled by a noun phrase in the main clause (19a), while prepositions do not. So (19b) is possible only if we anthropomorphize the truck to control the subject of the verb *seeing*, whereas the preposition *after* in (20) does not have any controlled subject:

(19) a. Seeing the cow, he swerved violently.
b. #Seeing the car, the truck swerved violently.

(20) After the discovery, the price of gold began to drop.

We can observe the following development. The earlier cases clearly have a participial verb with the basic motion sense of *follow* (21a). By extension, it later became common to use *follow* to indicate place (21b) or time (21c). In these examples, it is still plausible to regard a noun phrase in the main clause (including the unexpressed subject of the imperative in (21b)) as coreferential with the subject of *following*, but such an assumption is not necessary: we could replace *following* with the preposition *after* with virtually no change in interpretation. This ambiguity of analysis allowed the temporal use to become generalized, and *following* began to be used as a preposition with no possible controller (21d–e):

(21) a. They moved slowly, toward the main gate, following the wall.
b. Repeat the instructions following the asterisk.
c. This continued most of the week following that ill-starred trip to church.
d. He bled profusely following circumcision.
e. Following a telephone call, a little earlier, Winter had said ...

There is a tradition in linguistics of imposing categoricity on data. One sees this also in NLP. For example, the tagging guidelines for the Penn Treebank (Santorini 1990) declare about such cases that

putative prepositions ending in *-ed* or *-ing* should be tagged as past participles (VBN) or gerunds (VBG), respectively, not as prepositions (IN).

According/VBG to reliable sources
Concerning/VBG your request of last week

Maintaining this as an arbitrary rule in the face of varying linguistic usage is essentially meaningless. One can avoid accepting gradual change by stipulating categoricity. But the results of such moves are not terribly insightful: it seems that it would be useful to explore modeling words as moving in a continuous space of syntactic category, with dense groupings corresponding to traditional parts of speech (Tabor 2000).[28]

8.5 Explaining More: Probabilistic Models of Syntactic Usage

Much of the research in probabilistic NLP has worked with probabilities over rules, such as the PCFG models discussed by Bod (this volume). Such an approach probably seems old-fashioned to most linguists who have moved to thinking in terms of constraints on linguistic representations. In this section, I adopt such an outlook and examine the space of

probabilistic and categorical linguistic models, not from the point of view of the substantive claims of particular theories, but from the perspective of how constraints interact.

8.5.1 A Linguistic Example

I integrate with the theory discussion of a particular syntactic example drawn from the interaction of passive, person, and topicality. This example is drawn from joint work on stochastic syntax with Joan Bresnan and Shipra Dingare; see Bresnan, Dingare, and Manning 2001 and Dingare 2001 for more details and references. The example is very much simplified for expository purposes, but still has enough structure to permit comparing and contrasting different approaches and to build example probabilistic models. The central aim of the example is to support the following proposition: The same categorical phenomena that are attributed to hard grammatical constraints in some languages continue to show up as soft constraints in other languages. This argues for a grammatical model that can account for constraints varying in strength from soft to hard within a uniform formal architecture, and for the explanatory inadequacy of a theory that cannot. Compare these remarks by Givón (1979, 28) contrasting a categorical restriction on indefinite subjects in Krio with the dispreference for them in English:

> But are we dealing with two different kinds of facts in English and Krio? Hardly. What we are dealing with is apparently the very same *communicative tendency*—to reserve the subject position in the sentence for the *topic*, the old-information argument, the "continuity marker." In some languages (Krio, etc.) this communicative tendency is expressed at the *categorial* level of 100%. In other languages (English, etc.) the very same communicative tendency is expressed "only" at the *noncategorial* level of 90%. And a transformational-generative linguist will then be forced to count this fact as competence in Krio and performance in English.

Ideas of stronger and weaker constraints are common in the typological and functionalist syntactic literatures, but until now there has been a dearth of formal syntactic models that can describe such a situation. Rather, the soft constraints are treated as some sort of performance effect, which bears *no* theoretical relation to the categorical rules of other languages, which would be seen as deriving, for instance, from parameter settings in Universal Grammar.

A traditional linguist could argue that the competence system is categorical, with only hard constraints, but that there is a lot of noise in the performance system, which means that ungrammatical sentences are

sometimes produced. I presented the first argument against this position in section 8.3.1, where the problem of the nonidentifiability of the competence grammar (where to draw the line between grammatical and ungrammatical sentences) was discussed. In this section, I concentrate on the stronger argument that this model is unexplanatory because one would have to invest the performance system with soft constraints that should be treated uniformly with hard constraints of the competence grammar of other languages. At any rate, one needs a model of any noise attributed to the performance system; simply leaving it unspecified, as in most linguistic work, means that the model makes no empirical predictions about actual language use.

Consider an event where a policeman scolded me, where I'm part of the existing discourse (perhaps just by virtue of being a speech act participant), but mention of the policeman is new. This *input* semantic representation will be described by an extended *argument structure* as *see*⟨*policeman* **[new]**, *me* **[old]**⟩. In English, this idea would most commonly be expressed as an *output* by either a simple active or a simple passive sentence, as in (22). An equivalent pair of choices is available in many languages, but in others, one of these choices is excluded.

(22)

Constraints = Features = Properties	Linking	Discourse	Person
Input: *scold*⟨*policeman* **[new]**, *me* **[old]**⟩	f_1 ✓Subj/Ag *Nonsubj/Ag	f_2 ✓Subj/Older *Subj/Newer	f_3 ✓$1/2 > 3$ *$3 > 1/2$
a. *A policeman scolded me*	1	0	0
b. *I was scolded by a policeman*	0	1	1

What determines a speaker's choice here? There are many factors, but in the simplified example in (22), we will consider just three. The constraints can be stated either positively or negatively, and I provide mnemonic names for each polarity. Use of negative constraints is de rigeur in Optimality Theory (Prince and Smolensky 1993), but the choice is arbitrary, and for the generalized linear models that I discuss later, it may be more helpful to think of them as positive soft constraints ("The subject should be old information," etc.).

1. *A linking constraint:* $^{\checkmark}$Subj/Ag = *Nonsubj/Ag. It is preferable for the subject of the sentence to express the entity that is performing the action of the sentence, that is, its *agent*.
2. *A discourse constraint:* $^{\checkmark}$Subj/Older = *Subj/Newer. For purposes of topic continuity, it is better for the previously mentioned (i.e., older) entity to be the subject.
3. *A person constraint:* $^{\checkmark}1/2 > 3 = {}^{*}3 > 1/2$. Because the 1st/2nd person speech act participants are the central foci of discourse empathy, it is better for one of them to be the subject, the thing that the sentence is perceived to be about, than some other 3rd person entity.

These three constraints (also referred to as features or properties) are shown in (22), labeled with both their mnemonic names and the expressions f_1, f_2, and f_3. The first constraint would favor use of the active, while the other two would favor use of the passive. In (22), a 1 means a constraint is satisfied, and a 0 means that it is violated (i.e., a 0 corresponds to a * in an Optimality Theory tableau).

The constraints introduced here are all well supported crosslinguistically. Within Optimality Theory, their basis can be captured via the concept of *harmonic alignment* (Prince and Smolensky 1993). When there are two scales of prominence, one structural (e.g., syntactic positions) and the other substantive (e.g., semantic features), harmonic alignment specifies a way of building an (asymmetric) partially ordered set of constraints in the form of linear scales, referred to as *constraint subhierarchies*. These give putatively universal limits on constraint ordering/strength. For instance, following Aissen (1999), we could perform harmonic alignment between a surface grammatical relations hierarchy and a thematic hierarchy to derive the importance of the linking constraint:

(23)

Hierarchies	*Harmonic alignment*	*Constraint subhierarchies*
Agent $\succ$ Patient	Subj/Ag > Subj/Pt	*Subj/Pt $\gg$ *Subj/Ag
Subj $\succ$ Nonsubj	Nonsubj/Pt > Nonsubj/Ag	*Nonsubj/Ag $\gg$ *Nonsubj/Pt

The second subhierarchy, restated positively, gives the result that universally a feature rewarding agents that are subjects must be ranked higher than a feature rewarding patients that are subjects. For simplicity, in the remainder, I only use the higher-ranked constraint, $^{\checkmark}$Subj/Ag = *Nonsubj/Ag. Similarly, it is well supported in the typological literature that there is a person hierarchy, where 1st and 2nd person outrank 3rd;

that is, *local* person outranks *nonlocal* (Silverstein 1976; Kuno and Kaburaki 1977; Givón 1994). This constraint could be modeled by harmonic alignment of the person hierarchy with the grammatical relations hierarchy, but I have expressed it as a simple relational constraint—just to stress that the models allow the inclusion of any well-defined property and are not restricted to one particular proposal for defining constraint families. At any rate, in this manner, although our constraints are possibly soft, not categorical, we can nevertheless build in and work with substantive linguistic hypotheses.

8.5.2 Categorical (Constraint-Based) Syntactic Theories

The standard but limiting case of a syntactic theory is the categorical case, where outputs are either grammatical or ungrammatical. The grammar assumes some kind of representation, commonly involving tree structures and/or attribute-value representations of grammatical features. The grammar explains how the representation is used to describe an infinite set of sentences, and it defines a number of hard (categorical) constraints on those representations. Generative approaches assume a common set of representations and principles, Universal Grammar (UG), underlying all languages, plus language-specific constraints, which are conjoined with the ones given by UG, and/or parameterized constraints, which are part of UG but can vary on a language-particular basis. A grammatical sentence must satisfy all the constraints. Sentences that do not are regarded as syntactically ill formed or ungrammatical. For instance, there is an *agreement constraint*, which ensures that, for the passive sentence in (22), (24a) is grammatical, while (24b) is not:

(24) a. I was scolded by a policeman.
b. *I were scolded by a policeman.

A grammar of a language is the conjunction of a large set of such constraints over the representations. One can determine the grammatical sentences of the language by solving the resulting large constraint satisfaction problem (Carpenter 1992; Stabler 2001). In conventional, categorical NLP systems, ungrammatical sentences do not parse and so cannot be processed.[29]

For all the main forms of formal/generative syntax, the grammar of English will do no more than say that both the active and the passive outputs in (22) are possible grammatical sentences of English. Since none

of the constraints $f_1, \ldots, f_3$ are categorical in English, none would be part of the grammar of English. In other words, the grammar says nothing at all about why a speaker would choose one output or the other.

However, in many languages of the world, one or another of these constraints is categorical, and a speaker would not be able to use sentences corresponding to both of the English outputs in the situation specified in (22).[30] In general, passives are marked (Greenberg 1966; Trask 1979): many languages lack a passive construction, passives are normally more restricted language internally, and normally they are morphologically marked. While this could be partly due to historical happenstance, I see it as reflecting the constraint f_1. If the constraint f_1 is categorical in a language, then passive forms will never occur:[31]

(25)

	scold⟨*policeman* **[new]**, *me* **[old]**⟩		*Nonsubj/Ag
☞	Active: S_{ag}, O_{pt}	*A policeman scolded me*	
	Passive: S_{pt}, Obl_{ag}	*I was scolded by a policeman*	*!

The effects of the person hierarchy on grammar are categorical in some languages, most famously in languages with inverse systems, but also in languages with person restrictions on passivization. In Lummi (Coast Salishan, United States and Canada), for example, if one argument is 1st/2nd person and the other is 3rd person, then the 1st/2nd person argument must be the subject. The appropriate choice of passive or active is obligatory to achieve this (26)–(27). If both arguments are 3rd person (28), then either active or passive is possible (Jelinek and Demers 1983, 1994):[32]

(26) a. *'The man knows me/you.'
b. xˇči-t-ŋ=sən/=sx^{w} ə cə swəyʔqəʔ
know-TR-PASS = 1/2.NOM.SUBJ by the man
'I am/You are known by the man.'

(27) a. xˇči-t=sən/=sx^{w} cə swəyʔqəʔ
know-TR = 1/2.NOM.SUBJ the man
'I/You know the man.'
b. *'The man is known by me/you.'

(28) a. xčí-t-s cə swəy?qə? cə swi?qo?əɬ
know-TR-3.ERG.SUBJ the man the boy
'The man knows the boy.'

b. xčí-t-ŋ cə swi?qo?əɬ ə cə swəy?qə?
know-TR-PASS the boy by the man
'The boy is known by the man.'

Such interactions of person and voice (or inversion) occur in a considerable number of Native American languages, and also more widely (Bresnan, Dingare, and Manning 2001). To account for this, we can say that in Lummi the person constraint f_2 is categorical (i.e., part of the grammar of Lummi), but the others are not. Some of the Lummi cases then look like this:

(29)

	scold⟨ policeman **[new]**, *me* **[old]**⟩			*3 > 1/2
	Active:	S_3, O_1	*A policeman scolded me*	*!
☞	Passive:	S_1, Obl_3	*I was scolded by a policeman*	

(30)

	scold⟨ policeman **[new]**, *Fred* **[old]**⟩			*3 > 1/2
☞	Active:	S_3, O_3	*A policeman scolded Fred*	
☞	Passive:	S_3, Obl_3	*Fred was scolded by a policeman*	

In a categorical model, if constraints come into conflict, then the form is ungrammatical. Therefore, if a language had a categorical version of both the linking constraint and the person constraint, there would be no way to express the idea *scold⟨ policeman*, *me*⟩. Expressive gaps of this sort may occasionally occur in language (perhaps, an example is expressing ideas that violate relative clause *wh*-extraction constraints—Pesetsky 1998), but are extremely rare. In general, languages seem to conspire to avoid such gaps. In a categorical grammar, the linguist has to build such conspiracies into the grammar by restricting the basic constraints, either by adding complex negated conditions on constraints by hand or by making use of ideas like the elsewhere principle (Kiparsky 1973).

The typological factors of person and linking are present in the analysis of passive and voice in Kuno and Kaburaki 1977. Again, we see in categorical work a tendency to view as ungrammatical structures that are

simply very marked, though perhaps licensed in special discourse circumstances. For example, Kuno and Kaburaki star as ungrammatical sentences such as (31):

(31) *Mary was hit by me.

On their account, it violates the Ban on Conflicting Empathy Foci: on the grammatical relations hierarchy, the choice of *Mary* as subject implies that the speaker empathizes with Mary, while the speech-act participant hierarchy dictates that the speaker must empathize most with himself. A conflict is possible in an unmarked active sentence, but not in a marked sentence type. However, Kato (1979) takes Kuno and Kaburaki to task for such claims, providing corpus evidence of passives with 1st person agents, such as (32):

(32) Gore [Vidal] never lacked love, nor was he abandoned by me. (*Time*)

It is somewhat unclear what Kuno and Kaburaki were claiming in the first place: while in some cases they explain a * ungrammaticality mark by invoking the Ban on Conflicting Empathy Foci, in others they suggest that special contexts can make violations of empathy constraints possible. Their conclusion (p. 670) points to the kind of approach to optimization over soft constraints that I am advocating here: "[V]iolations of empathy constraints sometimes yield totally unacceptable sentences; at other times, especially when other factors make up for the violations, they yield only awkward or even acceptable sentences." Such situations can only be written about informally when working in a categorical formal framework. When working in a probabilistic syntactic framework, they can be *formally modeled*.

8.5.3 Optimality Theory

Standard Optimality Theory (OT) is not a probabilistic framework,[33] but it is a useful in-between point as we proceed from categorical to probabilistic grammars. OT differs from standard categorical grammars by assuming that optimization over discrete symbol structures via ranked, violable constraints is fundamental to the cognitive architecture of language. Human language is described by a set of universal constraints, which are hypothesized to be present in all grammars, but they are more or less active depending on their ranking relative to other constraints. The (unique) grammatical output for an input is the one that optimally

satisfies the ranked constraints of a language, where satisfying a higher-ranked constraint is judged superior to satisfying any number of lower-ranked constraints. Consider again (22). If all three constraints were present and categorical, there would be no output: the theory would simply be inconsistent. However, under the OT conception of ranked, violable constraints, we can maintain all the constraints and simply rank the person constraint highest. As the person constraint is categorical for Lummi (*undominated* in OT terminology), it will determine the passive output for (22), as shown in (33b). When there are two 3rd person arguments, Lummi falls back on other constraints. Since passivization is still possible here, we conclude that the linking constraint is ranked low and that other information structure constraints, such as the discourse constraint, determine the optimal output. For example, if the agent is the topic, that is, the older information, then the active will be chosen over the passive (33a), and vice versa.

(33) a.

	Input: v⟨ag/3/old, pt/3/new⟩	*3 > 1/2	*Subj/Newer	*Nonsubj/Ag
☞	Active: S_{ag}, O_{pt}			
	Passive: S_{pt}, Obl_{ag}		*!	*

b.

	Input: v⟨ag/3/new, pt/1/old⟩	*3 > 1/2	*Subj/Newer	*Nonsubj/Ag
	Active: S_{ag}, O_{pt}	*!	*	
☞	Passive: S_{pt}, Obl_{ag}			*

The introduction of constraint ranking can potentially give a better description of Lummi syntax. It shows how secondary constraints come into play when the most important constraints are satisfied (a phenomenon referred to as ''emergence of the unmarked''). The output form does not have to obey all these constraints: it is most important that person be observed, but if this constraint does not differentiate forms, then lower-ranked constraints show through and determine the grammatical output. This automatically gives us the kind of *elsewhere hierarchy* that has been widely observed in linguistic systems (Kiparsky 1973). This ranking of

constraints, and the ability of less highly ranked constraints to show through, is the hallmark of OT.

A problem with OT is that it flies in the face of Sapir's words with which this chapter began: OT predicts that there should always be a unique output for every input. The highest-ranked differentiating constraint always determines things, giving a unique output. This is tenable in some areas of linguistics, but it goes against widespread *variation* in the use of language, not just across members of a community, but usually also by each individual (Guy 1981; Labov 1994; Kroch 2001). Because of the generally problematic nature of this theoretical assumption, a small industry has arisen trying to work variable outputs into OT (Nagy and Reynolds 1997; Boersma 1997; Pesetsky 1998; Müller 1999; Anttila 2001). For the particular case here, there are existing OT accounts of diathesis alternations based on typological evidence (Legendre, Raymond, and Smolensky 1993; Ackema and Neeleman 1998; Aissen 1999; Lødrup 1999), but once one incorporates the role of discourse and information structure into diathesis prediction, the consequent variation suggests the use of a stochastic account. Below, I will discuss Boersma's stochastic OT, the best-motivated and most thoroughly probabilistic extension to OT.

It is interesting that, historically (for at least Paul Smolensky, one of the primary architects of OT), OT is a retreat from a quantitative framework that does harmony maximization over numerical soft constraints. The earlier approach of Harmonic Grammar (Smolensky 1986; Legendre, Miyata, and Smolensky 1990), although couched in a connectionist network, is mathematically extremely close to the loglinear models discussed below (Smolensky and Legendre 2000). Prince and Smolensky (1993, 198) motivate OT by noting that in practice, "[o]rder, not quantity (or counting), is the key in Harmonic Grammar–based theories. In Optimality Theory, constraints are ranked, not weighted; harmonic evaluation involves the abstract algebra of order relations rather than numerical adjudication between quantities." This strict domination ordering of constraints is argued by Prince and Smolensky to be the basic, fundamental, ubiquitous, universal, and unmarked means of constraint interaction within linguistics.

It is true that such a ranking is often sufficient (just as it is true for some purposes that categorical constraints are sufficient), but the need to handle variability is one key reason for believing that sometimes more is needed. The other reason is to be able to handle the phenomenon of

"ganging up," where multiple lesser constraint violations are deemed to make an output worse than another with just one, more serious constraint violation. There is partial recognition of and a partial treatment for this phenomenon within OT via the theory of local conjunction (Smolensky 1993, 1997), whereby a conjunction of two constraints can be given a position in the constraint ordering separate from (i.e., higher than) the two individual constraints. But as soon as one wishes to allow a general conception of constraints having a combined effect (i.e., "ganging up"), then one needs to bring back the numbers. An ordering alone is insufficient to assess when multiple lesser constraints will or will not overrule higher-ranked constraints, or (in terms of the local conjunction perspective) where a local conjunction should be placed within the overall constraint ranking. While Smolensky motivates local conjunction as a limited but necessary deviation from the basic method of linguistic constraint interaction, it goes somewhat against the spirit of OT and suggests the model is not quite right: at the end of the day, a combination of constraint violations is deemed worse than a violation of just the highest-ranked individual constraint.

8.5.4 The Linguistic Example, Continued

Considering again (22), none of the three constraints shown are categorical in the grammar of English, but all of them play a role. All else being equal, old information is more commonly the subject, local persons are more commonly the subject, and agents are more commonly the subject.

Quantitative data can demonstrate that a language exhibits soft generalizations corresponding to what are categorical generalizations in other languages. A probabilistic model can then model the strength of these preferences, their interaction with each other, and their interaction with other principles of grammar. By giving variable outputs for the same input, it can predict the statistical patterning of the data. Beyond this, the model allows us to connect such soft constraints with the categorical restrictions that exist in other languages, naturally capturing that they are reflections of the same underlying principles. This serves to effectively link typological and quantitative evidence.

Bresnan, Dingare, and Manning (2001) collected counts over transitive verbs from parsed portions of the Switchboard corpus of conversational American English (Godfrey, Holliman, and McDaniel 1992), analyzing for person and active versus passive. Switchboard is a database of spontaneous telephone conversations between anonymous callers spread

across the United States. We chose Switchboard because it is a large, parsed, spoken corpus (about 1 million words are parsed). The parsing made our data collection significantly easier. Not only is spoken material more natural; it also includes numerous instances of 1st and 2nd person, whereas in many of the (mainly written) corpora available to us, 1st and 2nd person are extremely rare. On the other hand, full passives (ones with *by* phrases) turn out to be very rare, and so even though we counted around 10,000 clauses, the results table discussed below still has a couple of zeroes in it.[34]

English does not have a categorical constraint of person on passivization, but the same phenomenon is nonetheless at work as a soft constraint. We found that the same disharmonic person/argument associations that are avoided categorically in languages like Lummi also depress or elevate the relative frequency of passives in the Switchboard corpus. Compared to the rate of passivization for inputs of 3rd persons acting on 3rd persons (1.2%), the rate of passivization for 1st or 2nd person acting on 3rd is substantially depressed (0%) while that for 3rd person acting on 1st or 2nd (2.9%) is substantially elevated; see table 8.1.[35] The leftmost column in the table gives the four types of inputs (local person acting on local, local acting on nonlocal, etc.). For each input, we calculate the rate of passivization from the number of times that input was realized as passive. The percentage of full passives in spoken English is very small: most passives involve suppression of the agent (a further 114 examples with a local person patient, and 348 examples with a 3rd person patient).[36] Person is only a small part of the picture in determining the choice of active/passive in English (information structure, genre, and the like are more important—but we left consideration of information structure to further research because it is much more difficult to clas-

Table 8.1
Counts of actives and full passives in the Switchboard corpus, broken down by agent and patient person

Event roles	# act.	# pass.	% act.	% pass.
v⟨ag/1,2; pt/1,2⟩	179	0	100.00	0.00
v⟨ag/1,2; pt/3⟩	6,246	0	100.00	0.00
v⟨ag/3; pt/3⟩	3,110	39	98.76	1.24
v⟨ag/3; pt/1,2⟩	472	14	97.11	2.89
Total/Mean	10,007	53	99.47	0.53

sify). Nevertheless, there is a highly significant effect of person on active/passive choice.[37] The very same hard constraint found in Lummi appears as a soft constraint in English.

8.5.5 Stochastic Optimality Theory

Stochastic OT (Boersma 1997, 1998; Boersma and Hayes 2001) basically follows OT, but differs from it in two essential ways. First, constraints are not simply ordered, but in fact have a value on the continuous scale of real numbers. Constraints are specific distances apart, and these distances are relevant to what the theory predicts. Second, there is stochastic evaluation, which leads to variation and hence to a probability distribution over outputs from the grammar for a certain input. At each evaluation, the value of each constraint is perturbed by temporarily adding to its ranking value a random value drawn from a normal distribution. For example, a constraint with a rank of 99.6 could be evaluated at 97.1 or 105. It is the constraint ranking that results from this perturbation that is used in evaluation. Hence, the grammar constrains but underdetermines the output. One could think that this model of random perturbation is rather strange: does a speaker roll dice before deciding how to express him- or herself? There may be some inherent randomness in human behavior, but principally the randomness simply represents the incompleteness of our model and our uncertainty about the world (Bresnan and Deo 2000). Linguistic production is influenced by many factors that we would not wish to put into a syntactic model.[38] We cannot know or model all of these, so we are predicting simply that if one averages over all such effects, then certain outputs will occur a certain proportion of the time.

For instance, for the constraints in figure 8.2, *Nonsubj/Ag and *Subj/Newer would sometimes rerank, while a reranking of *Nonsubj/Ag and $^{*}3 > 1/2$ would be quite rare, but still occasionally noticeable.[39] In other

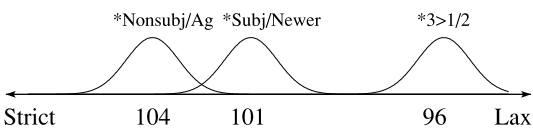

Figure 8.2
A stochastic OT model of the English passive. This is a constructed model, not based on real data. Note that the scale is reversed so the highest-ranked constraint is to the left, as in standard OT.

words, the situation would resemble that observed for the English active/passive choice.[40] Active forms would usually result, as the constraint to disprefer passives usually wins; but sometimes the discourse constraint would determine the optimal output, with passives being used to get old information into the subject position; and occasionally the person constraint would win out, causing a preference for passivization when the patient is local person, even when the outputs differ in how well they meet the discourse constraint. In some of these last rare cases, the evaluation constraint ranking for English would look just like the regular constraint ranking in Lummi shown in (33). This exemplifies the overarching point that variable outputs within grammars can reflect the variation between grammars in a way that theories of soft constraints can illuminate. Indeed, intraspeaker variation appears to be constrained by the same typological markedness factors that play a role in interspeaker variation (Cheshire, Edwards, and Whittle 1989, 1993; Ihalainen 1991; Cheshire 1991, 1996; Schilling-Estes and Wolfram 1994; Anderwald 1999; Bresnan and Deo 2000).

The advantages of the generalization to stochastic evaluation over standard OT include a robust learning algorithm and the capability of learning frequency distributions, which permits a unified account of both variable and categorical data (Boersma and Hayes 2001). In syntax and semantics, this approach has been adopted to explain problems of optionality and ambiguity that arise with nonstochastic OT (Bresnan and Deo 2000; Bresnan, Dingare, and Manning 2001; Koontz-Garboden 2001; Clark 2001; Boersma, in press). Like OT, stochastic OT can explain both crosslinguistic implicational generalizations (i.e., typological asymmetries) and within-language "emergence of the unmarked" effects. Stochastic OT thus preserves the core typological predictions of ordinal OT arising from universal subhierarchies of the kind briefly mentioned in section 8.5.1. This has the consequence that substantive linguistic claims (such as that making an agent a nonsubject must be more marked than making a patient a nonsubject) are still embodied in the theory. As a consequence, many types of languages (those that categorically violate hierarchies) are not learnable under the theory, thereby accounting for the typological gaps. A stochastic OT grammar with the constraint set shown cannot be made to learn an "anti-Lummi" where passivization occurs when there is a local person agent and a 3rd person patient.

Although this has only just begun to be explored (Bresnan, Dingare, and Manning 2001; Clark 2001), stochastic OT also seems quite promis-

ing for modeling historical change. If a process of historical change is modeled by the movement in strength of a constraint along the ranking scale, as implied by stochastic OT, then (all else being equal) smooth changes in the relative frequencies of usage are predicted, just as is generally observed in practice. In particular, assuming that the choice between output *A* and *B* for an input *I* is crucially dependent on the difference in ranking of two constraints α and β, and that α is moving at a constant rate to meet and then cross β in ranking, then stochastic OT predicts that the ratio between outputs *A* and *B* over time will be a logistic curve, of the sort shown in figure 8.4. That is, we would expect to see the kind of S-curve between the proportion of occurrences of the two outputs that has been widely noted in historical and sociolinguistics (Weinreich, Labov, and Herzog 1968; Bailey 1973; Kroch 2001). A stochastic grammatical model is in many ways a more plausible model for syntactic change than the competing-grammars model prevalent in generative grammar (Kroch 2001). By design (reflecting the orthodoxy of generative grammar), the competing-grammars model excludes variation from within grammars and places it as a choice among several competing grammars. Such a model can only easily generate variable outputs characterized by covariation between phenomena, captured by the choice of whether grammar *A* or grammar *B* was selected to generate a sentence. Where parameters vary independently or change at different rates, the competing-grammars model requires an exponential number of competing grammars (Guy 1997; Bresnan and Deo 2000), and within-sentence variation (as in (2j)) requires intrasentential switching among them. A stochastic grammar is both more economical, by localizing points of difference, and more expressive, by allowing variation within a single grammar.

A potential limitation of stochastic OT is that it still does not provide for doing optimization over the combination of all the constraint values—some number of highly ranked constraints at evaluation time will determine the winner, and the rest will be ignored. In particular, lower-ranked constraint violations cannot "gang up" to defeat a higher-ranked constraint. Each will individually occasionally be ranked over the higher constraint, and having multiple violations will cause this to happen more often, but there is no true "ganging up." Providing that the constraints are well spaced, a form that violates 10 lesser-ranked constraints will still almost always lose to one that violates 1 high-ranked constraint. This is potentially problematic as there is some evidence for ganging-up

effects in syntax (Keller 2000) and phonetics/phonology (Guy 1997; Flemming 2001). Further research is needed to see whether the constrained model of constraint interaction provided by stochastic OT is adequate for all linguistic systems. Kuhn (2001a,b) suggests that a stochastic OT model is quite promising in generation, when choosing on linguistic grounds among a fairly limited number of candidates, but seems less plausible as a parsing/interpretation model where in general most of the readings of an ambiguous sentence can be made plausible by varying context and lexical items in a way not easily modeled by an OT approach (the decisive evidence can come from many places). This would fit with the facts on the ground, where OT models (stochastic or otherwise) have been mainly employed for generation (though see Kuhn 2001a for discussion of bidirectional OT), whereas work in NLP, which focuses mainly on parsing, has tended to use more general feature interaction models.

8.5.6 Loglinear Models

Within speech and natural language processing, there has been a large movement away from categorical linguistic models to statistical models. Most such models are based on either markov chain models or branching process models (Harris 1963), such as PCFGs (Bod, this volume). An important advance in this area has been the application of loglinear models (Agresti 1990) to modeling linguistic systems (Rosenfeld 1994; Ratnaparkhi 1998). In particular, such statistical models can deal with the many interacting dependencies and the structural complexity found in modern syntactic theories, such as constraint-based theories of syntax, by allowing arbitrary features to be placed over linguistic representations and then combined into a probability model (Abney 1997; Johnson et al. 1999; Riezler et al. 2000).

In the above OT and stochastic OT frameworks, the output (''grammatical'') linguistic structures are those that are optimal among the subset of possible linguistic structures that can be generated to correspond to a certain input. For example, in OT generation, the grammatical linguistic structures are precisely those that are optimal with respect to all other possible structures with the same semantic interpretation or meaning. This corresponds to a *conditional probability distribution* in the case of loglinear models, in which the probability of an output is conditioned on the input semantic interpretation. For example (22), the probability of different outputs o_k for an input i would be modeled as a conditional

distribution in terms of weights w_j given to the different features f_j as follows:

$$P(o_k|i) = \frac{1}{Z(i)} e^{w_1 \cdot f_1(o_k,i) + w_2 \cdot f_2(o_k,i) + w_3 \cdot f_3(o_k,i)}.$$

Here $Z(i)$ is just a normalization constant—a technical way of making sure a probability distribution results, by scaling the exponential terms to make sure that the sums of the probabilities for all the o_k add to one. Such *exponential models* are also called *maximum entropy models* because an exponential distribution maximizes the entropy of the probability distribution subject to the given constraints, and *loglinear models* because taking the log of both sides results in the linear model:

$$\log P(o_k|i) = w_1 \cdot f_1(o_k,i) + w_2 \cdot f_2(o_k,i) + w_3 \cdot f_3(o_k,i) - \log Z(i).$$

While the features used can have arbitrary values, in our example (22), the features are just binary. Especially in this case, the above formula has an easy interpretation: the log of the probability of an output is straightforwardly related to the sum of the weights for the features that are satisfied (i.e., in OT terms, unviolated).

For example, if we assume the weights $w_1 = 4$, $w_2 = 3$, and $w_3 = 2$, then we have that

$$\log P(\text{active}) = w_1 - \log Z = 4 - \log Z,$$
$$\log P(\text{passive}) = w_2 + w_3 - \log Z = 5 - \log Z.$$

Assuming that these are the only possible outputs for this input, we can easily calculate the normalization term Z (using $\frac{1}{Z}(e^5 + e^4) = 1$ to obtain Z) and get the result that:

$$P(\text{active}) = 0.27,$$
$$P(\text{passive}) = 0.73.$$

There are actually two ways we could use this result. By optimizing over loglinear models—that is, by taking the most probable structure, $\arg\max_k P(o_k|i)$—one can determine a unique output that best satisfies multiple conflicting constraints, in a manner similar to OT, but allowing for arbitrary ganging up of features. That is, for this particular input and these weights, the model would always choose a passive output. But note crucially that it would do this because constraints f_2 and f_3 gang up to beat out constraint f_1, which is the highest-weighted constraint. Alternatively, by using the model as a probability distribution over outputs, we would get variable outputs in the style of stochastic OT, but again

with more flexibility in constraint interaction than the systems currently employed in linguistics. The model would predict that one would get a passive output about 3/4 of the time and an active output 1/4 of the time for this configuration. Given actual data on how often actives and passives occur for various input feature specifications, we can use fitting algorithms to automatically find the weights w_i that best fit the data. We still get the effect of "emergence of the unmarked": if all outputs share the same constraint violations or satisfactions for highly weighted constraints, then the optimal candidate (under the first interpretation) or the candidate that is most commonly output (under the second interpretation) will be determined by low-ranked constraints on which the various output candidates differ. However, there has so far not been much research into how typological restrictions on the space of possible grammars can be built into loglinear models.

8.5.7 Generalized Linear Models

In the last couple of decades, there has been a revolution in the way statisticians standardly model categorical count/proportion data of the kind most commonly found in linguistics (i.e., under the assumption that one's categories are discrete). The traditional tools were some simple parametric and nonparametric test statistics, of which by far the best known is the (Pearson) chi-square (χ^2) test. These approaches sat as an underdeveloped sibling beside the sophisticated theory of linear models for continuous variables. Most people will have seen at least simple linear regression, analyzing a set of points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, as in figure 8.3. We would like to find the line

$$f(x) = mx + b$$

with parameters m and b that fits these points best (in the sense of minimizing the squared error indicated by the arrow lengths in the figure).

This can easily be extended to a multiple linear regression model, where the response variable (y_i) is predicted from a variety of explanatory x_{ij} variables, or even suitable functions of them. For example, we might have multiple linear regression models that look like one of these:

$$f(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$
$$f(y|x_1, x_2) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_3 x_2^2.$$

Indeed, such a multiple linear regression model was suggested for variable rules in Labov's early sociolinguistic work (Labov 1969):[41]

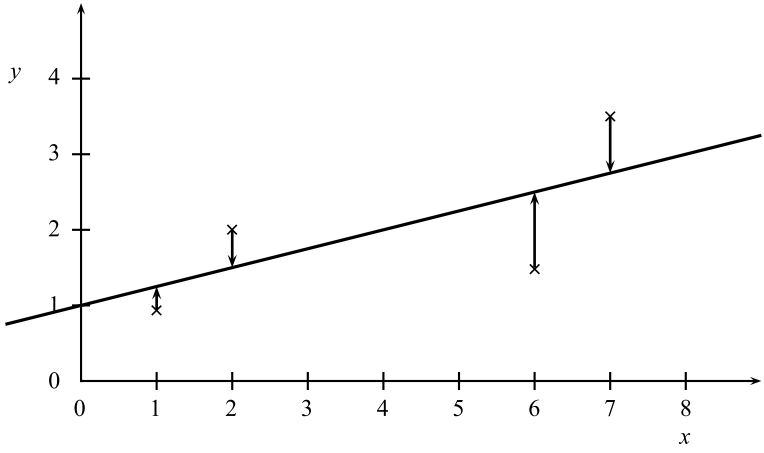

Figure 8.3
An example of linear regression. The line $y = 0.25x + 1$ is the best least-squares fit for the points $(1, 1)$, $(2, 2)$, $(6, 1.5)$, $(7, 3.5)$. Arrows show the y values for each x value given by the model.

$$p = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Here the x_i are indicator variables associated with a certain environmental feature, the β_i are weights, and p is the probability of a rule applying. However, this model is inappropriate when the aim is for the response variable to be a probability of rule application. Suppose there are three explanatory variables that individually strongly suggest a certain result, say with a probability of .8. If we set the corresponding β_i weights to around .8, the problem is that if all three of these features were present, then, when linearly combined, the value for the response variable would exceed one—an invalid value for a probability. Since this cannot be allowed, the best fit values in this model would be to set the β_i to a value around .3, which would mean that when only one of the features is present, the predictions of the model would be very poor (predicting only a 30% rather than 80% chance of the result). The immediate response to this was to consider multiplicative models of application and non-application (Cedergren and Sankoff 1974), but these have generally been replaced by logistic regression models (Rousseau and Sankoff 1978a; Sankoff 1988).

The logistic regression model is one common case of a *generalized linear model*, suitable for modeling binary response variables. The approach

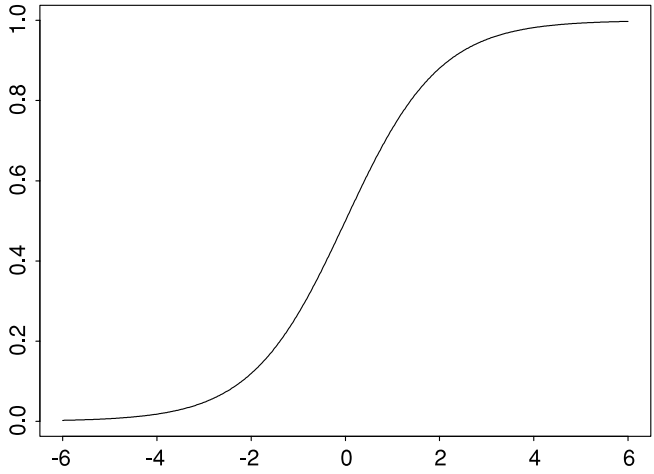

Figure 8.4
The logistic function

of generalized linear models has transformed the study of categorical data by bringing all the tools of traditional linear regression to the categorical case. This is done by generalizing the linear model by allowing the combined value of the explanatory variables to be equal to some function of the response variable, termed the *link function* g. So we have

$$g(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

For traditional linear regression (with a normally distributed response), g is the identity function. By choosing a log function for g, we obtain the loglinear models of the preceding section—a product of factors gives a sum of log factors. For a binary response variable, the appropriate link function is the logit function $\text{logit}(\pi) = \log[\pi/(1-\pi)]$. The inverse of the logit function is the logistic function. If $\text{logit}(\pi) = z$, then

$$\pi = \frac{e^z}{1+e^z}.$$

The logistic function will map any value of the right-hand side (z) to a proportion value between 0 and 1, as shown in figure 8.4. It is this logit model that has been the basis of VARBRUL sociolinguistic modeling since the late 1970s. An example relevant to the study in this section is Weiner and Labov's (1983) study of factors that predict active and passive in "agentless" sentences.

There is not space here for a detailed treatment of generalized linear models. There are many good books that cover these models in detail, though not always in a way approachable to linguists (Bishop, Fienberg, and Holland 1977 is the classic; Agresti 1990 and McCullagh and Nelder 1990 are standard references; Lloyd 1999 is more recent; Fienberg 1980 and Agresti 1996 are more approachable; Powers and Xie 1999 is directed toward social scientists; and some recent general statistical methods books such as Ramsey and Schafer 1997 give elementary coverage). However, since they are likely to continue to be so important to and useful in the development of probabilistic models over linguistic data, it is helpful to understand at least the main idea and how other models relate to them.

The logistic regression model is appropriate for binary response variables or proportions based on binomial counts—that is, when the number of trials is fixed. It is thus the natural model to use in the variable rules approach, where one is predicting the application or nonapplication of a particular rule. However, another very natural method is to simply collect a contingency table of counts for how often various linguistic tokens appear, organized into dimensions according to how they do or do not satisfy various constraints. For such a model, there is no natural limit to the counts in one cell, and the loglinear model, with a log link function, is appropriate. This model can be fit as a multinomial model by using iterative proportional fitting methods (Darroch and Ratcliff 1972) or general minimization techniques, such as conjugate gradient descent (Johnson et al. 1999).

Within a generalized linear model approach, an OT grammar is a special case, where the weight of each successive constraint is so much smaller than the weight of preceding constraints that there is no opportunity for ganging up—the highest-weighted differentiating feature always determines the outcome. The details differ slightly between the logistic and loglinear cases, but this connection has been noted multiple times. Rousseau and Sankoff (1978a, 66–67) discuss the possibility of this as a "whole hierarchy of successively weaker knockout features." Prince and Smolensky (1993, 200) point out that "Optimality Theory ... represents a very specialized kind of Harmonic Grammar, with exponential weighting of the constraints." Johnson (1998a) shows how standard OT (with a bound on the number of constraint violations) corresponds to loglinear models with the constraint weights being far enough apart that constraints do not interact. The last two are equivalent in that Harmonic

Grammar connectionist nets approximate loglinear models (Smolensky and Legendre 2000).

The example presented here is excessively simple (for expository purposes): there are just three constraints, designed to work over simple transitive sentences. I should therefore stress that these models can scale up. This has not been much demonstrated for linguistic goals, but within NLP, loglinear models with well over 100,000 features are regularly being deployed for tasks of parsing and disambiguation (Ratnaparkhi 1999; Toutanova and Manning 2000). Essentially, the models are based on *sufficient statistics*, which are counts of how often certain things happen or fail to happen in a sentence. These constraints can be any evaluable function, which can probe arbitrary aspects of sentence structure. For instance, Johnson et al. (1999) place features over grammatical relations, argument/adjunct status, low versus high attachment, and so on. Essentially, we are left with the customary linguistic task of finding the right constraints on linguistic goodness. Once we have them, they can be automatically weighted within a loglinear model.

8.6 Conclusion

There are many phenomena in syntax that cry out for noncategorical and probabilistic modeling and explanation. The opportunity to leave behind ill-fitting categorical assumptions and to better model probabilities of use in syntax is exciting. The existence of "soft" constraints within the variable output of an individual speaker, of exactly the same kind as the typological syntactic constraints found across languages, makes exploration of probabilistic grammar models compelling. One is not limited to simple surface representations: I have tried to outline how probabilistic models can be applied on top of one's favorite sophisticated linguistic representations. The frequency evidence needed for parameter estimation in probabilistic models requires much more data collection and much more careful evaluation and model building than traditional syntax, where one example can be the basis of a new theory, but the results can enrich linguistic theory by revealing the soft constraints at work in language use. This is an area ripe for exploration by the next generation of syntacticians.

Notes

My thanks to Bas Aarts, Farrell Ackerman, Harald Baayen, Rens Bod, Joan Bresnan, Ariel Cohen, Shipra Dingare, Mark Johnson, Dan Klein, Mike Maxwell, Simon Musgrave, John Paolillo, and Hidetosi Sirai for comments on a draft

of this chapter, or discussions of issues related to it. This chapter draws analyses and insight from joint research with Joan Bresnan and Shipra Dingare. In particular, although the presentation is different, the central example in the latter part of the chapter is drawn from Bresnan, Dingare, and Manning 2001. Taking linguistic usage seriously requires paying more attention to sociolinguistic variation than a syntactician traditionally did: I have found Chambers 1995 very helpful, and a few ideas and quotations are borrowed from that work. Parts of the research presented here were supported by the National Science Foundation under grant BCS-9818077.

1. For instance, Gleason 1961, a standard text of the period, devotes a chapter to a nontechnical introduction to Information Theory and suggests that it is likely to have a big impact on linguistics.

2. All italics in the cited examples are mine. Sources: (2a) *NYT* newswire, 1995/09/01 article by Dave Ahearn quoting Alan Greenspan; (2b) *NYT* newswire, 1995/11/29; (2c) John McHale, Economics 1415, Reform of the Public Sector, Fall 1999, ⟨http://icg.harvard.edu/~ec1415/lecture/lecturenote4.pdf⟩; (2d) Ed Zimmer, ⟨http://tenonline.org/art/9506.html⟩; (2e) State of California, Uniform Building Code, Title 24, Section 3306(r), ⟨http://www.johnsonite.com/techdata/section2/TITLE24C.HTM⟩; (2f) Ron Downing (Alaskan second-generation halibut fishing boat captain), March 10, 1998, ⟨http://www.alaska.net/~gusto/goodbye.txt⟩; (2g) Matthew Hindson (composer) catalogue notes, ⟨http://members.ozemail.com.au/~mhindson/catalogue/pnotes-pi.html⟩; (2h) *Sunday Business Times*, South Africa, ⟨http://www.btimes.co.za/99/0425/comp/comp04.htm⟩; (2i) Leigh Strope, Associated Press, 2000/11/07, ⟨http://www.theindependent.com/stories/110700/ele_turnout07.html⟩; (2j) Jeffrey M. Reisner, ⟨http://www.allcities.org/Articles/Ar_Reisner.htm⟩.

3. See also Fillmore 1992 for an amusing account of the strengths and weaknesses of corpus linguistics.

4. See, for instance, the research methods discussed in McEnery and Wilson 2001. For Windows computers, Mike Scott's Wordsmith Tools (⟨http://www.liv.ac.uk/~ms2928/wordsmith/⟩) is a leading recent example of a corpus linguistics tool based around concordances.

5. The parsed sentences in the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993) provide one useful source of data for complex syntactic queries (used, for example, by Wasow (1997) and Bresnan, Dingare, and Manning (2001)). But the tgrep access software supplied with some versions is limited to Unix and has not been actively maintained (though see ⟨http://www-2.cs.cmu.edu/~dr/Tgrep2/⟩ for an improved Tgrep2 program—also for Unix/Linux). The recently released ICE-GB corpus comes with a friendly graphical tool for Windows, ICECUP, which allows searching of parse trees (Wallis, Aarts, and Nelson 1999; Nelson, Wallis, and Aarts, in press). See ⟨http://nlp.stanford.edu/links/statnlp.html⟩ for a listing of corpora and NLP tools.

6. Recently, considerably more attention has been paid to the reliability of intuitive judgments and good methodologies for gathering them (Schütze 1996; Cowart 1997), but most linguists in practice pay little attention to this area.

7. In English, most verbs can be used either transitively or intransitively, and in introductory syntax classes we grope for examples of verbs that are purely transitive or intransitive, using verbs such as those in (3). Linguistic examples are standardly informally rated for exclusively syntactic well-formedness on the scale * (ungrammatical) > ?* > ?? > ? (questionable) > unmarked (grammatical). I use # to indicate a sentence that is somehow deviant, without distinguishing syntactic, semantic, or contextual factors.

8. It is certainly not my aim here to single out Pollard and Sag for reproach. I choose them as a source simply because they actually provide clear testable claims, as part of a commitment to broad observational adequacy. This has become increasingly rare in generative syntax, and I would not want to give the impression that another theory that assumes a categorical model of argument frames, but does not provide any examples of them, is superior.

9. The source for these and following examples is Linguistic Data Consortium *New York Times* newswire: (6a) 1994/08/04; (6b) 1994/12/07; (6c) 1994/10/03; (6d) 1995/01/16; (6e) 1995/08/04; (8a) 1995/12/15; (8b) 1996/05/15; (10) 1995/05/01; (12) 1994/09/07.

10. The exact numbers in this graph should be approached with caution, but I hope it is adequate to illustrate the general point. It would be necessary to do larger counts to get accurate probability estimates of rare patterns. Also, as is discussed by Roland and Jurafsky (1998), one should note that the frequency of different subcategorization patterns varies greatly with the genre of the text.

In this graph, forms with *as* followed by a verbal past/passive participle were collapsed with *as* plus adjective, and passive examples were grouped with the subcategorization for the corresponding active form.

11. Sampson (1987; 2001, chap. 11) also argues against a grammatical/ungrammatical distinction on the basis of a cline of commonness, though I think his evidence is less compelling (see Taylor, Grover, and Briscoe 1989; Culy 1998).

12. The *OALD* is one of several dictionaries that attempt to list subcategorization information (in a traditional form). Other sources of wide-coverage subcategorization information include resources constructed for NLP, such as COMLEX (Grishman, Macleod, and Meyers 1994). The discussion in this section expands the discussion of *retire* in Manning 1993.

13. All italics mine. These examples are from Linguistic Data Consortium *Wall Street Journal* newswire: (13a) 1994/11/23, (13b) 1994/11/20, (13c) 1987/04/07, (13d) 1986/12/08, (13e) 1994/12/12.

14. These are rough estimates from 1987 *WSJ*. Recall note 10—in particular, the extremely high frequency of transitive uses of *retire* would be unusual in many genres, but companies often retire debt in the *WSJ*.

15. Do these models estimate the entire frame jointly, or the parts of it independently? Some combination of these methods is generally best, since the joint distribution gives better information, when it can be estimated, but the sparseness of available data means that it is commonly better to work with the simpler but more robust independent estimates.

16. A semantic selection model of roughly the required sort has been explored in NLP models by Resnik (1996).

17. The judgments for these examples are quite murky and certainly wouldn't convince anyone who doesn't already believe in the argument/adjunct distinction. The distinctions are often sharper in nominalizations, which provide less freedom of ordering than verb phrases (e.g., contrast *Bill Clinton's retirement from politics in 2001* with ?*Bill Clinton's retirement in 2001 from politics*). See Radford 1988 for more convincing examples.

18. Ross's work "squishes" is another productive source of examples (Ross 1972, 1973a,b,c).

19. Of course, I would not want to suggest that human grammatical knowledge stipulates exactly this. Rather, it should follow from more general principles of information structure.

20. Contrast this expression with the language-understanding conditional probability from section 8.1.

21. The concept of "marginalizing out" in probability refers to removing a variable that we are not interested in, by summing (in the discrete case) a probability expression over all values of that variable. For instance, if there were just three contexts, c_1, c_2, and c_3, we could marginalize out the context by computing

$$P(f,m) = P(f,m,c_1) + P(f,m,c_2) + P(f,m,c_3) = \sum_{c_i} P(c_i)P(f,m|c_i).$$

Computationally efficient methods for manipulating individual variables and marginalizing out ones that are not of interest are at the heart of work on using Bayesian networks for reasoning (Jensen and Jensen 2001).

22. The Z term is needed for renormalization so that a probability distribution results. See the example of renormalization in section 8.5.6.

23. Though I should note that there has also also been considerable progress in categorical learning. A number of good books on statistical and machine learning are available (Bishop 1995; Ripley 1996; Mitchell 1997; Hastie, Tibshirani, and Friedman 2001), and conferences such as CoNLL (⟨http://ilk.kub.nl/~signll/⟩) focus on natural language learning.

24. See Bod 1998 for further evidence of human sensitivity to linguistic frequencies.

25. Concern about overly powerful learning methods is, admittedly, an issue that has been felt more keenly in linguistics than in most other areas, which have concentrated on finding sufficiently powerful learning methods. Smolensky (1999) attributes to Dave Rumelhart the notion that "linguists think backwards." Nevertheless, the goal of explaining the extensive but limited typological variation of human language is a real and valid one.

26. A couple of other examples appear in chapter 1 of Manning and Schütze 1999.

27. For example, "The wood being preserv'd dry will dure a very long time" (Evelyn 1664).

28. Another approach is to make progressively finer discrete subdivisions of word types, as is possible within a hierarchical lexicon approach (Malouf 2000). If the subdivisions are fine enough, it becomes difficult to argue against such an approach by linguistic means (since there are normally only a finite number of linguistic tests to employ). Indeed, such a model approaches a continuous model in the limit. Such an approach is not necessarily more insightful or constrained, though; indeed, I believe that locating words in a continuous space may be a more natural way to approach modeling variation and change.

29. Although, in practical NLP systems, people have commonly added some kind of heuristic repair strategy.

30. Below, I discuss the linking and person constraints. The presented discourse constraint is somewhat oversimplified, but nevertheless represents the essentials of another widespread phenomenon. Not only is it a soft constraint in English, but in many Philippine languages, such as Tagalog, there is a categorical specificity constraint on subject position, which resembles (but is more complex than) the discourse constraint given here (Kroeger 1993).

31. Here, adopting Optimality Theory notations, * indicates a constraint violation, ! means it is fatal, and the pointing hand indicates a valid form (which is unique for Optimality Theory, but might be many-fold for a categorical grammar).

32. Note that the Lummi pattern holds for bound pronouns; full pronouns designating speaker and hearer are focused and formally 3rd person expressions (Jelinek and Demers 1994, 714). If both arguments are 1st/2nd person, an active form is required.

33. Except in the trivial sense where all probabilities are 0 or 1.

34. Again, though, these are not "structural" zeroes representing impossible occurrences: an example of a sentence that would be counted in the top right cell was given in (32).

35. Previous studies also show evidence of person/voice interactions in English (Svartvik 1966; Estival and Myhill 1988). However, they only provide various marginals (e.g., counts of how many actives and passives there are in texts), which gives insufficient information to reconstruct the full joint distribution of all the variables of interest. Estival and Myhill (1988) *do* provide the kind of information needed for animacy and definiteness, but they provide person frequencies only for the patient role. We want to be able to predict the overall rate of the different systemic choices (Halliday 1994) that can be made for a certain input. That is, we want to know, for instance,

$P(form = passive | ag = 1, pt = 3)$.

We can determine the conditional frequencies needed from the full joint distribution (Bod, this volume).

36. It seems reasonable to put aside the short passives (without *by* phrases) as a separate systemic choice, and at any rate the person of the unexpressed agent cannot be automatically determined with certainty. If we include them, since they

overwhelmingly have 3rd person agents, the results are not materially affected: the person constraint shines through at least as clearly.

37. $\chi^2 = 116$, $p < .001$, though this test is not necessarily appropriate given the zero cell entries. The more appropriate Fisher's exact test gives $p < 10^{-8}$.

38. For example, famously Carroll, Bever, and Pollack (1981) showed that people's judgment of active versus passive sentences is affected by whether they are sitting in front of a mirror at the time.

39. Boersma (1997, 1998) and Boersma and Hayes (2001) assume normal curves with a fixed standard deviation of 2 (1 is more standardly used, but the choice is arbitrary). At 2 standard deviations apart (4 points on the scale), reranking occurs about 8% of the time. Note that according to this model, nothing is actually categorical; but if constraints are far enough apart, reversals in constraint ranking become vanishingly rare (perhaps one time in a billion a Lummi speaker does utter a sentence with a 3rd person subject and a local person object).

40. This is just an example grammar with my simple constraint system. See Bresnan, Dingare, and Manning 2001 for constructing actual grammars of the English passive using a more developed set of constraints.

41. Also, Keller (2000) uses multiple linear regression to model gradient grammaticality in his linear Optimality Theory model. (Noticing that speakers' magnitude estimation judgments are log-transformed, one could say Keller is using a loglinear model, as below. However, since the output values are arbitrary, rather than being probabilities, no normalization is necessary, and hence the fitting problem remains standard least squares, rather than requiring more complex techniques, as for either Harmonic Grammar or the loglinear models discussed above.)

This page intentionally left blank

Chapter 9

Probabilistic Approaches to Semantics

Ariel Cohen

9.1 Introduction

This chapter discusses the application of probability to issues in semantics. First, though, let's consider the devil's advocate warning. Probability, the devil's advocate would say, really has nothing to do with semantics. The probability calculus tells us how to calculate with probabilities, but not what such calculations *mean*. We know, for example, that if the probability that some event will occur is α, the probability that it will not occur is $1 - \alpha$. But what does it mean to say that the probability is α? What are the truth conditions of this statement? What ought the world to be like for it to be true? And how can we find out? The mathematical theory of probability is silent on these issues. Hence, the devil's advocate would say, probability judgments themselves are in need of a semantics. Therefore, saying that the meaning of a particular expression is a probability judgment gives us no insight into what it means, because the probability judgment itself is just as much in need of a semantics as the original expression is.

This argument has merit. Indeed, it probably accounts for the fact that there is less use of probability in semantics than perhaps there ought to be. There are, however, two types of response to this argument. One is that although the *mathematical* theory of probability provides no semantics for probability judgments, the *philosophical* theory of probability does. In fact, philosophers have been debating this very issue for centuries. They considered questions such as what it actually *means* to say that the probability that a coin will land "heads" is .5, or that the probability of getting cancer is higher if one smokes. There have been many proposals for the semantics of the term $P(A|B)$, that is, the conditional probability of A given B, or the probability for a B to be an A. If we accept one or another

of these solutions, we may use it to provide a semantics for probability judgments. As we will see, researchers have used different interpretations of probability to account for various phenomena.

Another response is applicable to cases where the use of probability in semantics is less direct. Rather than saying that probability judgments are the meaning of some expression, one can use a semantic account that is inspired by probability, that borrows some concepts and ways of thinking from it, without using it directly. We will discuss such use of probability later in this chapter.

This review is not meant to be exhaustive. The goal is to examine a number of case studies in some depth, considering to what extent probability can shed light on the problem under discussion.

In this chapter, I will assume familiarity with basic concepts in formal semantics. The reader who lacks this background may wish to consult one of the many textbooks on this topic (Allwood, Andersson, and Dahl 1977; Dowty, Wall, and Peters 1981; Gamut 1991; Heim and Kratzer 1998; de Swart 1998; Carpenter 1999; Chierchia and McConnell-Ginet 2000).

9.2 Generics and Frequency Adverbs

Perhaps the phenomenon that has received the greatest number of probabilistic accounts (as well as many nonprobabilistic ones) is that of generics and frequency statements.[1] These constructions occur frequently in natural language. Much of our knowledge about the world is expressed using such sentences—a glance at an encyclopedia will easily provide myriads of examples. Yet it is far from clear what such sentences mean, that is, what a given sentence entails, what it presupposes, and what it is that makes it true or false.

Perhaps the most puzzling fact about these sentences is that they are, in a sense, both very strong and very weak. On the one hand, generics and frequency statements are stronger than simple quantified statements, in being *lawlike*. On the other hand, they are weak: they are contingent on properties of the actual world and (except, perhaps, for *always*) are even weaker than universal statements, since they allow for exceptions.

The amount of work on this topic is quite substantial (see Carlson and Pelletier 1995, Cohen 1996, and Pelletier and Asher 1997 for overviews and references). In this chapter, I will only consider accounts that make use of probability.

The idea to be considered is that generics and frequency statements make probability judgments. Thus, it seems plausible that for (1) to be true, the probability that an arbitrary bird flies must be one:

(1) Birds always fly.

Similarly, *never* would require this probability to be zero, and *sometimes* would require it to be nonzero. It is less clear what the probability ought to be for adverbs such as *usually*, or *often*, and even less clear what it would be for generics such as (2):

(2) Birds fly.

But whatever the exact value is, it is attractive to consider such sentences, too, as expressions of probability judgments.

A number of authors have attempted to account for generics and frequency statements in terms of probability, or using concepts that arise from an analysis of probability.

It turns out that different theories of generics assume different interpretations of probability. Let us consider them in turn.

9.2.1 Ratio Theories

An old idea, which can be traced as far back as Galileo (see Hacking 1975), is that probability judgments express ratios of sets of individuals. $P(A|B)$ is the ratio of the set of individuals that satisfy both A and B to the set of those that satisfy B. Hence, the probability that a fair coin comes up "heads" would be .5 just in case in half of all observed tosses, the coin came up "heads."

Åquist, Hoepelman, and Rohrer (1980) have applied this conception of probability to the semantics of frequency adverbs. Frequency adverbs, according to this study, are relations between sets. For example, (1) is a relation between the set of birds and the set of flying objects, and its logical form would be as shown in (3):

(3) **always**(λx**bird**(x), λx**fly**(x))

The sets need not be sets of individuals. They may also be sets of tuples of individuals, or tuples of individuals and times (and presumably events as well). A classic example is (4a), discussed by Lewis (1975). Its logical form, according to Åquist, Hoepelman, and Rohrer, would be something like (4b):

(4) a. Riders on the Thirteenth Avenue line seldom find seats.
b. **seldom**($\lambda x \lambda t$(**person**(x) $\wedge$ **time**(t) $\wedge$ **rider**(x, t)),
$\lambda x \lambda t$(**person**(x) $\wedge$ **time**(t) $\wedge$ **find-seat**(x, t)))

The truth conditions Åquist, Hoepelman, and Rohrer provide for such adverbs of frequency are probabilistic. Specifically:

- **always**(B, A) is true iff $P(A|B) = 1$
- **very-often**(B, A) is true iff $P(A|B) \geq .9$
- **often**(B, A) is true iff $P(A|B) \geq .7$
- **fairly-often**(B, A) is true iff $P(A|B) > .5$
- **fairly-seldom**(B, A) is true iff $P(A|B) < .5$
- **seldom**(B, A) is true iff $P(A|B) \leq .3$
- **very-seldom**(B, A) is true iff $P(A|B) < .1$
- **never**(B, A) is true iff $P(A|B) = 0$
- **sometimes**(B, A) is true iff $P(A|B) \neq 0$

Though they do not discuss generics, presumably some suitable value of probability may be found for generics as well.

The heart of Åquist, Hoepelman, and Rohrer's proposal lies in their definition of probability. This is done in terms of ratios. That is to say, A and B are taken to be sets, and $P(A|B) =_{\text{def}} |A \cap B|/|B|$.

9.2.2 Logical Relation Theories

The idea underlying the logical relation interpretation of probability, originating with Carnap (1950), is that $P(A|B)$ is the ratio of possible worlds in which both A and B hold to those where B holds.[2] The probability that a coin comes up "heads," then, is the ratio of worlds in which it comes up "heads" to worlds in which it is tossed. Of course, in order to make sense of the idea of ratios of possible worlds, one needs to define some measure function on worlds. An appropriate measure function needs to assign a real number to each world, in such a way that the sum of the measures on all worlds is finite (typically one). A variety of such functions have been proposed by Carnap and his followers.

Two kinds of logical relation theories have been applied to generics.

9.2.2.1 Probability Distribution over Worlds Schubert and Pelletier (1989) propose an account of the meaning of generics that makes use of a logical relation theory of probability. I should emphasize that probability plays a relatively minor role in their account, whose emphasis is on the dynamic behavior of generics. For our purposes, the intricacies of their

dynamic logic are not relevant, and I will therefore explain the spirit, rather than the letter, of their ideas. Moreover, while the authors reserve their probabilistic truth conditions for generics only, it seems that they could be easily applied to frequency statements too.

According to Schubert and Pelletier's interpretation, (2) is true just in case in "most" pairs of possible worlds and birds, the bird flies in that world, "where 'most' is to be interpreted in terms of some probability distribution favouring worlds w' similar to w [the world of evaluation, usually the actual world] with regard to the 'inherent' or 'essential' nature of things" (pp. 259–260).

Schubert and Pelletier admit that they leave open exactly what this means; we will be able to say a little more about the implications of this theory after we consider another logical relation theory of probability.

9.2.2.2 Normality A different application of a logical relation interpretation of probability to genericity has been proposed by a number of scholars (Delgrande 1987; Morreau 1992; Asher and Morreau 1995; Krifka 1995; Pelletier and Asher 1997, to name just a few). These accounts make use of Kratzer's (1981) theory of probability judgments. Kratzer herself does not deal with generics; her main concern is with qualitative probability judgments, such as these:

(5) a. There is a good possibility that Gauzner-Michl was the murderer.
b. There is, however, still a slight possibility that Kastenjakl was the murderer.
c. Gauzner-Michl is more likely to be the murderer than Kastenjakl.
d. It is probable that Gauzner-Michl was the murderer.

In her treatment of such sentences, Kratzer proposes that they are modal statements, similar to statements of necessity and possibility. Modal statements, according to Kratzer, are evaluated with respect to two components: a modal base and an ordering source.

The *modal base* is the set of accessible worlds, and it is provided by the context. The role of the modal base is to indicate the type of modality: logical, physical, epistemic, deontic, and so on. In a sense, the modal base provides the definition of what is considered possible: something may be considered possible because it is logically possible, because it is allowed by the laws of nature, because it is possible as far as the speaker knows, because it is permissible, and so on, and the modal base provides this criterion for possibility.

In addition to the modal base, an *ordering source* is defined over worlds. The intuitive idea behind this ordering source is that if $w_1 \leq w_2$, then w_2 is closer to the ideal, or more "normal" than w_1. In the ideal worlds, everything is as it should be, and there is no unexpected turn of events. Such a world is, as Kratzer puts it, a "complete bore."

Using the machinery of the modal base and the ordering source, Kratzer provides an account of the truth conditions of the sentences in (5). Sentence (5d) expresses what she calls *human necessity*. She proposes that a proposition ϕ is humanly necessary iff it is true in all worlds closest to the ideal. More formally, if W is the modal base, then for all $w \in W$ there is $w' \in W$ such that

1. $w \leq w'$, and
2. for all $w'' \in W$, if $w' \leq w''$ then ϕ is true in w''.

Thus, (5d) is true iff in all worlds in which events turned out in the normal, expected way, Gauzner-Michl was the murderer.

Sentence (5a) expresses what Kratzer calls *human possibility*. A proposition ϕ is a human possibility iff $\neg\phi$ is not a human necessity. Thus, (5a) is true just in case it is not probable that Gauzner-Michl was not the murderer.

Sentence (5b) is an example of a *slight possibility*. A proposition ϕ is slightly possible iff

1. there is at least one $w \in W$ in which ϕ is true, and
2. $\neg\phi$ is a human necessity.

Thus, (5b) means that it is probable that Kastenjakl was not the murderer, but it is still possible that he was.

Comparative possibility is expressed by (5c). ϕ_1 is more possible than ϕ_2 iff for every world in which ϕ_2 holds, there is a world where ϕ_1 holds that is at least as normal, but there is a world where ϕ_1 holds for which there is no world at least as normal in which ϕ_2 holds. More formally,

1. for all $w \in W$, if ϕ_2 holds in w then there is $w \leq w'$ such that ϕ_1 holds in w', and
2. there is $w \in W$ such that ϕ_1 holds in w and for no $w \leq w'$, ϕ_2 holds in w'.

Thus, the truth of (5c) entails that for every world in which Kastenjakl is the murderer, there is a world at least as normal where Gauzner-Michl is; but there is at least one world w where Gauzner-Michl is the murderer,

and in all the worlds that are at least as normal as *w*, Kastenjakl is not the murderer.

Kratzer's theory has been applied to generics using the idea that generics express human necessity. Thus, for example, (2) is true just in case in the most normal worlds, all birds fly.

Note that, whatever merit this idea has with respect to generics, it cannot be applied to frequency adverbs. This is because Kratzer's technique cannot apply to arbitrary probability judgments, but only to a restricted class of them, exemplified by the sentences in (5). In particular, quantitative judgments, which are arguably necessary for an account of frequency adverbs, are not handled by Kratzer's account.

9.2.3 Relative Frequency Theories

A view that goes back to Poisson (1837) is that the probability judgment $P(A|B)$ expresses a statement of limiting relative frequency, that is, the mathematical limit of the proportion of *B*s that are *A*s as the number of *B*s approaches infinity.

There are a number of ways to formalize frequency theories mathematically (see, e.g., Reichenbach 1949; von Mises 1957; van Fraassen 1980), but the underlying idea is simple. If we want to know how likely smokers are to get lung cancer, we count the number of cancer patients among smokers and divide by the total number of smokers in our sample. We do this for large samples, over long periods of time. The larger the sample and the longer the duration of the study, the closer the ratio will get to the desired probability. The limit of this ratio as the sample size approaches infinity *is* the probability. Similarly, the limit of the frequency of "heads" in an infinite sequence of coin tosses, according to this view, is exactly the probability of "heads," namely, .5 (if the coin is fair). One conclusion of the frequency interpretation is that it is impossible for a fair coin to come up "heads" on each and every one of an infinite sequence of coin tosses; sooner or later, it is felt, the coin will come up "tails," and, in the long run, both outcomes will have occurred with equal frequency. Of course, we cannot actually examine an infinite number of smokers; but we can extrapolate from a finite sequence of examinations to what the limit might be. The longer the actual sequence is, the more confidence we should have in the correctness of the extrapolation.

It should be noted that most philosophers do not consider relative frequency theories to be an adequate interpretation of the general notion of probability. However, what we are concerned with here is to investigate

the adequacy of such theories for the specific probability judgments expressed by generics and frequency statements.

In Cohen 1999, I propose such an account of generics and frequency statements. The intuitive idea is that sentences such as (1) and (2) are evaluated with respect to infinite sequences of birds. Such sentences are evaluated in models with branching time, so that for any given time there is more than one possible future. There is a course of time where the world is destroyed in the year 3000, there is a course of time where you become the President of Dalmatia, there is a course of time where I never finish writing this chapter, and so on. Following Thomason (1970), each linear course of time is called a *history*. Sentences (1) and (2) are evaluated with respect to histories that admit infinite sequences of birds. Sentence (1) is true iff in every *admissible* such sequence, the limit of relative frequency of flying birds among birds is 1; (2) is true iff this limit is greater than .5.

Note that if every possible history were admissible, this theory of probability would turn out to be equivalent to a logical relation theory, since every history could be seen as a possible world. However, this is not really the case.

Since it is impossible to observe infinitely long sequences in the actual world, these must be *extrapolated* from the actual history. For this extrapolation to be of use, we need to assume that the observed instances provide a good statistical sample. That is to say, we need to assume that the relative frequency over the sample we do have is close to the value of the probability, that is, the relative frequency over infinitely long sequences. In order for us to believe this, any sequence we consider ought to be such that any sufficiently large sample taken from it is a good sample. Then, if our sample is sufficiently large, we can extrapolate from it with some confidence. Thus, the admissible histories are such that the relative frequency of flying birds over any sufficiently long subsequence will be close to that over the sequence as a whole. In particular, (the relevant part of) the actual history must be part of any admissible history, and the relative frequency over the actual history must be close to that of the admissible history as a whole.

What sort of thing is an admissible history, then? With respect to a generic or a frequency statement $Q(\psi, \phi)$, it is a history where the proportion of ϕs among ψs remains pretty much the same. With respect to (1) and (2), for example, an admissible history is one where the proportion of flying birds among birds is pretty much constant. There may be

relatively brief fluctuations, but, on the whole, there will not be any significantly long intervals of time where the proportion of flying birds changes drastically, and there will not be prolonged discontinuities. Thus, a history in which birds suddenly lose the faculty of flight will *not* be admissible. And since admissible histories must contain the actual history, the limiting relative frequency of flying birds in all admissible infinite sequences will be close to that in the actual world.

To take another example, let us consider (6):

(6) John (often) jogs in the park.

Here, an admissible history is one in which John's jogging in the park continues with pretty much the same frequency. It is possible that for a few days in a row he might be ill and not jog, or that during some week or other he might feel very energetic and jog every day. But, on the whole, his jogging frequency should remain pretty much constant throughout the history. A history in which John jogs furiously in the park just before summer in order to lose weight, and then stays idle the rest of the year, will not be admissible. The reason is that there would be a sufficiently long time interval where the frequency of John's jogging is very high, and another sufficiently long time interval where that frequency is very low.

According to Salmon (1977), a reference class B is homogeneous iff for every partition of B, and every $B' \subseteq B$ induced by the partition, $P(A|B)$ is equal to $P(A|B')$. If we relax Salmon's definition so that we require only that the probabilities be roughly the same, and only consider partitions along the temporal dimension, it follows that the domain of generics and frequency adverbs is homogeneous. This is because they are evaluated with respect to admissible histories, so that for every sufficiently long interval, the relative frequency is close to that of the limiting relative frequency over the history as a whole.

It is obviously possible to partition a domain in other ways, not simply according to time. Do other partitions play a role in the evaluation of generics and frequency statements?

The answer to this question, it turns out, is yes. While frequency adverbs require their domain to be homogeneous only with respect to the time partition, generics require homogeneity with respect to a great number of other partitions as well. This notion seems to correspond rather well to the pretheoretical notion of what a generic sentence means. Lowe (1991), for example, considers the following sentence:

(7) Fido chases cats.

He writes:

> The sort of empirical evidence which *would* support or undermine the truth of [(7)] is a matter that is broadly familiar to us all (though it is by no means a *simple* matter). If Fido is found to chase a good many different individual cats, of varying sizes, colours, ages and temperaments, then, *ceteris paribus*, we shall consider [(7)] to be empirically well supported; if not, then not. (p. 295; original emphases)

Lowe's observation makes intuitive sense. Note that what this means is that it is not sufficient for Fido to chase many cats, but that he should chase cats of many varieties. That is to say, the domain of the generic has to be homogeneous with respect to many partitions, depending on cats' sizes, colors, and so on.

What exactly the partitions are with respect to which a generic requires homogeneity is a matter for further research, and will probably require a consideration of cognitive factors,[3] but one can get an idea from examples such as these:

(8) a. Israelis live on the coastal plain.
    b. People are over three years old. (Henk Zeevat, personal communication)
    c. Primary school teachers are female.

The majority of Israelis live on the coastal plain, yet (8a) is ruled out, because if the domain of Israelis is partitioned according to the geographical regions in which they live, there will obviously be subsets of this domain whose members do not live on the coastal plain (e.g., those who live in Jerusalem or Beer Sheva). Hence, the domain is not homogeneous with respect to the property of living on the coastal plain, and (8a) is ruled out.

The majority of people are clearly over three years old; yet (8b) is bad. If we partition people according to their ages, there will be some subsets of people (e.g., babies) such that the probability that their members are over three years old is zero. Hence, the domain of the generic quantifier is not homogeneous, and sentence (8b) is ruled out.

Although the majority of primary school teachers are female, (8c) is odd. The reason is that, if we partition the set of teachers according to their gender, there will obviously be a set, the probability of whose members to be female is zero—the set of male teachers. Therefore, the set of teachers is not homogeneous with respect to the property of being female.

In contrast to generics, if the adverb *usually* is inserted into the sentences in (8), they become acceptable, in fact true:

(9) a. Israelis usually live on the coastal plain.
b. People are usually over three years old.
c. Primary school teachers are usually female.

These facts demonstrate that frequency adverbs, unlike generics, only require that their domain be homogeneous with respect to the time partition.

9.2.4 Evaluating Interpretations of Probability

The question is, of course, which, if any, of the interpretations of probability should we choose? L. J. Cohen (1989) argues that different interpretations of probability are appropriate for different types of probability judgment. Thus, it is possible that some interpretation of probability will be appropriate for the semantics of generics and frequency adverbs, whereas another interpretation may be appropriate for a different phenomenon.

How can we choose an interpretation of probability suitable for a given application? L. J. Cohen characterizes probability judgments using four parameters and points out that different interpretations are appropriate for different settings of these parameters. Let us evaluate the theories discussed in this section with respect to these parameters.

9.2.4.1 Necessity versus Contingency The first parameter indicates whether a probability judgment expresses a necessary or contingent statement. In general, generics and frequency statements are true or false contingently; they may be true in the real world, yet false in other worlds, and vice versa. It just happens to be the case that, in the actual world, (10) is true, but it might have been otherwise:

(10) Birds (generally) fly.

This is quite compatible with a ratio or relative frequency theory, since there is no necessity for a ratio or a relative frequency to have a certain value.

Things are more complicated with logical relation theories. These theories were originally developed to account for logically necessary probability judgments (specifically, the relation between evidence and

conclusion). Hence, they face a problem when applied to the analysis of generics and frequency adverbs.

Schubert and Pelletier (1989) offer to solve this problem, as we have seen, by assuming that the probability distribution is sensitive to inherent or essential properties of the actual world. Thus, the effect of other worlds on the truth conditions is negligible. If an eventual explication of this notion allows us to understand "inherent" and "essential" properties as contingent (though, perhaps, not accidental) properties, we would have an account of the contingency of generics and frequency statements: they are evaluated with respect to worlds that are similar to the actual world in terms of contingent properties.

Similarly, if normality is interpreted as a contingent rather than a necessary property, normality-based approaches can account for the contingency of generics and frequency statements.

9.2.4.2 Propositions versus Properties The second parameter indicates whether, in the probability judgment $P(A|B)$, the terms A and B stand for "whole sentences, expressing fully determinate propositions, or just sentence-schemata or parts of sentences" (L. J. Cohen 1989, 84). I take this parameter to indicate a distinction between probability judgments relating propositions to those relating properties.

Generics and frequency statements relate properties; (1) and (2) relate the property of flying to the property of being a bird, rather than relating, say, the proposition that Tweety flies to the proposition that Tweety is a bird.

Ratio and frequency theories require A and B to designate properties, since they are concerned with the ratio or the frequency with which individuals that satisfy B also satisfy A. Therefore, they satisfy the second parameter with respect to generics and frequency adverbs.

Logical relation theories, on the other hand, are problematic. Bacchus (1990) claims that they are not appropriate for dealing with generics and frequency statements.[4] The reason is that generics and frequency statements relate properties, not propositions; but only propositions, not properties, can be said to hold in a world.

Take, again, (2). According to logical relation theories, (2) means that some proposition has a high probability, relative to some probability distribution over possible worlds. But what would such a proposition be? One possibility that immediately suggests itself is (11):

(11) $\forall x(\mathbf{bird}(x) \rightarrow \mathbf{fly}(x))$

But this is much too strong a requirement. For (11) to have a high probability, it is required that it be true in "most" possible worlds. This means that in those worlds there will be no nonflying birds, which, in turn, means that the probability for the claim that there are nonflying birds is low. However, the truth of (2) does not require that we assign low probability to the statement that there are nonflying birds; one can readily accept the existence of nonflying birds and still consistently believe (2). Bacchus shows that other alternatives to (11) are also inadequate and concludes that logical relation theories of probability are not appropriate for the type of probability judgments exemplified by (2).

Schubert and Pelletier (1989) amend the logical relation theory so as to solve this problem. As we have seen, (2) expresses quantification over pairs of possible worlds and birds, not just possible worlds. In general, the probability distribution is defined not simply over possible worlds, but over pairs of possible worlds and *contexts*. For our purposes, we can treat contexts simply as assignment functions. Then, for example, (2) is true just in case for most pairs $\langle w, g\rangle$ (where w is a world and g an assignment function) such that $[\![\mathbf{bird}(x)]\!]^{w,g} = 1$, it holds that $[\![\mathbf{fly}(x)]\!]^{w,g} = 1$. This has the effect that for most pairs of birds and worlds, the bird flies in that world.

Thus, in effect, Schubert and Pelletier propose an account of probability based on possible worlds, while maintaining that generics and frequency adverbs relate properties, not propositions. While a property cannot be given a truth value with respect to a given world, it can be given a truth value with respect to a world and an assignment function.

The application of Kratzer's (1981) theory to generics (e.g., the claim that (2) is true just in case (11) is true in the most normal worlds) is also open to Bacchus's objection. Pelletier and Asher (1997) attempt to solve this problem by making normality relative to an individual and a property (and a world). For example, (2) is true iff, for every individual bird b, in all worlds that are most normal with respect to b and the property of being a bird, b flies. Other birds in these worlds may fail to fly, thus enabling us to assign human necessity to the statement that some birds do not fly. Thus, for each bird there might be a different set of normal worlds. Pelletier and Asher do not, however, suggest any way of determining these normal worlds.

9.2.4.3 Substitutivity The third and perhaps crucial parameter involves the conditions under which other terms can be substituted for A or B without changing the probability $P(A|B)$. Suppose we have two extensionally equivalent properties, B_1 and B_2. That is to say, the set of individuals that satisfy B_1, at this moment in time and in the actual world, is identical to the set of individuals that satisfy B_2. If generics and frequency adverbs behave extensionally, we would expect $Q(B_1, A)$ and $Q(B_2, A)$ to have the same truth conditions for every property A and frequency adverb Q.

It seems that this does not hold in general. Consider the following example (from Carlson 1989):

(12) A computer (always) computes the daily weather forecast.

Carlson observes that

> "the daily weather forecast" requires an *intensional* interpretation, where its meaning cannot be taken as rigidly referring to the present weather forecast, e.g. the one appearing in today's copy of the *Times* predicting light rain and highs in the upper thirties. (p. 179; emphasis added)

Suppose, for example, that today's weather forecast predicts a severe blizzard and is consequently the main news item. So the terms *the daily weather forecast* and *the main news item* have the same extension today, yet (13), when interpreted generically, would be false:

(13) A computer (always) computes the main news item.

The intensionality exhibited here—it is important to note—is with respect to the time index, but not with respect to possible worlds.[5] Suppose that the weather report is John's favorite newspaper feature. Then (14) would have the same truth conditions as (12), although there are any number of worlds where John never so much as glances at the daily weather forecast:

(14) A computer (always) computes John's favorite newspaper feature.

I should make it clear what I am not saying here. I am definitely not claiming that the property of being the daily weather report, or being John's favorite newspaper feature, has the same extension in all possible worlds; clearly, in different possible worlds, there may be different weather conditions, and John may have different preferences. What I *am* claiming is that the truth conditions of a generic sentence relating two properties do not depend on their extensions in any other world but the

real one, though they do depend on the extensions of the properties at different times.

To give another example, suppose that John fears all bats but no other animal. The set of bats is equivalent to the set of animals John fears, although the intensions of the respective terms differ; there are any number of possible worlds where John feels nothing but love toward bats. However, it seems that a generic or frequency statement involving bats has the same truth conditions as a similar sentence about animals John fears:

(15) a. Bats (usually) fly.
b. Animals that John fears (usually) fly.

Similarly, there is no logical necessity for the whale to be the largest animal on Earth, or for the quetzal to be Guatemala's national bird; yet (16a) and (16b) have the same truth conditions as (17a) and (17b), respectively:

(16) a. The whale suckles its young.
b. The quetzal has a magnificent, golden-green tail.

(17) a. The largest animal on Earth suckles its young.
b. Guatemala's national bird has a magnificent, golden-green tail.

Generics and frequency adverbs, then, are parametric on time, but not on possible worlds; if two properties have the same extension throughout time, they can be freely exchanged in a generic or a frequency sentence *salva veritate*.

I should emphasize that whether or not two properties have the same extension in the actual world may depend on other worlds, if the properties contain intensional contexts. The following examples are due to Rich Thomason (personal communication):

(18) a. Americans know that Indiana is in the Midwest.
b. Americans know that Guanabara is in Brazil.

The propositions that Indiana is in the Midwest and that Guanabara is in Brazil have the same extension in the actual world—both are true; yet (18a) may be true and (18b) false, or vice versa. The reason is that *know* is an intensional verb, and therefore the properties of knowing that Indiana is in the Midwest and knowing that Guanabara is in Brazil do not necessarily have the same extension *in the actual world*, since one may know one fact without knowing the other.

Given that generics and frequency adverbs are parametric on time but not possible worlds, an account of probability is needed that satisfies this requirement—in other words, where different descriptions of the same timeless property can be interchanged without a change in truth value.

Theories that regard probabilities as ratios of sets allow any terms that are extensionally equivalent at the present time to be substituted for each other without changing the truth value, since sets are purely extensional categories. Hence, they wrongly predict that generics and frequency adverbs are completely extensional.

According to logical relation theories, whether or not two terms can be substituted for one another depends on whether they have the same extension in all possible worlds that are similar to the actual world (Schubert and Pelletier 1989) or in all possible worlds that are in the modal base (Kratzer 1981). Thus, such theories predict that generics and frequency adverbs are parametric on possible worlds. This is, again, an incorrect prediction.

The relative frequency view is the one that successfully predicts the behavior of generics. Generics and frequency statements are parametric on time because they are evaluated with respect to courses of time—histories.

However, they are not parametric on possible worlds, because the histories are restricted to admissible ones. There are any number of worlds where the whale is not the largest animal on Earth and where the quetzal is not Guatemala's national bird, and correspondingly there are any number of histories where a larger animal than the whale evolves and where Guatemala decides to change its national bird. But such histories will not be admissible. In order for them to be admissible, they would have to continue the actual history; but in the actual history the whale *is* the largest animal on Earth and the quetzal *is* Guatemala's national bird, so these histories would fail to be homogeneous. Only histories in which things happen pretty much the way they occur in the actual world will be admissible; hence, generics and frequency statements are not parametric on possible worlds.

9.2.4.4 Extensibility The fourth parameter indicates whether or not the probability judgment $P(A|B)$ would remain the same if the number of Bs were greater than it actually is.[6] Suppose that,

> on the assumption that he is a forty-year-old asbestos worker, a man has a .8 probability of death before the age of sixty. . . . [T]he same judgment of probability would presumably hold good even if the number of asbestos workers were larger

than it in fact is. Only thus might someone legitimately infer a reason for not going to work in an asbestos factory. But to accept a .6 probability, for any person picked out at random at a certain conference, that he is staying in the Hotel Excelsior is not to accept a probability that could be relied on to have held good if more people had been at the conference. Perhaps the additional participants would all have had to stay elsewhere because the Hotel Excelsior had no more unoccupied rooms. (L. J. Cohen 1989, 89)

In this sense, generics and frequency adverbs are extensible; (1) and (2) would keep their respective truth values even if there were more birds than there actually are.

Theories regarding probabilities as ratios of sets are not applicable to extensible judgments, since changing the number of instances might very well change the ratio.

Logical relation theories, on the other hand, would pass this test, since they take probability judgments to be about possible worlds, rather than the actual world; thus, if the number of birds in the actual world were different, this would not necessarily change the truth values of (1) and (2).

Relative frequency theories also predict the extensibility of generics and frequency adverbs, since they take probability to be evaluated with respect to sequences that already contain infinitely many *B*s.

9.2.4.5 Summary Ratio theories predict the facts that generics and frequency adverbs are contingent and relate properties; however, they erroneously predict them to be fully extensional and cannot account for their extensibility.

Logical relation theories account for the extensibility of generics and frequency adverbs and can, with some modifications, account for the fact that they relate properties. However, explaining the contingency of these constructions is not easy, and these theories erroneously predict that generics and frequency adverbs are fully intensional.

Relative frequency theories account for the behavior of generics and frequency adverbs with respect to all four parameters: they are correctly predicted to be contingent, to relate properties, to be extensible, and to be parametric on time but not on possible worlds.

9.3 Conditionals

We have seen that generics and frequency adverbs are extensible. This means that if (19) is true, then, if there were more birds than there actually are, they would probably fly:

(19) Birds (usually) fly.

It follows that if something that is not currently a bird were a bird, it would probably fly. In other words, (19) entails the conditional (20):

(20) If Dumbo were a bird, he would probably fly.

One of the nagging problems for any truth-conditional semantics of natural language is an account of conditionals. What does it mean to say that *If A then B*? The truth or falsity of a conditional appears to be more than simply a function of the truth values of *A* and *B*. Surely, the meaning of conditionals is not simply the material conditional, or else all counterfactual conditionals would be trivially true.

An attractive idea, one that goes back to Stalnaker 1970, is to use conditional probabilities in the analysis of conditionals.[7] The idea is the following. We know that in a given situation, a sentence such as (21) is either true or false:

(21) It is raining.

We can meaningfully talk about the probability that (21) is true. Indeed, this is presumably what weather forecasters do. So speakers may assign certain probabilities to sentences. The probabilities of complex sentences may be dependent on, or constrained by, the probabilities of their parts. For example, our semantics ought to make sure that the probability of (22a) is one minus the probability of (21), and that the probability of (22b) is not greater than the probability of (21):

(22) a. It is not raining.
b. It is raining and it is cold.

Indeed, the semantics of negation and conjunction satisfy these requirements.

The idea is to find a definition of the truth conditions of the conditional *If A then B* such that its probability will be the conditional probability $P(B|A)$. Thus, for example, we seek a semantics for (23a) such that its probability will be (23b):

(23) a. If it rains tomorrow, the track will be muddy.
b. P(the track is muddy|it rains)

On the face of it, this seems almost obvious. As van Fraassen (1976) puts it:

[T]he English statement of a conditional probability sounds exactly like that of the probability of a conditional. What is the probability that I throw a six if I throw an even number, if not the probability that: if I throw an even number, it will be a six? (pp. 272–273)

The rather surprising fact of the matter is that such a definition cannot be found, as proved by Lewis (1976). More specifically, Lewis has shown that, given some basic constraints on probability functions, only probability functions that have at most four values will satisfy such a definition.

Some researchers have taken Lewis's negative results (and other, similar results; see Hájek and Hall 1994) to show that probability can play no role in the explication of the semantics of conditionals. However, all is not really lost. One possibility is to devise a system where the conditional probability $P(B|A)$ is equal, not to the probability of the conditional *If A then B*, but to some other value associated with this conditional and related to probability.

One such approach has been developed by Stalnaker and Jeffrey (1994). The idea is to treat all propositions as random variables. For propositions that do not involve conditionals, the corresponding random variable has only two possible values: 1 if the proposition is true, 0 if the proposition is false. Now take the conditional *If A then B*, written as $A > B$. If A is true and B is true, the value of the corresponding random variable is 1; that is, the conditional is true. If A is true and B is false, the conditional is false, so the value of the corresponding random variable is 0. But what if A is false? In this case, Stalnaker and Jeffrey suggest that the value of the random variable is $P(B|A)$. So, conditionals are propositions that may receive not only the values 1 and 0, but also some number between these extremes.

Since the value of a conditional is a random variable, it makes sense to talk about the expected value of this variable:

$$E(A > B) = 1 \times P(A \wedge B) + 0 \times P(A \wedge \neg B) + P(B|A) \times P(\neg A). \quad (24)$$

Stalnaker and Jeffrey show that for many important cases of conditionals, this expectation is equal to the conditional probability:

$$E(A > B) = P(B|A). \quad (25)$$

Take the following example (due to Edgington 1991). Suppose $P(A) = .7$ and $P(A \wedge B) = .56$. Then $P(B|A) = .8$, and $P(A \wedge \neg B) = P(A) - P(A \wedge B) = .14$. Now $A > B$ is 1 if $A \wedge B$, 0 if $A \wedge \neg B$ and .8 otherwise. So,

$$E(A > B) = 1 \times .56 + 0 \times .14 + .8 \times .3 = .8 = P(B|A). \quad (26)$$

So, indeed, the expected value of $A > B$ is equal to the conditional probability of B given A.

While theories such as that of Stalnaker and Jeffrey may have interesting technical and philosophical implications, it is not clear that they correspond to our intuitions about the meaning of conditionals. In particular, things become less intuitive when embedded conditionals are considered. Here is an example, due to Edgington (1991; see also Lance 1991):

Consider the conditional (27a), represented schematically as (27b):

(27) a. If the match is wet, then if you strike it it will light.
b. $W > (S > L)$

What is its expected value? According to the definition, we want it to be $P(S > L|W)$. Intuitively, this probability ought to be zero, or very close to zero: given that the match is wet, the conditional that if you strike it it will light ought to have a very low probability. However, this is not the result that we get if we follow the calculation through.

Suppose the match is wet, so we have W. Then if $S \wedge L$, $S > L$ has the value 1. Hence, the probability that $S > L$ has the value 1 is $P(S \wedge L|W)$. Similarly, the probability that $S > L$ has the value 0 is $P(S \wedge \neg L|W)$, and the probability that $S > L$ has the value $P(L|S)$ is $P(\neg S|W)$. Therefore, the expected value is

$$1 \times P(S \wedge L|W) + 0 \times P(S \wedge \neg L|W) + P(L|S) \times P(\neg S|W). \quad (28)$$

Now, let us assign some numbers to the probability judgments:

- The probability of the match being wet is $P(W) = .55$.
- If the match is wet, it will certainly not light: $P(L|W) = 0$, therefore $P(S \wedge L|W) = 0$.
- In general, there is a .9 chance of the match lighting if struck: $P(L|S) = .9$.
- There is a fifty-fifty chance of the match being struck: $P(S) = P(\neg S) = .5$.
- If the match is not struck, it is certainly wet: $P(W|\neg S) = 1$. By Bayes' rule,

$$P(\neg S|W) = P(W|\neg S) \times \frac{P(\neg S)}{P(W)} = 1 \times \frac{.5}{.55} \approx .91.$$

So we get

$$P(S > L|W) = 1 \times 0 + 0 + .9 \times .91 \approx .82. \quad (29)$$

So, instead of 0, we got a rather high value, .82.

To conclude this section, we can say that the jury is out regarding whether a probabilistic approach to natural language conditionals is possible. While there are some interesting approaches, most of the results are negative. It might turn out that although conditionals *look* very much like conditional probabilities, there is, in fact, no relation between them. This, indeed, was Lewis's (1976) original conclusion. On the other hand, it may turn out that probabilistic notions, such as expectations, while not determining the semantics of conditionals, may yet enable us to derive deep insights about the meaning of this construction.

9.4 Vagueness

We have seen that it is hard to determine not only the truth conditions of conditionals, but even their truth *values*. That is to say, it is often not clear whether, in a given situation, the conditional *If A then B* is true, false, or somewhere in between. It is for this reason that a probabilistic account of conditionals appears, at least initially, so appealing.

Another phenomenon where judgments of truth value are hard, and, consequently, a probabilistic account is attractive, is that of vague predication. Many terms in natural language, probably the vast majority, are vague. For example, how tall does a person need to be in order to be considered tall? We may all agree that a seven-foot person is tall and a five-foot person is not tall. But what about people with heights in between?

Kamp (1975) addresses these issues. His motivation is to give an account of comparatives in terms of positive adjectives (e.g., to analyze *taller* in terms of the meaning of *tall*). This calls for a semantics where a predicate can hold of an entity to a certain degree. Then we can say that John is taller than Bill just in case the predicate *tall* holds of John to a greater degree than it does of Bill.

Kamp argues against using a multivalued logic (i.e., a logic where propositions can have truth values ranging between 0 and 1) to account for degrees. He maintains that it would be impossible to provide combinatory rules for such a logic that will satisfy our intuitions.

For example, what would the value of $[\![\neg\phi]\!]$ be? It ought to be $1 - [\![\phi]\!]$. But then what do we do with $[\![\phi \wedge \psi]\!]$? Suppose that $[\![\phi]\!] = [\![\psi]\!] = .5$. Then, of course, $[\![\phi \wedge \psi]\!] \leq .5$, but what is its precise value? We cannot define $[\![\phi \wedge \psi]\!] = .5$, for then we would get $[\![\phi \wedge \neg\phi]\!] = .5$, where, in fact, we want $[\![\phi \wedge \neg\phi]\!] = 0$. Nor can we define $[\![\phi \wedge \psi]\!] = 0$, for then we would get $[\![\phi \wedge \phi]\!] = 0$, which is absurd. And any number between 0 and .5 will give the wrong results for both $[\![\phi \wedge \neg\phi]\!]$ and $[\![\phi \wedge \phi]\!]$.

Although Kamp does not use a multivalued logic, he finds it instructive to consider a simple case of such a logic, a three-valued logic: true, false, and undefined. In this logic, models may be *partial*. In classical logic, the interpretation function F assigns to each symbol of arity n an n-place relation on the universe U. Thus, constants are assigned individuals, one-place predicates are assigned sets of individuals, two-place relations are assigned sets of ordered pairs, and so on. However, in a partial model the interpretation function assigns to each symbol of arity n an ordered pair of n-place relations: $F(Q^n) = \langle F^+(Q^n), F^-(Q^n)\rangle$. The idea is that $F^+(Q^n)$ is where Q^n definitely holds, $F^-(Q^n)$ is where Q^n definitely does not hold, and the rest is where Q^n is undefined. For example, suppose we have three individuals, Aleksandra, Bart, and Caroline. Aleksandra is definitely tall, Caroline is definitely not tall, and Bart is neither definitely tall nor definitely not tall. Then:

(30) $F(\textbf{tall}) = \langle\{\text{Aleksandra}\}, \{\text{Caroline}\}\rangle$

There is a theory that provides the right truth conditions for tautologies and contradictions in partial models: this is the theory of supervaluation (van Fraassen 1969). The idea is this. Suppose we have a formula that contains one or more truth value gaps: the truth value of some (perhaps all) of its parts is undefined. Then we consider all the possible completions, that is, all models where the gaps are "filled" in a consistent manner. So we have a set of models that are not partial, where every formula is either true or false. Then we look at our formula: if it is true in all completions, we consider it true, and if it is false in all completions, we consider it false; otherwise, its truth value is undefined. For example, $p \wedge \neg p$ will be false, because in all models in which p has a definite truth value, $p \wedge \neg p$ is false. For similar reasons, $p \vee \neg p$ is true.

This still does not help in the general case, but it is a step forward. We can associate with every sentence the set of all completions that make it true. Sentences that are already true, or tautologies, will be associated with the set of all completions; sentences that are already false, or con-

tradictions, will be associated with the empty set. And intermediate sentences will have an intermediate set associated with them. Intuitively, the larger the set associated with a sentence, the "truer" it is. Kamp's idea is to have a probability function defined over these sets to make this intuition precise.

Kamp defines a *vague model* $\mathscr{M}$ for a language L (for simplicity, L is taken to be extensional) as a quadruple $\langle M, \mathscr{L}, \mathscr{F}, P\rangle$ where

1. M is a partial model for L;
2. $\mathscr{L}$ is a set of completions of L;
3. $\mathscr{F}$ is a field of subsets over $\mathscr{L}$;
4. for each formula $\phi \in L$ and assignment g, $\{M' \in \mathscr{L} : [\![\phi]\!]^{M',g} = 1\} \in \mathscr{F}$;
5. P is a probability function over $\mathscr{F}$.

Now, the degree to which a sentence ϕ (with respect to assignment g) is true can be defined as the value of the probability function over those completions where it is true:

(31) $[\![\phi]\!]^{\mathscr{M},g} = P(\{M' \in \mathscr{L} : [\![\phi]\!]^{M',g} = 1\})$

Thus, we capture the intuition that a sentence is "truer" the larger the set of completions that make it true. Because P is a probability function, if ϕ is definitely true in M (or is a tautology), it is definitely true in $\mathscr{M}$, and if it is definitely false in M (or a contradiction), it is definitely false in $\mathscr{M}$. The reason is that if ϕ is definitely true or a tautology, then all completions make it true, and $P(\mathscr{L}) = 1$; and if ϕ is definitely false or a contradiction, then no completion makes it true, and $P(\varnothing) = 0$.

Let us consider the role probability plays in Kamp's system. We have already seen the importance of using probability to capturing correctly the truth conditions of nonvague sentences, tautologies, and contradictions, something that a multivalued logic on its own is claimed to be incapable of doing. But Kamp notes another advantage. Given a vague predicate, we can, in principle, decide arbitrarily whether a given individual falls into its extension or not. But once we have made such a decision, decisions regarding other individuals are not necessarily arbitrary. For example, suppose there are two persons, $\mathbf{p}_1$ and $\mathbf{p}_2$, whose respective heights are h_1 and h_2, and suppose $h_1 < h_2$. Then, we do not have to decide that $\mathbf{p}_1$ is tall; but if we do, we *must* say that $\mathbf{p}_2$ is tall too. This means that there is no completion where $\mathbf{p}_1$ is tall but $\mathbf{p}_2$ is not. Therefore,

$$\{M' \in \mathscr{L} : [\![\mathbf{tall}(\mathbf{p}_1)]\!]^{M',g} = 1\} \subseteq \{M' \in \mathscr{L} : [\![\mathbf{tall}(\mathbf{p}_2)]\!]^{M',g} = 1\}, \qquad (32)$$

and hence

$$[\![\mathbf{tall}(\mathbf{p}_1)]\!]^{\mathcal{M},g} \leq [\![\mathbf{tall}(\mathbf{p}_2)]\!]^{\mathcal{M},g}. \quad (33)$$

That is to say, it is "truer" that $\mathbf{p}_2$ is tall than that $\mathbf{p}_1$ is tall. Kamp uses this (after further refinements of his system, which do not concern us here) as the truth conditions of (34):

(34) $\mathbf{p}_2$ is taller than $\mathbf{p}_1$.

The use of probabilities means that one cannot have a simple formula for calculating, say, the truth value of $\phi \wedge \psi$ based on the truth values of ϕ and ψ. However, in special cases, this can be done. We know that $P(\phi \wedge \psi) = P(\phi) \times P(\psi)$ if ϕ and ψ are independent, that is, if $P(\phi|\psi) = P(\phi)$.

Kamp does not propose an interpretation of his probability function. In particular, he does not propose an interpretation of the notion of independence. What does it mean to say that two propositions are independent of each other? Edgington (1996) considers this question.

One interpretation of probability that appears to be attractive for the meaning of vagueness is the subjectivist, or Bayesian, interpretation, according to which the probability of ϕ indicates our degree of belief that ϕ is true. If we are certain that John is tall, we would assign to (35) the value 1; if we are certain that he is not tall, we would assign it the value 0; and if we are not certain, we would assign it an intermediate value:

(35) John is tall.

Thus, we could interpret independence as follows: ϕ is independent of ψ iff knowing ϕ gives us no rational reason to change our belief about ψ. For example, if we know that John is tall, this does not affect our belief about how intelligent he is; hence, (35) and (36) are independent of each other:

(36) John is intelligent.

However, Edgington argues that belief is not the right sort of interpretation for vagueness—that there is a difference between being uncertain about whether ϕ holds and ϕ's holding to a certain degree. For example, imagine we would like to drink coffee, and we have to decide whether to visit Shirley or Roger. Shirley may serve us either coffee or tea; we do not know what beverage it would be, so the probability that we will drink coffee is .5. Roger, on the other hand, will definitely serve us a drink that

is an odd mixture of coffee and tea; it is indeterminate whether it can be properly called "coffee" or "tea." Clearly, the two choices are not equivalent. We may decide that even tea is better than drinking Roger's concoction, and visit Shirley; or we may decide that the prospect of drinking tea is so abhorrent, that Roger's novelty drink is the safest choice. But the point is that having a beverage that is coffee to the degree .5 is not the same as having a .5 chance of drinking coffee.

How, then, are we to interpret the statement that ϕ and ψ are independent? According to Edgington, $P(\phi|\psi)$ is to be interpreted as the truth value of ϕ if we decide that ψ is definitely true. For example, suppose John is good at math to the degree d_1 and intelligent to the degree d_2. Then, if we decide that John is definitely good at math, this may affect the truth value of his being intelligent. Hence, the two statements are not independent. On the other hand, if we decide that John is definitely tall, this would probably not affect our judgment of how intelligent he is. Hence, being tall and being intelligent are independent.

Does Kamp's account of vagueness, which makes use of probability, fit our intuition? The answer is not very clear. Suppose, for example, that John is tall to the degree .5 and intelligent to the degree .5. If the properties of being tall and being intelligent are independent of each other, it follows that John has the property of being both tall and intelligent to the degree $.5 \times .5 = .25$. It is not immediately clear whether this is or is not a reasonable conclusion.

9.5 *Many*

A vague word that has received considerable attention is *many* (and its counterpart, *few*). How many is many? There is obviously no clear boundary between what is considered many and what is not.

The problem of *many* is complicated further by the fact that, on top of its vagueness, it is also ambiguous, having at least two readings (Partee 1988).[8] Consider the following sentence, for example:

(37) Many Nobel laureates watched the Olympic games.

Sentence (37) may mean simply that the number of Nobel laureates who watched the Olympic games was large, compared with some norm n; this is the *cardinal* reading. Under this reading, (37) is false, simply because there aren't many people who have won a Nobel prize. But (37) can also mean that the proportion of Nobel laureates who watched the Olympic

games was large, compared with some norm k; this is the *proportional* reading. Under this reading, (37) may very well be true.

The two readings of *many* can be formalized as follows:

1. Cardinal: **many**(A, B) iff $|A \cap B| > n$
2. Proportional: **many**(A, B) iff $|A \cap B| > k \times |A|$

Partee argues at length that the two readings are indeed distinct and have different truth conditions. One of the distinctions between the two readings is that only the cardinal reading is symmetric; the proportional one is not. Thus, if we understand (37) as saying simply that the number of Nobel laureates who watched the Olympic games is high, (38) is true. If, more plausibly, we take (37) to indicate a high percentage of Nobel laureates who are interested in the Olympic games, (38) is false:

(38) Many people who watched the Olympic games were Nobel laureates.

As noted above, besides being ambiguous, *many* is also vague; the parameters n and k do not have a precise value. Where would these values come from? Fernando and Kamp (1996) suggest that these values depend on our expectations: we may not expect Nobel laureates to be interested in anything as mundane as the Olympic games, and hence even a small number or percentage of them will count as *many*; on the other hand, we may expect sports fans to be very interested in the Olympic games, so that much higher norms would be required to make (39) true:

(39) Many sports fans watched the Olympic games.

Informally, Fernando and Kamp propose that *Many As are Bs* is true just in case *It could well have been the case that fewer As are Bs*. Taking *many* to be dependent on expectation naturally calls for a probabilistic account: something is expected if its probability is high, unexpected if its probability is low. But what sort of probability account will be appropriate? As in the case of generics and frequency statements, the crucial question appears to be that of intensionality. Is *many* extensional or intensional?

At first glance, it appears that *many* must be extensional. If we look at its definition above, we see that, on both its readings, the only thing about the sets A and B that matters is their cardinalities. Once the extensions of A and B are known, we can compute the cardinality of their intersection (and, for the proportional reading, also the cardinality of A), and we are done.

However, things are not that simple. Keenan and Stavi (1986) produce the following minimal pair:

(40) a. Many lawyers attended the meeting last year.
b. Many doctors attended the meeting last year.

They point out that even if all the lawyers were doctors and all the doctors were lawyers (so the predicates *lawyer* and *doctor* were coextensional), the truth values of the two sentences might differ. For example, if the meeting in question is a meeting of a medical association, we would expect doctors to attend, but not lawyers; hence, if the meeting is small, we may still affirm (40a) while denying (40b).

A similar point is made by Kamp and Reyle (1993), whose examples are these:

(41) a. Many houses in X burned down last year.
b. Many houses in X were insured against fire last year.

Even if the all the houses that burned were insured, and all the insured houses burned down, the truth values of the sentences might differ. The expectation that a house will be insured is much greater than the expectation that it will burn down (indeed, this is the rationale behind the insurance business). Hence, after a given fire, we may affirm (41a) yet deny (41b).

How can one account for this seemingly intensional behavior of *many*? The explanation is that once the values of n and k are fixed, *many* is, indeed, extensional. But these same values depend on the intensions of the arguments of *many*. Thus, the truth conditions of *many* do, after all, depend on the intensions of its arguments.

Consequently, the probability judgment expressed by *many* cannot be extensional, so Fernando and Kamp argue that it cannot be a type of ratio theory. They propose instead a type of logical relation theory, that is, an account of probability in terms of possible worlds.

Since, according to their theory, *many* is intensional, they replace the sets A and B with the formulas ϕ and ψ, and consider the formula $\mathbf{many}_x(\phi(x), \psi(x))$. Its cardinal reading amounts to saying that there is some number n such that there are at least n individuals that are both ϕ and ψ, and that this n counts as many:

(42) $\mathbf{many}_x(\phi(x), \psi(x))$ iff
$$\bigvee_{n \geq 1} ((\exists_{\geq n} x(\phi(x) \wedge \psi(x)) \wedge n\text{-}\mathbf{is\text{-}many}_x(\phi(x), \psi(x))$$

The expression $\exists_{\geq n}x$ is an abbreviation for "there are at least n xs." More interesting is the second conjunct, which expresses the claim that this n is considered many with respect to ϕ and ψ. It is here that probability plays a role.

Let $|\alpha(x)|_{x,w}$ indicate the number of individuals that satisfy α in world w: $|\{u : [\![\alpha(x)]\!]^{w,g[u/x]} = 1\}|$. If w is the actual world, this will be written simply $|\alpha(x)|_x$. Then, the truth conditions of n-**is-many** can be represented as follows:

(43) n-**is-many**$_x(\phi(x), \psi(x))$ iff $P(\{w : |\phi(x) \wedge \psi(x)|_{x,w} < n\}) > c$

In words: take the probability of the set of worlds where the number of individuals satisfying both ϕ and ψ is less than n; this probability is greater than some parameter c. Now, **many**$_x(\phi(x), \psi(x))$ means that there is some number n, such that there are n individuals that are both ϕ and ψ, and there could well have been fewer than n.

Note that, under the cardinal reading, **many**$_x(\phi(x), \psi(x))$ is a function of (the intension of) $\phi \wedge \psi$, hence its symmetry.

Let us now turn to the proportional reading. The only difference is in the definition of n-**is-many**. Here, instead of unconditional probability, conditional probability is used:

(44) n-**is-many**$_x(\phi(x), \psi(x))$ iff
$P(\{w : |\phi(x) \wedge \psi(x)|_{x,w} < n\} | \{w : |\phi(x)|_{x,w} = |\phi(x)|_x\}) > c$

In words: n is many iff there could well have been fewer than n xs that satisfy $\phi(x) \wedge \psi(x)$, given that there are $|\phi(x)|_x$ xs that satisfy $\phi(x)$.

Since ϕ has a privileged position (it forms the reference class of the conditional probability), this reading is asymmetric, as desired.

9.6 *Even*

We have just seen a probabilistic account of *many*, which views probability as expectation. If something is expected, it has a high probability; if something is surprising, we assign it a low probability. One would expect, then, to have probabilistic accounts of natural language expressions associated with expectation.

One such expression, whose semantics and pragmatics have received some attention, is *even*. Consider, for example, a sentence such as (45):

(45) The food was so good, that even Denise finished everything on her plate.

This sentence conveys, in some sense, that we do not expect Denise to finish everything on her plate, that she is unlikely to do so. Thus, the fact that she did finish eating everything is used as a demonstration of the truth of the claim that the food was good. If, in contrast, Denise were expected to eat anything put before her, regardless of its quality, there would be no reason to infer that the food was exceptionally good.

Exactly how *even* conveys this notion of surprise or unlikelihood is a matter of some debate; most researchers claim it is a matter of presupposition or conventional implicature, though it has also been suggested that this meaning is conveyed through conversational implicature (Kay 1990) or even as part of the truth conditions of *even* (Lycan 1991). This controversy is not really our concern here. What is more interesting is whether probability can be used as a way to model this surprise.

As it happens, no thorough account of *even* in explicit probabilistic terms has been proposed, as far as I know; but such an account is sketched by Chierchia and McConnell-Ginet (1990, 387). They consider sentences of the form in (46):

(46) Even N S

The sentence S is subjectless, and its logical form, S′, is an open formula. N is a proper noun, and its logical form, N′, is a constant. Then the logical form of (46) is (47):

(47) $\exists x(x = \mathrm{N}' \wedge \mathrm{S}')$

If $[\![\mathrm{N}']\!]^{M,g} = a$, then (47) is, of course, satisfied just in case (48) holds:

(48) $[\![\mathrm{S}']\!]^{M,g[a/x]} = 1$

For example, the (relevant part of) the logical form of (45) is (49),

(49) $\exists x(x = \mathbf{d} \wedge \mathbf{finish\text{-}all\text{-}food}(x))$

which is satisfied just in case (50) holds:

(50) $[\![\mathbf{finish\text{-}all\text{-}food}(x)]\!]^{M,g[\mathrm{Denise}/x]} = 1$

The interesting part, of course, is the presupposition of (46). According to Chierchia and McConnell-Ginet, this presupposition is satisfied in a context iff the context makes salient some probability measure, and some set A of individuals, such that for every $a' \in A$ (with the exception of a itself), the probability that $[\![S']\!]^{M,g[a'/x]} = 1$ is greater than the probability that $[\![S']\!]^{M,g[a/x]} = 1$. Thus, (45) presupposes that everybody else is more likely to finish all the food on their plate than Denise is.

As mentioned above, Chierchia and McConnell-Ginet's proposal is more of a note than a thorough treatment of *even*, and the authors willingly admit that "[m]uch more needs to be done, of course, for an adequate analysis" (p. 387). Indeed, this issue is not taken up in the second edition of their textbook (Chierchia and McConnell-Ginet 2000). Why has there been no detailed probabilistic treatment of *even*?

Francescotti (1995) proposes an answer. He argues that probability is the wrong sort of concept to be used in formalizing the meaning of *even*. He considers the following pair of sentences:

(51) a. Granny was accused of kidnapping, and even murder.
b. Granny was accused of murder, and even kidnapping.

Supposing that murder is more common than kidnapping, the probability that Granny committed murder is higher than that she committed a kidnapping; hence, Francescotti claims, if surprise were accounted for in terms of probability, we would expect Granny's being accused of kidnapping to be more surprising than her being accused of murder. We would therefore expect (51b) to be felicitous and (51a) to be odd. But the facts are exactly the reverse: (51a) is fine, but (51b) is odd.

Francescotti explains these judgments by saying that surprise is sensitive to more than probability—for example, to the moral and legal significance of an act. Francescotti's statement is not entirely clear, and there are two ways to understand it.

One possible interpretation is that even if A is more probable than B, it can be more surprising, because its implications are more significant. This is the interpretation preferred by Kay (1990, 84). He presents the following example:

(52) A: It looks as if Mary is doing well at Consolidated Widget.
George [the second vice president] likes her work.
B: That's nothing. Even Bill [the president] likes her work.

According to Kay, (52) is quite felicitous even in a context where there is no reason to think that Bill is less likely than George to like Mary's work. The reason why *even* is felicitous is that Bill's liking Mary's work is a stronger statement: it shows evidence of a higher level of success.

Another way to interpret evidence such as (51) and (52) is that *even* is, in fact, dependent on probability, but that the probability judgments with respect to which a sentence is evaluated are not always determined

directly by the sentence. Though Francescotti is not entirely clear on this point, he seems to favor this interpretation.

For example, if we evaluate the sentences in (51) with respect to the respective probabilities of murders and kidnappings, we would, indeed, erroneously predict that (51a) is odd and (51b) is fine. But if, instead, we notice that committing murder puts one in more trouble than kidnapping, and if we assume that people are not likely to cause themselves more trouble than is necessary, we would be "more surprised to find that one would do what would put one in more serious moral or legal trouble. It is for this reason that we find Granny's committing murder more surprising than her kidnapping" (Francescotti 1995, 167).

This probability judgment may be affected by the specific character of Granny herself: "[s]uppose we know that Granny has the propensity, not only for getting in trouble, but also performing actions that afford her the maximal amount of trouble. In this case, we would not be surprised at all to find that Granny has placed herself in moral and legal jeopardy, and [(51b)] would therefore be more felicitous than [(51a)]" (p. 167).

At this point, it appears that in the case of *even*, as in the case of conditionals, more research is needed before we can conclude whether a probabilistic account is desirable, let alone feasible.

9.7 Indirect Use of Probability

The previous sections discussed several cases where probability is explicitly introduced into the semantics in order to provide an account of some phenomenon. As mentioned in the introduction, however, there are also less direct uses of probability. There are semantic accounts that do not use probability as such, but are inspired by probabilistic notions.

A good example of such a theory is van der Does and van Lambalgen's (2000) logic of perception reports. While their logic does not use probability as such, it does use ideas from probability theory, in particular the notion of conditional expectation.

The problem that van der Does and van Lambalgen set out to solve is the fact that, in general, perception reports are not veridical. Take (53), for example:

(53) I see this arm.

Suppose we represent (53) as (54):

(54) $\mathbf{see}(\mathbf{I}, x)$

Given an assignment function that assigns the individual arm in question to x, (54) would be satisfied just in case I do, in fact, see this arm.

However, these are not really the truth conditions of (53). For this sentence to be true, all that is required is that something would appear like this arm to me. It may indeed be this arm, but it may be another arm, or a leg, or a loaf of bread, or, indeed, nothing at all—a hallucination. Intuitively, the actual nature of the object I see depends on how well I can see and interpret what I see, which in turn depends both on my vision and on the current conditions (lighting, occluding objects, etc.). If I have perfect vision, the conditions are ideal, nothing interferes with or occludes my vision, and there is no possibility of hallucination, then it would indeed follow from (53) that there exists an arm that I see. In contrast, if my vision is really poor, then nothing at all can be concluded about the object I see. In this case, all we can say is that there is something I see—but what it really is, we cannot say:

(55) $\exists x\ \mathbf{see}(\mathbf{I}, x)$

Van der Does and van Lambalgen use Marr's (1982) theory of vision. According to Marr, the same object may be represented at various levels of detail. For example, we can view the arm as simply a sort of cylindrical shape; or, in a more detailed way, as two cylinders, corresponding to the forearm and the upper arm; or as a more detailed picture still, composed of three cylinders, corresponding to the upper arm, the forearm, and the hand; or with additional cylinders, corresponding to the fingers; and so on. At one extreme, the representation of an object is a blur where we can distinguish nothing about the object; at the other extreme, the representation is an exceedingly specific picture, where all the details are discernible.

Viewed in this way, vision involves approximation: we start from the actual object, which is infinitely detailed, and we arrive at some sort of mental representation that filters out many of these details. A consequence of this is that in the context of a perception report, we cannot simply talk about an object being seen to have some property; rather, we must talk about whether, given a certain level of detail, it is seen to have this property.

Applying Marr's idea to a concrete example, we can represent the predicate **arm** using a series of models; in each of them the predicate is interpreted at a different level of detail, but in each of them it is an arm. Van der Does and van Lambalgen call this system a *refining system*.

Thus, for example, in a coarse model, we may have an individual a that is an arm; but in a more refined model, this individual will turn out to be a composite $\{u, f\}$, where u is the upper arm and f is the forearm. We can order these models according to how refined they are; intuitively, the limit of this series is reality, that is, the completely detailed representation of the object.

Consider (53) again. Suppose that at the current level of detail, I cannot distinguish an arm from a leg, yet I am still quite able to distinguish an arm from a loaf of bread. In this case, we do know something about the thing I see, but not its exact identity. So we cannot have a logical form like (54), where the variable is totally free, a case representing perfect knowledge; nor do we want (55), where the variable is totally bound, representing no knowledge at all. We want, intuitively, a case where the variable is "partially" bound, representing partial knowledge.

It is at this point that a notion from probability theory becomes helpful. Probability theory offers a device to represent partial knowledge, namely, conditional expectation.

Conditional expectation is used to represent cases when we cannot measure a random variable completely accurately, instead having only partial knowledge of its value. Suppose, for example, that a person is picked at random, using some probability distribution. This person has a precise height; let the value of the random variable X be this height. Suppose we need to make our best guess about the value of X. If our vision is infinitely accurate (say, we have some sort of bionic eye), then we can know every person's height just by looking—we have perfect information. So, in this case we do not need to guess—we know the precise value of X. In the other extreme case, our vision is very poor; we cannot make out the person's shape at all. What is our best guess about this person's height? Our best strategy is to use the average height of people—namely, the expected value of X—and guess that this is the person's height.

Now, suppose we can see the person, but cannot assess the height precisely. All we can distinguish is whether the person is tall, of middle height, or short. Then, if the person is tall, we should guess the average height of tall people, that is, the expected value of X among tall people. If the person is of middle height, we should guess the expected value of X among middle-sized people; and if short, the expected value among short people.

Mathematically, if $\mathcal{G}$ is a σ-field generated by the sets of tall people, middle-sized people, and short people, then our guess of the height of the person is the conditional expectation of X given $\mathcal{G}$, $E(X|\mathcal{G})$. Note that $E(X|\mathcal{G})$ is itself a random variable: its value will be either the average height of tall people, the average height of medium-sized people, or the average height of short people. Thus, it is "smoother" than the original random variable X: it filters out distinctions that are real, but cannot be perceived.

Van der Does and van Lambalgen propose a counterpart of conditional expectation for their logic of vision, which they call *conditional quantification*. Let M be a model, $\mathcal{F}$ the set of assignments on M, $\mathcal{G}$ an algebra of subsets of $\mathcal{F}$, and ϕ a formula. Let us identify with a formula ϕ the set of assignments that make it true, $\{f \in \mathcal{F} : [\![\phi]\!]^{M,f} = 1\}$. Then, the set of assignments that corresponds to the conditional quantifier $\exists(\phi|\mathcal{G})$ is $\bigwedge\{\mathcal{C} \in \mathcal{G}|\phi \subseteq \mathcal{C}\}$.

The idea is that $\exists(\phi|\mathcal{G})$ is the best estimate of the assignments that make ϕ true on the basis of the information available in $\mathcal{G}$. $\mathcal{G}$ contains those propositions (sets of assignments) whose truth we can verify at the current level of detail. So, an assignment makes $\exists(\phi|\mathcal{G})$ true just in case it makes true those statements entailed by ϕ that we can verify at the current level of detail.

As an example, let us go back to (53). Its logical form, according to van der Does and van Lambalgen, is (56):

(56) $\exists(\mathbf{arm}(x)|\mathcal{G})$

The algebra $\mathcal{G}$ represents the (visual) knowledge the speaker has. Now, suppose there are three individuals in the universe, a, l, and b; and we know that a is an arm, l is a leg, and b is a loaf of bread.

Now suppose the speaker has perfect vision; every property can be identified. In this case, $\mathcal{G}$ is the algebra of the power set of $\mathcal{F}$, the set of all assignments. So, an assignment will make (56) true just in case it makes true all its entailments that the speaker can verify. But since the speaker can verify everything, an assignment will make (56) true just in case it will make $\mathbf{arm}(x)$ true, as desired. Formally:

$$\bigwedge\{\mathcal{C} \in \mathcal{G}|\mathbf{arm}(x) \subseteq \mathcal{C}\} = \bigwedge\{\mathcal{C} \subseteq \mathcal{F}|\mathbf{arm}(x) \subseteq \mathcal{C}\} = \mathbf{arm}(x)$$
$$= \{f : f(x) = a\}. \tag{57}$$

To take the other extreme case, suppose the speaker cannot distinguish any properties. The algebra $\mathcal{G}$ will be simply $\{\varnothing, \mathcal{F}\}$. Then, the set

of assignments that make (56) true will be those that make true all the entailments of **arm**(x) that the speaker can verify. But since the speaker can verify nothing, this will be the set of all assignments. Formally:

$$\bigwedge\{\mathscr{C} \in \mathscr{G} | \mathbf{arm}(x) \subseteq \mathscr{C}\} = \mathscr{F} = \{f : f(x) = a \text{ or } f(x) = l \text{ or } f(x) = b\}. \quad (58)$$

So, indeed, in a case of no information at all, we are not able to distinguish any object from any other object.

Now, suppose the speaker can identify the property **body-part**, but not **arm**. So the speaker is able to distinguish a from b, but not from l. The algebra will be generated by {**body-part**(x)}. An assignment will satisfy (56) just in case it satisfies its entailments that the speaker can verify, namely, {**body-part**(x)}. Formally:

$$\bigwedge\{\mathscr{C} \in \mathscr{G} | \mathbf{arm}(x) \subseteq \mathscr{C}\} = \{\mathbf{body\text{-}part}(x)\} = \{f(x) = a \text{ or } f(x) = l\}. \quad (59)$$

This is the desired result.

9.8 Conclusion

In this chapter, I have discussed the role that probability, in a variety of ways and interpretations, plays in the study of semantic phenomena. I have considered a number of proposals, each setting out to answer a specific question about a specific phenomenon; let the reader decide to what extent each specific proposal is successful.

Considering all these theories as a whole, a pattern suggests itself: probability appears as a tool that aids traditional truth-conditional semantics, not as a replacement for it. Probabilistic notions do not change the fundamentals of semantic theory, but interact with them in specific cases in order to solve a specific problem.

It is conceivable that a more radical use of probability might be discovered, one that plays a role in the fundamentals of the theory. Let me sketch what direction such a theory might take.

Traditional semantics treats the meaning of a sentence as a function from situations (often considered to be possible worlds) to truth values. The intuition is that in order to demonstrate understanding of the meaning of a sentence, one must be able to judge its truth or falsity in any situation. Crucially, the description of the situation must be complete. That is, all the facts about the situation must be known; otherwise, one may understand a sentence yet still be unable to judge its truth value. For

instance, we all presumably understand (60); yet we are unable to judge its truth value with certainty, because we lack sufficient information. But our understanding the meaning of the sentence implies that if we did have complete information, we would know if it were true or false.

(60) There is life on Mars.

Now, this is a rather idealized view of what understanding the meaning of a sentence is. In practice, we hardly ever have complete information; the way by which we demonstrate understanding of a sentence involves, by necessity, judgments in cases of incomplete information. If we wish to take the real rather than the idealized conditions as the definition of meaning, we would need to use probability, or something like it, at a fundamental level of our semantics. Then, we will be talking about judgments of probability, rather than judgments of truth values; to demonstrate understanding of the meaning of a sentence, one would need to be able to judge the likelihood that it is true in any given situation. The meaning of a sentence would, then, not be a function from possible worlds to truth values; rather, it would be a function from states of knowledge to probability values. We can take the usual approach of representing knowledge in terms of a set of possible worlds; the meaning of a sentence, then, would be a function from sets of possible worlds to probabilities. This would capture formally the idea that understanding the meaning of a sentence is the ability, given a situation, to assess its probability.

Such a semantics has never, to my knowledge, been carefully proposed. There is, of course, a good reason for this: such a theory would be considerably more complicated than truth-conditional semantics, and it is not clear that it would have advantages over the more traditional semantics. But it might.

Notes

I would like to thank Fritz Hamm, who gave me much detailed and enlightening advice, and Chris Manning, who reviewed the chapter and made many useful suggestions on both content and form.

1. Following de Swart (1991) and others, I distinguish between frequency adverbs, such as *usually* and *always*, and other adverbs of quantification, such as *twice*. I refer to a sentence containing a frequency adverb as a *frequency statement*.

2. Actually, Carnap uses *state descriptions* rather than possible worlds, but his approach can naturally be recast in terms of the latter.

3. See Cohen 2002; compare Fisher's (1959) requirement that the reference class be "subjectively homogeneous and without recognizable stratification" (p. 33).

4. Or, more precisely, with what he calls "statistical probability."

5. For the distinction between the two types of intensionality, depending on the index involved, see Landman 1989.

6. L. J. Cohen refers to this property as "counterfactualizability," but I believe this term is somewhat confusing, as well as being rather cumbersome.

7. See Edgington 1991 for a good survey of the issues involved.

8. I say *at least* because it has been claimed that, in fact, *many* has additional readings (Westerståhl 1985; Lappin 1988, 1993; Cohen 2001); but these will not concern us here.

This page intentionally left blank

Glossary of Probabilistic Terms

Bayes' rule Bayes' rule calculates the **conditional probability** $P(A|B)$ from $P(B|A)$. This is useful when the former quantity is difficult to determine. From the equivalence $P(A, B) = P(B, A)$ (see **joint probability**), we derive that $P(A)\ P(B|A) = P(B)\ P(A|B)$, from which Bayes' rule follows:

$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)}.$$

Bayes' rule is often used to determine the most probable hypothesis H given some evidence E: $\text{argmax}_H\ P(H|E)$. Since $P(E)$ is a constant, we may ignore $P(E)$ in maximizing $P(H)\ P(E|H)/P(E)$. In that case, Bayes' rule reduces to

$$\text{argmax}_H\ P(H|E) = \text{argmax}_H\ P(H)\ P(E|H).$$

The probability $P(H)$ is usually called the *prior* probability that H occurs, while $P(E|H)$ is called the *posterior* probability.

Bias of an estimator The bias of an **estimator** is the difference between the expected value (or **expectation**) of the estimator and the true value of the **parameter**.

Bigram A bigram is a sequence of two consecutive items such as phonemes, letters, or words.

Binomial distribution The binomial distribution is a discrete **distribution** that results from a series of trials with only two outcomes (also called Bernoulli trials), each trial being independent from the other ones. Repeatedly tossing a coin is an example of a binomial distribution, but also repeatedly selecting a certain word (type) from a corpus. The binomial distribution gives the number r of successes out of n trials, where the probability of success in each trial is p:

$$P(r; n, p) = \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r}.$$

In this formula, n and p are the *parameters* (see **distribution**) of the binomial distribution (parameters are separated from the random variable by a semicolon in the left-hand side of the function).

Central limit theorem The central limit theorem states that the sum of a large number of independent, identically distributed random variables follows approximately a **normal distribution**.

Chain rule The chain rule is a generalization of the **product rule**. It gives the **joint probability** for n events $A_1, A_2, \ldots, A_n$. That is,

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \times P(A_2|A_1) \times \cdots \times P(A_n|A_1, A_2, \ldots, A_{n-1}).$$

Note that for two events, the chain rule is equivalent to the product rule:

$$P(A_1, A_2) = P(A_1) \times P(A_2|A_1).$$

For **independent** events, the chain rule reduces to

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \times P(A_2) \times \cdots \times P(A_n).$$

Chi-square test The chi-square or χ^2 test is a statistical test used to compare observed frequencies with frequencies expected for **independence**. If the difference between observed and expected frequencies is large, then the **null hypothesis** of independence can be rejected.

Condition The verb *condition* refers to using something on the right-hand side of a **conditional probability** expression to delimit the range of circumstances over which the probability of the thing on the left-hand side of the conditional probability expression is evaluated.

Conditional expectation The notion of conditional **expectation** is useful if we cannot measure a **random variable** completely accurately, but have only partial knowledge of its value. The conditional expectation $E(X|Y)$ is defined as the function of the random variable Y whose value at $Y = y$ is

$$E(X|Y = y).$$

Thus, $E(X|Y)$ is itself a random variable. A very useful property of conditional expectation is that for all random variables X and Y,

$$E(X) = E(E(X|Y)).$$

Conditional independence Two events A and B are conditionally **independent** given event C when

$$P(A, B|C) = P(A|C)\ P(B|C).$$

Conditional probability The notion of conditional probability is useful if we cannot determine the probability of an event directly, but have only partial knowledge about the outcome of an experiment. The conditional probability of an event A given that an event B has occurred is written as $P(A|B)$. For $P(B) > 0$, it is defined as

$$P(A|B) = \frac{P(A, B)}{P(B)},$$

where $P(A, B)$ is the **joint probability** of A and B.

Confidence interval See **Hypothesis testing**.

Contingency table A contingency table allows for presenting the frequencies of the outcomes from an experiment in which the observations in the sample are classified according to two criteria.

Correspondence analysis Correspondence analysis is a statistical technique used to analyze simple two-way and multi-way tables containing some correspondence between the rows and columns.

Dependence See **Independence**.

Disjoint Two sets or **events** are disjoint if they have an empty intersection.

Distribution A probability distribution (or **probability function**) is a function that distributes a mass of 1 throughout the **sample space** Ω. Distributions can be grouped into families that differ only in the value of a so-called *parameter*. A parameter is a constant if one is looking at a specific function, while it is a variable if one is looking at a whole family. There is a distinction to be made between *discrete* distributions, having a finite number of outcomes, and *continuous* distributions, having a continuous domain. The three principal distributions, which occur in a great variety of problems, are the **binomial distribution**, the **Poisson distribution**, and the **normal distribution**.

EM algorithm See **Expectation Maximization algorithm**.

Entropy The entropy H is the average uncertainty of a **random variable** X and is defined as

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

Entropy is normally measured in bits. In Shannon's *Information Theory*, entropy is understood as the amount of information in a random variable. It can be thought of as the average length of the message needed to transmit an outcome of that variable. For an accessible introduction to Information Theory and its application to natural language, see Manning and Schütze 1999.

Estimator An estimator is any **random variable** used to estimate some **parameter** of the underlying **sample space** from which the sample is drawn. A widely used estimator is the **maximum likelihood** estimator. Two important notions related to any estimator are **bias** and **robustness**.

Event An event is a subset of a **sample space** Ω.

Event space The event space is the set of all events of a **sample space** Ω, that is, the power set of Ω.

Expectation The expectation is the mean of a **random variable**. Let X be a random variable assuming the values $x_1, x_2, \ldots$ with corresponding probabilities $P(x_1), P(x_2), \ldots$. The expectation or expected value of X is defined by

$$E(X) = \sum_i x_i P(x_i).$$

The terms *mean*, *average*, and *expectation* are all synonymous.

Expectation Maximization algorithm The Expectation Maximization or EM algorithm provides a general procedure for estimating the values of hidden parameters of a model (e.g., a **hidden markov model**). This algorithm starts with an arbitrary initial hypothesis; it then repeatedly calculates the expected values of the hidden parameters under the assumption that the current hypothesis is correct, after which it recalculates the **maximum likelihood** hypothesis. This procedure is repeated until the change in the values of the parameters becomes negligible.

Gaussian A Gaussian is a probability distribution also known as **normal distribution**.

Hidden markov model A hidden markov model or HMM is a **markov model** that, instead of emitting the same symbol each time at a given state, has available a choice of symbols, each with a certain probability of being selected. In an HMM, the symbol emitted at a certain state is unknown or *hidden*.

HMM See **Hidden markov model**.

Hypothesis testing Hypothesis testing is an assessment of whether or not something is a chance event. It consists of assuming that a *null hypothesis* H_0, which we want to prove false, is really true. We calculate the probability of as extreme a statistical material as the one observed, given that H_0 is indeed true. If this probability is less than some predefined value p, we can discard the null hypothesis with a *significance level* p (e.g., 5%). The significance level is related to the confidence degree in that a method for constructing a *confidence interval* with confidence degree $1 - p$ (e.g., 95%) can be used to construct a hypothesis test with significance level p. Note that failure to discard the null hypothesis does not mean that we can conclude it is true. It may be false or highly unlikely, but the statistical material (the data) does not allow us to discard it at the desired significance level.

Independence Two events A and B are said to be *independent* if

$P(A, B) = P(A)\ P(B)$.

Unless $P(B) = 0$, this is equivalent to saying that

$P(A) = P(A|B)$.

In other words, knowing that B is the case does not affect the probability of A. Otherwise, events are *dependent*.

Independent See **Independence**.

Information Theory See **Entropy**.

Joint probability The joint probability of two events A and B is the probability of both A and B occurring. The joint probability is written as $P(A, B)$ and also as $P(A \cap B)$. Note that

$P(A, B) = P(B, A)$

and

$P(A \cap B) = P(B \cap A)$.

The joint probability is related to the **conditional probability** of A given B, $P(A|B)$, as follows:

$P(A, B) = P(A|B)\ P(B) = P(B|A)\ P(A)$.

Logit function The logit function is a function of the probability of an **event**; it is the natural logarithm of the ratio of the probability of the event occurring to the probability of that event not occurring, given by the formula

$$f(p) = \log_e \frac{p}{1-p}.$$

The logit function is widely used in sociolinguistic modeling (Mendoza-Denton, Hay, and Jannedy, this volume), but also in probabilistic syntax (Manning, this volume) and language change (Zuraw, this volume).

Markov chain See **Markov model**.

Markov model A markov model (also *markov chain* or *markov process*) can be thought of as a stochastic process generating a sequence of symbols, where the symbols are *not* independent but depend on a number of previous symbols in the sequence. This number of previous symbols corresponds to the *order* of a markov model. For example, in a *first-order* markov model, each symbol depends only on its preceding symbol; in a *second-order* markov model, each symbol depends on the two previous symbols; and so on (see also Bod, this volume). A markov model can be conveniently represented by a *state diagram* consisting of *states* and *transitions*. It is said that each state *emits* a symbol, and a transition between two states is labeled with the probability of this transition.

Markov process See **Markov model**.

Maximum likelihood The maximum likelihood hypothesis or **estimator** chooses the parameter values (see **distribution**) that assign the highest probability to the observed outcome. In estimating the probability of a word in a corpus or a context-free rule in a treebank, the maximum likelihood estimator is equal to the relative frequency (of the word or rule). For estimating the values of *hidden parameters* (e.g., in a **hidden markov Model**), the **Expectation Maximization algorithm** is often used. For many linguistic problems, the maximum likelihood estimator is unsuitable because of the sparseness of the data. If a certain linguistic phenomenon can be modeled by a given **distribution**, it is possible to estimate the probabilities of events that lie outside the observed data.

Multiplication rule See **Product rule**.

***n*-gram** An *n*-gram is a sequence of *n* consecutive items such as phonemes, letters, or words.

Nonparametric test A nonparametric test is a statistical test (see **hypothesis testing**) that makes no assumptions about the underlying distribution.

Normal distribution The normal or Gaussian distribution is a continuous **distribution** that can be applied to a wide range of experimental data (see **central limit theorem**) and is often used to approximate the discrete binomial distribution. It is more popularly known as the *bell curve*. The normal distribution has two **parameters**: the mean μ and the **standard deviation** σ. It is defined by

$P(x; \mu, \sigma) = (2\pi\sigma)^{-1/2} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}$.

Null hypothesis See **Hypothesis testing**.

Parameter See **Distribution**.

Parametric test A parametric test is a statistical test (see **hypothesis testing**) that assumes that the underlying distribution follows the **normal distribution**.

Poisson distribution The Poisson **distribution** is a convenient approximation of the **binomial distribution** in the case where the number of trials n is large and the

probability p small. The Poisson distribution is defined by

$$P(k;\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!},$$

where $P(k,\lambda)$ can be conceived as the probability of exactly k successes in an ideal experiment (and where the parameter λ is viewed as a physical **parameter**).

Posterior probability See **Bayes' rule**.

Principal components analysis Principal components analysis is a multivariate statistical procedure that is used to reduce the apparent dimensionality of the data. It leads to a unique solution.

Prior probability See **Bayes' rule**.

Probability distribution See **Distribution**.

Probability function A probability function (or probability distribution) distributes a probability mass of 1 throughout the **sample space** Ω. It is any function that ranges over [0, 1], such that $P(\Omega) = 1$ and for two **disjoint events** A and B,

$P(A \cup B) = P(A) + P(B)$.

(See also Bod, this volume.)

Probability mass function A probability mass function for a **random variable** X gives the probability that the random variable has certain numerical values. We write a probability mass function as $P(X = x)$, which reads as the probability that the random variable X has value x.

Probability space A probability space ties together the three main components of probability theory; it consists of a **sample space**, a **σ-field**, and a **probability function**. The probability space's main use is in providing a method of background assumptions for definitions and theorems.

Product rule The product rule (or *multiplication rule*) states that the **joint probability** of two events A and B is equal to the probability of A multiplied by the **conditional probability** of B given A:

$P(A,B) = P(A)\ P(B|A)$.

If A and B are **independent**, the product rule reduces to

$P(A,B) = P(A)\ P(B)$.

A derivation of the product rule is given in Bod, this volume.

Random variable Intuitively, a random variable is a real number determined by a (probabilistic) experiment. A random variable is defined by the function $X\colon \Omega \to R$, where R is the set of real numbers. Random variables are useful if we want to talk about probabilities of numerical values related to the **event space**. Since random variables have a numerical range, we can work directly with the values of a random variable, rather than with irregular events.

Regression Regression is a statistical procedure to determine the relationship between a dependent variable and one or more independent variables. See Manning, this volume, for some important regression methods.

Robustness of an estimator The robustness of an **estimator** is its breakdown point, that is, the fraction of outlying sample points that can corrupt the estimator.

Sample space A sample space Ω is the set of outcomes for an experiment.

σ-field A σ-field is a set with a maximal element Ω (a **sample space**) and arbitrary complements and unions. The power set of the sample space (i.e., the **event space**) is a σ-field.

Significance See **Hypothesis testing**.

Standard deviation The standard deviation of a **random variable** is closely related to the notion of **variance**. It is defined as the square root of the variance.

Stochastic The word *stochastic* is commonly used as a synonym for *probabilistic*, especially if it refers to a sequence of results generated by an underlying **probability distribution**.

Stochastic process A stochastic process is the generation of a sequence of results with a certain **probability distribution**.

Sum rule For **disjoint events**, the sum rule states that the probability of either event occurring is equal to the sum of the probabilities of the events. That is, for two events A and B, where $A \cap B = \varnothing$,

$$P(A \cup B) = P(A) + P(B).$$

For the general case, where $A \cap B \neq \varnothing$, the sum rule is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

(See also Bod, this volume.)

Trigram A trigram is a sequence of three consecutive items such as phonemes, letters, or words.

Variance The variance of a **random variable** is a measure of how much the values of the random variable vary over trials. Variance can be intuitively understood as how much on average the variable's values differ from the variable's **expectation**:

$$\mathrm{Var}(X) = E((X - E(X))^2).$$

This page intentionally left blank

References

Abney, S. 1996. Statistical methods and linguistics. In J. L. Klavans and P. Resnik, eds., *The balancing act: Combining symbolic and statistical approaches to language*. Cambridge, Mass.: MIT Press, pp. 1–26.

Abney, S. 1997. Stochastic attribute-value grammars. *Computational Linguistics* 23, 597–618.

Abney, S., McAllester, D., and Pereira, F. 1999. Relating probabilistic grammars and automata. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, Md., pp. 542–549.

Ackema, P., and Neeleman, A. 1998. Conflict resolution in passive formation. *Lingua* 104, 13–29.

Adams, E. W. 1998. *A primer of probability logic*. Stanford, Calif.: CSLI Publications.

Agresti, A. 1990. *Categorical data analysis*. New York: Wiley.

Agresti, A. 1996. *An introduction to categorical data analysis*. New York: Wiley.

Aha, D. W., Kibler, D., and Albert, M. 1991. Instance-based learning algorithms. *Machine Learning* 6, 37–66.

Ahrens, K. V. 1998. Lexical ambiguity resolution: Languages, tasks, and timing. In D. Hillert ed., *Sentence processing: A crosslinguistic perspective*. Syntax and Semantics 31. San Diego, Calif.: Academic Press.

Aissen, J. 1999. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17, 673–711.

Albright, A., Andrade, A., and Hayes, B. 2001. Segmental environments of Spanish diphthongization. In A. Albright and T. Cho, eds., *Papers in phonology 5*. UCLA Working Papers in Linguistics 7. Los Angeles: University of California, Los Angeles, Department of Linguistics, pp. 117–151.

Albright, A., and Hayes, B. 2000. An automated learner for phonology and morphology. Manuscript, University of California, Los Angeles.

Allegre, M., and Gordon, P. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40, 41–61.

Allwood, J., Andersson, L.-G., and Dahl, Ö. 1977. *Logic in linguistics*. Cambridge: Cambridge University Press.

Anderson, J. R. 1990. *The adaptive character of thought*. Hillsdale, N.J.: Erlbaum.

Anderson, S. R. 1981. Why phonology isn't "natural." *Linguistic Inquiry* 12, 493–539.

Anderwald, L. 1999. Negation in non-standard British English. Ph.D. thesis, University of Freiburg.

Andrews, S. 1989. Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15, 802–814.

Andrews, S. 1992. Frequency and neighborhood size effects on lexical access: Similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18, 234–254.

Anglin, J. M. 1993. Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*.

Anttila, A. T. 1997. Deriving variation from grammar. In F. Hinskens, R. van Hout, and L. Wetzels, eds., *Variation, change, and phonological theory*. Amsterdam: John Benjamins, pp. 35–68. (Downloadable from the Rutgers Optimality Archive, ROA 63-0000. ⟨http://roa.rutgers.edu/index.php.3⟩.)

Anttila, A. T. 2001. Variation and phonological theory. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes, eds., *The handbook of language variation and change*. Oxford: Blackwell, pp. 206–243.

Åquist, L., Hoepelman, J., and Rohrer, C. 1980. Adverbs of frequency. In C. Rohrer, ed., *Time, tense and quantifiers*. Tübingen, Germany: Niemeyer Verlag, pp. 1–18.

Aronoff, M. 1976. *Word formation in generative grammar*. Cambridge, Mass.: MIT Press.

Aronoff, M. 1994. *Morphology by itself*. Cambridge, Mass.: MIT Press.

Asher, N., and Morreau, M. 1995. What some generic sentences mean. In G. Carlson and F. J. Pelletier, eds., *The generic book*. Chicago: University of Chicago Press, pp. 300–338.

Atkins, B. T. S., and Levin, B. C. 1995. Building on a corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography* 8, 85–114.

Attneave, F. 1959. *Applications of information theory to psychology*. New York: Holt, Rinehart and Winston.

Baayen, R. H. 1994. Productivity in language production. *Language and Cognitive Processes* 9, 447–469.

Baayen, R. H. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, R. H., Burani, C., and Schreuder, R. 1997. Effects of semantic markedness in the processing of regular nominal singulars and plurals in Italian. In G. E.

Booij and J. van Marle, eds., *Yearbook of morphology 1996*. Dordrecht: Kluwer, pp. 13–34.

Baayen, R. H., Dijkstra, T., and Schreuder, R. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language* 36, 94–117.

Baayen, R. H., McQueen, J., Dijkstra, A., and Schreuder, R. In press. Dutch inflectional morphology in spoken- and written-word recognition. *Linguistics*.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. 1995. *The CELEX lexical database (CDROM)*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Baayen, R. H., and Renouf, A. 1996. Chronicling the *Times*: Productive innovations in an English newspaper. *Language* 72, 69–96.

Baayen, R. H., and Schreuder, R. 1999. War and peace: Morphemes and full forms in a non-interactive activation parallel dual route model. *Brain and Language* 68, 27–32.

Baayen, R. H., and Schreuder, R. 2000. Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society (Series A: Mathematical, Physical and Engineering Sciences)* 358, 1–13.

Baayen, R. H., Schreuder, R., de Jong, N. H., and Krott, A. 2002. Dutch inflection: The rules that prove the exception. In S. Nooteboom, F. Weerman, and F. Wijnen, eds., *Storage and computation in the language faculty*. Dordrecht: Kluwer, pp. 61–92.

Baayen, R. H., Schreuder, R., and Sproat, R. 2000. Morphology in the mental lexicon: A computational model for visual word recognition. In F. van Eynde and D. Gibbon, eds., *Lexicon development for speech and language processing*. Dordrecht: Kluwer, pp. 267–291.

Babby, L. 1980. The syntax of surface case marking. In W. Harbert and J. Herschensohn, eds., *Cornell working papers in linguistics*. Ithaca, N.Y.: Cornell University, Department of Modern Languages and Linguistics, pp. 1–32.

Bacchus, F. 1990. *Representing and reasoning with probabilistic knowledge*. Cambridge, Mass.: MIT Press.

Bailey, C.-J. 1973. *Variation and linguistic theory*. Washington, D.C.: Center for Applied Linguistics.

Bailey, C.-J. 1987. Variation theory and so-called "sociolinguistic grammars." *Language and Communication* 7, 269–291.

Bailey, T. M., and Hahn, U. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language* 4, 568–591.

Balota, D. A., and Chumbley, J. I. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance* 10, 340–357.

Barlow, M., and Kemmer, S., eds. 2000. *Usage-based models of language*. Stanford, Calif.: CSLI Publications.

Bates, E., and Devescovi, A. 1989. Crosslinguistic studies of sentence production. In B. MacWhinney and E. Bates, eds., *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press, pp. 225–256.

Bates, E., and MacWhinney, B. 1989. Functionalism and the competition model. In B. MacWhinney and E. Bates, eds., *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press, pp. 3–76.

Bauer, L. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Baxter, W., and Manaster Ramer, A. 1996. Review of Ringe 1992. *Diachronica* 13, 371–384.

Bayes, T. 1764. An essay towards solving a problem in the probability of chances. *Philosophical Transactions of the Royal Society of London* 53, 370–418.

Bayley, R. 2001. The quantitative paradigm. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes, eds., *The handbook of language variation and change*. Oxford: Blackwell, pp. 117–142.

Beckman, M. E., and Edwards, J. 2000. The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development* 71, 240–249.

Beddor, P., Harnsberger, J., and Lindemann, S. In press. Language specific patterns of vowel-to-vowel coarticulation. *Journal of the Acoustical Society of America*.

Beddor, P., and Krakow, R. A. 1999. Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *Journal of the Acoustical Society of America* 106, 2868–2887.

Behrens, H. 2001. How to learn a minority default: The acquisition of the German *-s* plural. Manuscript, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bell, A. 1984. Language style as audience design. *Language in Society* 13, 145–204.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. 2001. Form variation of English function words in conversation. Manuscript, University of Colorado, Lucent Bell Laboratories, Brown University, and University of Pennsylvania.

Berdan, R. 1996. Disentangling language acquisition from language variation. In R. Bayley and D. R. Preston, eds., *Second language acquisition and linguistic variation*. Amsterdam: John Benjamins, pp. 203–244.

Berent, I., and Shimron, J. 1997. The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition* 64, 39–72.

Berkley, D. M. 1994. The OCP and gradient data. *Studies in the Linguistic Sciences* 24, 59–72.

Berkley, D. M. 2000. Gradient OCP effects. Ph.D. thesis, Northwestern University.

Bertram, R., Baayen, R. H., and Schreuder, R. 2000. Effects of family size for complex words. *Journal of Memory and Language* 42, 390–405.

Bertram, R., Laine, M., Baayen, R. H., Schreuder, R., and Hyönä, J. 1999. Affixal homonymy triggers full-form storage even with inflected words, even in a morphologically rich language. *Cognition* 74, B13–B25.

Bertram, R., Schreuder, R., and Baayen, R. H. 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, 419–511.

Bever, T. G. 1970. The cognitive basis for linguistic structures. In J. R. Hayes, ed., *Cognition and the development of language*. New York: Wiley, pp. 279–352.

Biber, D., and Finegan, E. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65, 487–517.

Bickerton, D. 1971. Inherent variability and variable rules. *Foundations of Language* 7, 457–492.

Bickerton, D. 1973. Quantitative versus dynamic paradigms: The case of Montreal "que." In C.-J. Bailey and R. W. Shuy, eds., *New ways of analyzing variation in English*. Washington, D.C.: Georgetown University Press, pp. 23–43.

Bishop, C. M. 1995. *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bishop, Y., Fienberg, S. E., and Holland, P. W. 1977. *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

Black, E., Jelinek, F., Lafferty, J. D., Magerman, D. M., Mercer, R. L., and Roukos, S. 1993. Towards history-based grammars: Using richer models for probabilistic parsing. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 31–37.

Bloomfield, L. 1933. *Language*. Chicago: University of Chicago Press.

Bochi, S., Brossier, G., Celeux, G., Charles, C., Chifflet, R., Darcos, J., Diday, E., Diebolt, J., Fevre, P., Govaert, G., Hanani, O., Jacquet, D., Lechevallier, Y., Lemaire, J., Lemoine, Y., Molliere, J. L., Morisset, G., Ok-Sakun, Y., Rousseau, P., Sankoff, D., Schroeder, A., Sidi, J., and Taleng, F. 1980. *Optimisation en classification automatique. Tome 1, 2* (Optimization in automatic classification. Vol. 1, 2). Rocquencourt, France: Institut National de Recherche en Informatique et en Automatique (INRIA).

Bod, R. 1992. Data-Oriented Parsing. *Proceedings of COLING 1992*, Nantes, France, pp. 855–859.

Bod, R. 1993. Using an annotated corpus as a stochastic grammar. *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL 6)*, Utrecht, The Netherlands, pp. 37–44.

Bod, R. 1998. *Beyond grammar: An experience-based theory of language*. Stanford, Calif.: CSLI Publications.

Bod, R. 2000a. Combining semantic and syntactic structure for language modeling, *Proceedings of the Eighth International Conference on Spoken Language Processing (ICSLP-00)*, Beijing, pp. 298–301.

Bod, R. 2000b. The storage vs. computation of three-word sentences. Paper presented at Architectures and Mechanisms in Language Processing Conference 2000 (AMLaP '00), Leiden, The Netherlands. (Downloadable from ⟨http://staff.science.uva.nl/~rens/amlap00.ps⟩.)

Bod, R. 2001a. Sentence memory: The storage vs. computation of frequent sentences. Paper presented at the 2001 CUNY Sentence Processing Conference, Philadelphia, Pa. (Downloadable from ⟨http://staff.science.uva.nl/~rens/cuny2001.pdf⟩.)

Bod, R. 2001b. What is the minimal set of fragments that achieves maximal parse accuracy? *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 66–73.

Bod, R., and Kaplan, R. 1998. A probabilistic corpus-driven model for lexical-functional analysis. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and COLING 17 (ACL 36/COLING 17)*, Montreal, pp. 145–151.

Bod, R., and Kaplan, R. 2002. A data-oriented parsing model for lexical-functional grammar. In R. Bod, R. Scha, and K. Sima'an, eds., *Data-Oriented Parsing*. Stanford, Calif.: CSLI Publications.

Bod, R., Scha, R., and Sima'an, K., eds. 2002. *Data-Oriented Parsing*. Stanford, Calif.: CSLI Publications.

Boersma, P. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* 21, 43–58.

Boersma, P. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Boersma, P. In press. Phonology-semantics interaction in OT, and its acquisition. In R. Kirchner, W. Wikeley, and J. Pater, eds., *Papers in experimental and theoretical linguistics*. Vol. 6. Edmonton: University of Alberta.

Boersma, P., and Hayes, B. 2001. Empirical tests of the Gradual Learning Algorithm, *Linguistic Inquiry* 32, 45–86. (Downloadable from ⟨http://fonsg3.let.uva.nl/paul/⟩.)

Bonnema, R. 2002. Probability models for Data-Oriented Parsing. In R. Bod, R. Scha, and K. Sima'an, eds., *Data-Oriented Parsing*. Stanford, Calif.: CSLI Publications.

Booij, G. E. 1995. *The phonology of Dutch*. Oxford: Oxford University Press.

Booij, G. E. 1999. The syllable, views and facts. In H. van der Hulst and N. A. Ritter, eds., *Morpheme structure constraints and the phonotactics of Dutch*. Berlin: Mouton de Gruyter, pp. 53–68.

Booth, T. L. 1969. Probabilistic representation of formal languages. *IEEE Conference Record of the 1969 Tenth Annual Symposium on Switching and Automata Theory*, pp. 74–81.

Booth, T. L., and Thomson, R. A. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers* C-22, 442–450.

Brants, T. 1999. Cascaded markov models. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 9)*, Bergen, Norway.

Breiman, L. 1973. *Statistics with a view toward applications*. Boston: Houghton Mifflin.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. *Classification and regression trees*. Belmont, Calif.: Wadsworth.

Brent, M. R., and Cartwright, T. A. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61, 93–125.

Bresnan, J. 2001. *Lexical-functional syntax*. Oxford: Blackwell.

Bresnan, J., and Deo, A. 2000. "Be" in the *Survey of English dialects*: A stochastic OT account. Paper presented at the Symposium on Optimality Theory, English Linguistic Society of Japan, November 18, 2000, Kobe, Japan.

Bresnan, J., Dingare, S., and Manning, C. D. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt and T. Holloway King, eds., *Proceedings of the LFG 01 Conference*. Stanford, Calif.: CSLI Publications, pp. 13–32.

Bresnan, J., and Moshi, L. 1990. Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry* 21, 147–186.

Bresnan, J., and Zaenen, A. 1990. Deep unaccusativity in LFG. In K. Dziwirek, P. Farrell, and E. Mejías-Bikandi, eds., *Grammatical relations: A cross-theoretical perspective*. Stanford, Calif.: CSLI Publications, pp. 45–57.

Briggs, C. 1986. *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge: Cambridge University Press.

Briscoe, T. 1999. Grammatical acquisition and linguistic selection. Draft for Briscoe 2002.

Briscoe, T. 2000. An evolutionary approach to (logistic-like) language change. Manuscript, University of Cambridge.

Briscoe, T. 2002. Grammatical acquisition and linguistic selection. In T. Briscoe, ed., *Linguistic evolution through language acquisition: Formal and computational models*. Cambridge: Cambridge University Press.

Browman, C., and Goldstein, L. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3, 219–252.

Browman, C., and Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49, 155–180.

Burgess, C., and Hollbach, S. C. 1988. A computational model of syntactic ambiguity as a lexical process. *Proceedings of the 10th Annual Conference of the Cognitive Science Society (COGSCI-88)*, Budapest, pp. 263–269.

Burgess, C., and Lund, K. 1994. Multiple constraints in syntactic ambiguity resolution: A connectionist account of psycholinguistic data. *Proceedings of the 16th Annual Conference of the Cognitive Science Society (COGSCI-94)*, Atlanta, Ga., pp. 90–95.

Burzio, L. In press a. Missing players: Phonology and the past-tense debate. *Lingua*.

Burzio, L. In press b. Surface-to-surface morphology: When your representations turn into constraints. In P. Boucher, ed., *Many morphologies*. Somerville, Mass.: Cascadilla Press.

Bush, N. 1999. The predictive value of transitional probability for word-boundary palatalization in English. Master's thesis, University of New Mexico.

Butters, R. R. 1971. On the notion of "rule of grammar" in dialectology. *Papers from the Seventh Regional Meeting, Chicago Linguistic Society*, pp. 305–315.

Butters, R. R. 1972. Competence, performance, and variable rules. *Language Sciences* 20, 29–32.

Butters, R. R. 1990. Current issues in variation theory. Plenary address presented at the First International Congress of Dialectologists, University of Bamberg.

Bybee, J. L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, J. L. 1994. A view of phonology from a cognitive and functional perspective. *Cognitive Linguistics* 5, 285–305.

Bybee, J. L. 1995a. Diachronic and typological properties of morphology and their implications for representation. In L. B. Feldman, ed., *Morphological aspects of language processing*. Hillsdale, N.J.: Erlbaum, pp. 225–246.

Bybee, J. L. 1995b. Regular morphology and the lexicon. *Language and Cognitive Processes* 10, 425–435.

Bybee, J. L. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In M. Barlow and S. Kemmer, eds., *Usage-based models of language*. Stanford, Calif.: CSLI Publications, pp. 65–85.

Bybee, J. L. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. L. In press. Mechanisms of change in grammaticalization: The role of frequency. In R. D. Janda and B. D. Joseph, eds., *Handbook of historical linguistics*. Oxford: Blackwell.

Bybee, J. L., and Moder, C. L. 1983. Morphological classes as natural categories. *Language* 59, 251–270.

Bybee, J. L., and Pardo, E. 1981. Morphological and lexical conditioning of rules: Experimental evidence from Spanish. *Linguistics* 19, 937–968.

Bybee, J. L., Perkins, R., and Pagliuca, W. 1991. Back to the future. In E. C. Traugott and B. Heine, eds., *Approaches to grammaticalization*. Amsterdam: John Benjamins, pp. 17–58.

Bybee, J. L., Perkins, R., and Pagliuca, W. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.

Bybee, J. L., and Scheibman, J. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37, 575–596.

California Style Collective. 1993. Variation and personal/group style. Paper presented at New Ways of Analyzing Variation 22, University of Ottawa.

Cameron, D. 1990. De-mythologizing sociolinguistics: Why language does not reflect society. In J. Joseph and T. Taylor, eds., *Ideologies of language*. London: Routledge, pp. 79–93.

Caramazza, A., and Yeni-Komshian, G. H. 1974. Voice onset time in two French dialects. *Journal of Phonetics* 2, 239–245.

Carlson, G. 1989. On the semantic composition of English generic sentences. In G. Chierchia, B. H. Partee, and R. Turner, eds., *Properties, types and meaning*. Dordrecht: Kluwer, pp. 167–192.

Carlson, G., and Pelletier, F. J., eds. 1995. *The generic book*. Chicago: University of Chicago Press.

Carnap, R. 1950. *Logical foundations of probability*. Chicago: University of Chicago Press.

Carpenter, B. 1992. *The logic of typed feature structures*. Cambridge: Cambridge University Press.

Carpenter, B. 1999. *Type-logical semantics*. Cambridge, Mass.: MIT Press.

Carroll, J., and Weir, D. 2000. Encoding frequency information in lexicalized grammars. In H. Bunt and A. Nijholt, eds., *Advances in probabilistic and other parsing technologies*. Dordrecht: Kluwer, pp. 13–28.

Carroll, J. M., Bever, T. G., and Pollack, C. R. 1981. The non-uniqueness of linguistic intuitions. *Language* 57, 368–383.

Carterette, E., and Jones, M. H. 1974. *Informal speech*. Berkeley and Los Angeles: University of California Press.

Cedergren, H. J., and Sankoff, D. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50, 333–355.

Cena, R. M. 1978. When is a phonological generalization psychologically real? Bloomington: Indiana University Linguistics Club.

Chambers, J. K. 1995. *Sociolinguistic theory: Linguistic variation and its social significance*. Cambridge, Mass.: Blackwell.

Charniak, E. 1996. Tree-bank grammars. *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI '96)*, Menlo Park, Calif., pp. 1031–1036.

Charniak, E. 1997a. Statistical parsing with a context-free grammar and word statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI '97)*, Providence, R.I., pp. 598–603.

Charniak, E. 1997b. Statistical techniques for natural language parsing. *AI Magazine* 4, 33–43.

Charniak, E. 2000. A maximum-entropy-inspired parser. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, Seattle, Wash., pp. 132–139.

Chater, N., Crocker, M. J., and Pickering, M. J. 1998. The rational analysis of inquiry: The case of parsing. In M. Oaksford and N. Chater, eds., *Rational models of cognition*. Oxford: Oxford University Press, pp. 441–468.

Cheshire, J. 1987. Syntactic variation, the linguistic variable and sociolinguistic theory. *Linguistics* 25, 257–282.

Cheshire, J. 1991. Variation in the use of *ain't* in an urban British English dialect. In P. Trudgill and J. K. Chambers, eds., *Dialects of English: Studies in grammatical variation*. London: Longman, pp. 54–73.

Cheshire, J. 1996. Syntactic variation and the concept of prominence. In J. Klemola, M. Kytö, and M. Rissanen, eds., *Speech past and present: Studies in English dialectology in memory of Ossi Ihalainen*. Frankfurt: Peter Lang, pp. 1–17.

Cheshire, J., Edwards, V., and Whittle, P. 1989. Urban British dialect grammar: The question of dialect levelling. *English World-Wide* 10, 185–225.

Cheshire, J., Edwards, V., and Whittle, P. 1993. Non-standard English and dialect levelling. In J. Milroy and L. Milroy, eds., *Real English: The grammar of English dialects in the British Isles*. London: Longman, pp. 53–96.

Chi, Z., and Geman, S. 1998. Estimation of probabilistic context-free grammars. *Computational Linguistics* 24, 299–305.

Chiang, D. 2000. Statistical parsing with an automatically extracted tree adjoining grammar. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, Hong Kong, pp. 456–463.

Chierchia, G., and McConnell-Ginet, S. 1990. *Meaning and grammar*. Cambridge, Mass.: MIT Press.

Chierchia, G., and McConnell-Ginet, S. 2000. *Meaning and grammar*. 2nd ed. Cambridge, Mass.: MIT Press.

Chierchia, G., Partee, B. H., and Turner, R., eds. 1989. *Properties, types and meaning*. Dordrecht: Kluwer.

Chomsky, N. 1955. *The logical structure of linguistic theory*. Published 1975, New York: Plenum Press.

Chomsky, N. 1956. Three models for the description of language. *IRE Transactions on Information Theory* IT2, 113–124.

Chomsky, N. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, N. 1959. On certain formal properties of grammars. *Information and Control* 2, 137–167.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press.

Chomsky, N. 1969. Quine's empirical assumptions. In D. Davidson and J. Hintikka, eds., *Words and objections: Essays on the work of W. V. Quine*. Dordrecht: Reidel, pp. 53–68.

Chomsky, N. 1979. *Language and responsibility: Based on conversations with Mitsou Ronat*. Translated by John Viertel. New York: Pantheon.

Chomsky, N. 1981a. *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. 1981b. Principles and parameters in syntactic theory. In N. Hornstein and D. Lightfoot, eds., *Explanation in linguistics: The logical problem of language acquisition*. London: Longman, pp. 123–146.

Chomsky, N. 1986. *Barriers*. Cambridge, Mass.: MIT Press.

Chomsky, N. 1995. *The Minimalist Program*. Cambridge, Mass.: MIT Press.

Chomsky, N., and M. Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136–143.

Clahsen, H. 1999. Lexical entries and rules of language: A multi-disciplinary study of German inflection. *Behavioral and Brain Sciences* 22, 991–1060.

Clahsen, H., Eisenbeiss, S., and Sonnenstuhl-Henning, I. 1997. Morphological structure and the processing of inflected words. *Theoretical Linguistics* 23, 201–249.

Clahsen, H., and Rothweiler, M. 1992. Inflectional rules in children's grammars: Evidence from the development of participles in German. *Yearbook of Morphology 1992*, 1–34.

Clark, B. Z. 2001. A stochastic Optimality Theory approach to phrase structure variation and change in early English. Paper presented at the Berkeley Historical Syntax Workshop. Manuscript, Stanford University.

Clark, R., and Roberts, I. 1993. A computational model of language learnability and language change. *Linguistic Inquiry* 24, 299–345.

Clarkson, D., Fan, Y., and Joe, H. 1993. A remark on algorithm 643: FEXACT: An algorithm for performing Fisher's exact text in $r \times c$ contingency tables. *ACM Transactions on Mathematical Software* 19, 484–488.

Clifton, C., Jr., Frazier, L., and Connine, C. 1984. Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior* 23, 696–708.

Cohen, A. 1996. Think generic! The meaning and use of generic sentences. Ph.D. thesis, Carnegie Mellon University. Published 1999, Stanford, Calif.: CSLI Publications.

Cohen, A. 1999. Generics, frequency adverbs and probability. *Linguistics and Philosophy* 22, 221–253.

Cohen, A. 2001. Relative readings of *many*, *often*, and generics. *Natural Language Semantics* 9, 41–67.

Cohen, A. 2002. On the role of mental representations in a formal theory of meaning. Manuscript, Ben-Gurion University of the Negev.

Cohen, L. J. 1989. *An introduction to the philosophy of induction and probability*. Oxford: Clarendon Press.

Collins, M. J. 1996. A new statistical parser based on bigram lexical dependencies. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, Calif., pp. 184–191.

Collins, M. J. 1997. Three generative, lexicalised models for statistical parsing. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL 35/EACL 8)*, Madrid, pp. 16–23.

Collins, M. J. 1999. Head-driven statistical models for natural language parsing. Ph.D. thesis, University of Pennsylvania.

Collins, M. J. 2000. Discriminative reranking for natural language parsing. *Machine Learning: Proceedings of the Seventeenth International Conference (ICML 2000)*, Stanford, Calif., pp. 175–182.

Collins, M. J., and Duffy, N. 2001. Convolution kernels for natural language. *Proceedings of Neural Information Processing Systems 2001 (NIPS 2001)*, Vancouver.

Collins, M. J., and Duffy, N. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pa.

Connine, C., Ferreira, F., Jones, C., Clifton, C., and Frazier, L. 1984. Verb frame preference: Descriptive norms. *Journal of Psycholinguistic Research* 13, 307–319.

Copeland, B., and Proudfoot, D. 1999. Alan Turing's forgotten ideas in computer science. *Scientific American* 280, 76–81.

Corley, S., and Crocker, M. W. 1996. Evidence for a tagging model of human lexical category disambiguation. *Proceedings of the 18th Annual Conference of the Cognitive Science Society (COGSCI-96)*, pp. 272–277.

Corley, S., and Crocker, M. W. 2000. The modular statistical hypothesis: Exploring lexical category ambiguity. In M. W. Crocker, M. Pickering, and C. Clifton, eds., *Architectures and mechanisms for language processing*. Cambridge: Cambridge University Press, pp. 135–160.

Cowan, J. 1979. *Hans Wehr: A dictionary of modern written Arabic*. Wiesbaden, Germany: Otto Harrassowitz.

Cowart, W. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, Calif.: Sage Publications.

Crocker, M. W., and Brants, T. 2000. Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research* 29, 647–669.

Cuetos, F., Mitchell, D. C., and Corley, M. M. B. 1996. Parsing in different languages. In M. Carreiras, J. E. García-Albea, and N. Sebastián-Gallés, eds., *Language processing in Spanish*. Hillsdale, N.J.: Erlbaum, pp. 156–187.

Culy, C. 1998. Statistical distribution and the grammatical/ungrammatical distinction. *Grammars* 1, 1–13.

Cutler, A. 1981. Degrees of transparency in word formation. *Papers from the 26th Regional Meeting, Chicago Linguistic Society*, pp. 73–77.

Cutler, A., and Butterfield, S. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language* 31, 218–236.

Cutler, A., and Otake, T. 1996. Phonological contrasts and their role in processing. In T. Otake and A. Cutler, eds., *Phonological structure and language processing: Cross-linguistic studies*. Berlin: Mouton de Gruyter, pp. 1–12.

Cutler, A., and Otake, T. 1998. Assimilation of place in Japanese and Dutch. *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP-98)*, pp. 1751–1754.

Daelemans, W., Zavrel, J., van der Sloot, K., and van den Bosch, A. 2000. TiMBL: Tilburg memory based learner reference guide 3.0. Report 00-01. Tilburg, The Netherlands: Tilburg University, Computational Linguistics.

Daelemans, W., van den Bosch, A., and Weijters, T. 1997. IGTree: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review* 11, 407–423.

d'Arcais, G. B. F. 1993. The comprehension and semantic interpretation of idioms. In C. Cacciari and P. Tabossi, eds., *Idioms: Processing, structure, and interpretation*. Hillsdale, N.J.: Erlbaum, pp. 79–98.

Darroch, J. N., and Ratcliff, D. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics* 43, 1470–1480.

Daugherty, K., and Seidenberg, M. 1994. Beyond rules and exceptions: A connectionist approach to inflectional morphology. In S. Lima, R. Corrigan, and G. Iverson, eds., *The reality of linguistic rules*. Amsterdam: John Benjamins, pp. 358–388.

Davis, S. 1991. Coronals and the phonotactics of nonadjacent consonants in English. *Phonetics and Phonology* 2, 49–60.

De Boer, B. 2000. Self-organization in vowel systems. *Journal of Phonetics* 28, 441–466.

DeGroot, M. 1989. *Probability and statistics*. Reading, Mass.: Addison-Wesley.

de Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., and Baayen, R. H. 2002. The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. *Brain and Language* 81, 555–567.

de Jong, N. H., Schreuder, R., and Baayen, R. H. 2000. The morphological family size effect and morphology. *Language and Cognitive Processes* 15, 329–365.

Delgrande, J. P. 1987. A first-order conditional logic for prototypical properties. *Artificial Intelligence* 33, 105–130.

Dell, G. S. 1986. A spreading activation theory of retrieval in sentence production. *Psychological Review* 93, 283–321.

Dell, G. S. 1990. Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes* 5, 313–349.

Dell, G. S., Schwarts, M. F., Martin, N., Saffran, E., and Gagnon, D. A. 1997. Lexical access in aphasic and nonaphasic speakers. *Psychological Review* 104, 801–838.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39, 1–38.

Desmet, T., Brysbaert, M., and Baecke, C. D. In press. The correspondence between sentence production and corpus frequencies in modifier attachment. *Quarterly Journal of Experimental Psychology*.

de Swart, H. 1991. Adverbs of quantification: A generalized quantifier approach. Ph.D. thesis, Groningen University. Published 1993, New York: Garland.

de Swart, H. 1998. *Introduction to natural language semantics*. Stanford, Calif.: CSLI Publications.

Dingare, S. 2001. The effect of feature hierarchies on frequencies of passivization in English. Master's thesis, Stanford University.

Docherty, G., and Foulkes, P. 2000. Speaker, speech and knowledge of sounds. In P. C. N. Burton-Roberts and G. Docherty, eds., *Phonological knowledge: Conceptual and empirical issues*. Oxford: Oxford University Press, pp. 105–129.

Dowty, D. R., Wall, R. E., and Peters, S. 1981. *Introduction to Montague semantics*. Dordrecht: Reidel.

Dras, M., Harrison, K. D., and Kapicioglu, B. 2001. Agent-based simulation and microparametric variation: Modeling the evolution of vowel harmony. Paper presented at the 32nd Meeting of the North East Linguistic Society (NELS 32), City University of New York and New York University.

Dressler, W. 1985. On the predictiveness of natural morphology. *Journal of Linguistics* 21, 321–337.

Duarte, D., Galves, A., Garcia, N. L., and Maronna, R. 2001. The statistical analysis of acoustic correlates of speech rhythm. Paper presented at Zentrum für interdisciplinäre Forschung, University of Bielefeld. (Downloadable from ⟨http://www.physik.uni-bielefeld.de/complexity/duarte.pdf⟩.)

Eckert, P. 1989. The whole woman: Sex and gender differences in variation. *Language Variation and Change* 1, 245–267.

Eckert, P. 1999. *Linguistic variation as social practice*. Oxford: Blackwell.

Eckert, P., and McConnell-Ginet, S. 1992. Think practically and look locally. *Annual Review of Anthropology* 21, 461–490.

Eddington, D. 1996. Diphthongization in Spanish derivational morphology: An empirical investigation. *Hispanic Linguistics* 8, 1–13.

Edgington, D. 1991. The mystery of the Missing Matter of Fact. *Proceedings of the Aristotelian Society, supplementary volume 65*, pp. 185–209.

Edgington, D. 1996. Vagueness by degrees. In R. Keefe and P. Smith, eds., *Vagueness*. Cambridge, Mass.: MIT Press, pp. 294–316.

Eells, E., and Skyrms, B., eds. 1994. *Probability and conditionals: Belief revision and rational decision*. Cambridge: Cambridge University Press.

Eisner, J. 1996. Three new probabilistic models for dependency parsing: An exploration. *Proceedings of COLING 1996*, Copenhagen, pp. 340–345.

Ellegård, A. 1953. *The auxiliary* do: *The establishment and regulation of its use in English*. Stockholm: Almqvist and Wiksell.

Embleton, S. 1986. Statistics in historical linguistics. *Quantitative Linguistics* 30.

Engdahl, E. 1983. Parasitic gaps. *Linguistics and Philosophy* 6, 5–34.

Engstrand, O., and Krull, D. 1994. Durational correlates of quantity in Swedish, Finnish, and Estonian: Cross-linguistic evidence for a theory of adaptive dispersion. *Phonetica* 51, 80–91.

Ernestus, M., and Baayen, R. H. 2001. Choosing between the Dutch past-tense suffixes *-te* and *-de*. In T. van der Wouden and H. de Hoop, eds., *Linguistics in the Netherlands*. Amsterdam: John Benjamins, pp. 81–93.

Ernestus, M., and Baayen, R. H. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*. Forthcoming.

Ernestus, M., and Baayen, R. H. 2002. Analogical effects in regular past-tense formation in Dutch. *Linguistics*.

Estival, D., and Myhill, J. 1988. Formal and functional aspects of the development from passive to ergative systems. In M. Shibatani, ed., *Passive and voice*. Amsterdam: John Benjamins, pp. 441–491.

Evelyn, J. 1664. *Sylva, or A discourse of forest-trees, and the propagation of timber in His Majesties dominions, &c.* Printed by Jo. Martyn and Ja. Allestry, for the Royal Society, London.

Faber, A., and DiPaolo, M. 1995. The discriminability of nearly merged sounds. *Language Variation and Change* 7, 35–78.

Fasold, R. 1991. The quiet demise of variable rules. *American Speech* 66, 3–21.

Feagin, C. 2001. Entering the community: Fieldwork. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes, eds., *The handbook of language variation and change*. Oxford: Blackwell, pp. 20–40.

Feldman, J. A., Gips, J., Horning, J. J., and Reder, S. 1969. Grammatical complexity and inference. Technical report CS 125. Stanford, Calif.: Stanford University, Computer Science Department.

Feller, W. 1970. *An introduction to probability theory and its applications*. New York: Wiley.

Fernando, T., and Kamp, H. 1996. Expecting many. *Proceedings of the 6th Conference on Semantics and Linguistic Theory*, pp. 53–68.

Fidelholz, J. 1975. Word frequency and vowel reduction in English. *Papers from the 13th Regional Meeting, Chicago Linguistic Society*, pp. 200–213.

Fienberg, S. E. 1980. *The analysis of cross-classified categorical data*. 2nd ed. Cambridge, Mass.: MIT Press.

Filip, H., Tanenhaus, M. K., Carlson, G. N., Allopenna, P. D., and Blatt, J. 2002. Reduced relatives judged hard require constraint-based analyses. In P. Merlo and S. Stevenson, eds., *Sentence processing and the lexicon: Formal, computational, and experimental perspectives*. Amsterdam: John Benjamins, pp. 255–280.

Fillmore, C. J. 1992. "Corpus linguistics" or "computer-aided armchair linguistics." In J. Svartvik, ed., *Directions in corpus linguistics: Proceedings of Nobel Symposium 82: Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter, pp. 35–60.

Fisher, R. A. 1959. *Statistical methods and scientific inference*. 2nd ed. Edinburgh: Oliver and Boyd.

Flege, J. E., and Hillenbrand, J. 1986. Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. *Journal of the Acoustical Society of America* 79, 508–517.

Flemming, E. 1995. Auditory phonology. Ph.D. thesis, University of California, Los Angeles.

Flemming, E. 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18, 7–44.

Fodor, J. A. 1983. *Modularity of mind*. Cambridge, Mass.: MIT Press.

Fodor, J. D. 1978. Parsing strategies and constraints on transformations. *Linguistic Inquiry* 9, 427–473.

Fontaine, C. 1985. Application de méthodes quantitatives en diachronie: L'inversion du sujet en français. Ph.D. thesis, Université du Québec, Montréal.

Ford, M., Bresnan, J., and Kaplan, R. M. 1982. A competence-based theory of syntactic closure. In J. Bresnan, ed., *The mental representation of grammatical relations*. Cambridge, Mass.: MIT Press, pp. 727–796.

Forster, K. 1994. Computational modeling and elementary process analysis in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance* 20, 1292–1310.

Forster, K. 2000. Computational modeling and elementary process analysis in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance* 20, 1292–1310.

Forster, K., and Chambers, S. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* 12, 627–635.

Fowler, C., and Housum, J. 1987. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26, 489–504.

Fowler, G. H. 1987. The syntax of genitive case in Russian. Ph.D. thesis, University of Chicago.

Fowler, H. W. 1926. *A dictionary of modern English usage*. Oxford: Clarendon Press.

Francescotti, R. M. 1995. Even: The conventional implicature approach reconsidered. *Linguistics and Philosophy* 18, 153–173.

Francis, W. N., and Kučcera, H. 1982. *Frequency analysis of English usage*. Boston: Houghton Mifflin.

Frisch, S. A. 1994. Reanalysis precedes syntactic change: Evidence from Middle English. *Studies in the Linguistic Sciences* 24, 187–201.

Frisch, S. A. 1996. Similarity and frequency in phonology. Ph.D. thesis, Northwestern University.

Frisch, S. A., Large, N. R., and Pisoni, D. B. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42, 481–496.

Frisch, S. A., Large, N. R., Zawaydeh, B. A., and Pisoni, D. B. 2001. Emergent phonotactic generalizations. In J. L. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 159–180.

Frisch, S. A., and Zawaydeh, B. A. 2001. The psychological reality of OCP-place in Arabic. *Language* 77, 91–106.

Frost, R., Forster, K. I., and Deutsch, A. 1997. What can we learn from the morphology of Hebrew? A masker-priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory and Cognition* 23, 829–856.

Fu, K. S. 1974. *Syntactic methods in pattern recognition*. New York: Academic Press.

Gahl, S. 1999. Unergative, unaccusative, (in)transitive and (in)frequent. *Papers from the 34th Regional Meeting, Chicago Linguistic Society*, pp. 185–193.

Gahl, S., Menn, L., Ramsberger, G., Jurafsky, D., Elder, D., and Rewega, M. In press. Syntactic frame and verb bias in aphasia. *Brain and Cognition*.

Gal, S. 1978. Variation and change in patterns of speaking: Language use in Austria. In D. Sankoff, ed., *Linguistic variation: Models and methods*. New York: Academic Press, pp. 227–238.

Gamut, L. T. F. 1991. *Logic, language, and meaning*. Chicago: University of Chicago Press.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language* 37, 58–93.

Gazdar, G. 1976. Quantifying context. *York papers in linguistics*, pp. 177–133.

Gibson, E. 1991. A computational theory of human linguistic processing: Memory limitations and processing breakdown. Ph.D. thesis, Carnegie Mellon University.

Gibson, E. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68, 1–76.

Gibson, E., Schütze, C., and Salomon, A. 1996. The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research* 25, 59–92.

Gillund, G., and Shiffrin, R. M. 1984. A retrieval model for both recognition and recall. *Psychological Review* 91, 1–67.

Givón, T. 1979. *On understanding grammar*. New York: Academic Press.

Givón, T. 1994. The pragmatics of de-transitive voice: Functional and typological aspects of inversion. In T. Givón, ed., *Voice and inversion*. Amsterdam: John Benjamins, pp. 3–44.

Gleason, H. A. 1961. *An introduction to descriptive linguistics*. Rev. ed. New York: Holt, Rinehart and Winston.

Glymour, C., and Cheng, P. W. 1998. Causal mechanism and probability: A normative approach. In M. Oaksford and N. Chater, eds., *Rational models of cognition*. Oxford: Oxford University Press, pp. 296–313.

Godfrey, J., Holliman, E., and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing (IEEE ICASSP-92)*, San Francisco, pp. 517–520.

Gold, E. M. 1967. Language identification in the limit. *Information and Control* 10, 447–474.

Goldinger, S. D. 1997. Words and voices: Perception and production in an episodic lexicon. In K. Johnson and J. Mullenix, eds., *Talker variability in speech processing*. San Diego, Calif.: Academic Press, pp. 33–66.

Goldinger, S. D. 2000. The role of perceptual episodes in lexical processing. In A. Cutler, J. M. McQueen, and R. Zondervan, eds., *Proceedings of SWAP (Spoken Word Access Processes)*. Nijmegen: Max Planck Institute for Psycholinguistics, pp. 155–159.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237–264.

Goodman, J. 1998. Parsing inside-out. Ph.D. thesis, Harvard University.

Goodman, J. 2002. PCFG reductions of Data-Oriented Parsing. In R. Bod, R. Scha, and K. Sima'an, eds., *Data-Oriented Parsing*. Stanford, Calif.: CSLI Publications.

Goodsitt, J., Morgan, J., and Kuhl, P. 1993. Perceptual strategies in prelingual speech segmentation. *Journal of Child Language* 20, 229–252.

Gopnik, A., Meltzoff, A., and Kuhl, P. K. 1999. *The scientist in the crib*. New York: Morrow.

Gowers, E. 1948. *Plain words: A guide to the use of English*. London: Her Majesty's Stationery Office.

Grainger, J. 1990. Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language* 29, 228–244.

Grainger, J., and Jacobs, A. M. 1996. Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review* 103, 518–565.

Greenberg, J. 1966. *Language universals: With special reference to feature hierarchies*. The Hague: Mouton.

Greenberg, J. H. 1987. *Language in the Americas*. Stanford, Calif.: Stanford University Press.

Greenberg, J. H., and Jenkins, J. J. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20, 157–177.

Greenberg, J. H., and Ruhlen, M. 1992. Linguistic origins of Native Americans. *Scientific American* 267, 94–99.

Greenberg, S., Ellis, D., and Hollenback, J. 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceedings of the International Conference on Spoken Language Processing (ICSLP-96)*, Philadelphia, Pa., pp. S24–27.

Grefenstette, G. 2000. Very large lexicons from the WWW. Paper presented at the University of Brighton, U.K.

Gregory, M. L. 2001. Linguistic informativeness and speech production: An investigation of contextual and discourse-pragmatic effects on phonological variation. Ph.D. thesis, University of Colorado, Boulder.

Gregory, M., Raymond, W., Fosler-Lussier, E., and Jurafsky, D. 2000. The effects of collocational strength and contextual predictability in lexical production. *Papers from the 35th Regional Meeting, Chicago Linguistic Society*, pp. 151–166.

Grenander, U. 1967. Syntax controlled probabilities. Technical report. Providence, R.I.: Brown University, Division of Applied Mathematics.

Grimshaw, J. 1979. Complement selection and the lexicon. *Linguistic Inquiry* 10, 279–326.

Grimshaw, J. 1990. *Argument structure*. Cambridge, Mass.: MIT Press.

Grimshaw, J., and Vikner, S. 1993. Obligatory adjuncts and the structure of events. In E. Reuland and W. Abraham, eds., *Knowledge and language*. Vol. II. Dordrecht: Kluwer, pp. 143–155.

Grishman, R., Macleod, C., and Meyers, A. 1994. COMLEX syntax: Building a computational lexicon. *Proceedings of COLING 1994*, Kyoto, 268–272.

Grodzinsky, Y. 2000. The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences* 23, 47–117.

Grosjean, F. 1980. Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics* 28, 267–283.

Guy, G. 1980. Variation in the group and the individual. In W. Labov, ed., *Locating language in time and space*. New York: Academic Press, pp. 1–36.

Guy, G. 1981. Linguistic variation in Brazilian Portuguese: Aspects of the phonology, syntax, and language history. Ph.D. thesis, University of Pennsylvania.

Guy, G. 1997. Violable is variable: OT and linguistic variation. *Language Variation and Change* 9, 333–348.

Guy, J. 1980a. *Experimental glottochronology: Basic methods and results*. Canberra: Australian National University, Research School of Pacific Studies.

Guy, J. 1980b. *Glottochronology without cognate recognition*. Canberra: Australian National University, Research School of Pacific Studies.

Hacking, I. 1975. *The emergence of probability: A philosophical study of early ideas about probability*. Cambridge: Cambridge University Press.

Haeri, N. 1998. *The sociolinguistic market of Cairo: Gender, class and education*. London: Kegan Paul.

Haiman, J. 1994. Ritualization and the development of language. In W. Pagliuca, ed., *Perspectives on grammaticalization*. Amsterdam: John Benjamins, pp. 3–28.

Hájek, A., and Hall, N. 1994. The hypothesis of the conditional construal of conditional probability. In E. Eells and B. Skyrms, eds., *Probability and conditionals: Belief revision and rational decision*. Cambridge: Cambridge University Press, pp. 75–111.

Hale, J. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, Philadelphia, Pa., pp. 159–166.

Hale, K., and Keyser, S. J. 1993. On argument structure and the lexical expression of syntactic relations. In K. Hale and S. J. Keyser, eds., *The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger*. Cambridge, Mass.: MIT Press, pp. 53–109.

Halliday, M. A. K. 1994. *An introduction to functional grammar*. 2nd ed. London: Edward Arnold.

Hare, M. L., Ford, M., and Marslen-Wilson, W. D. 2001. Ambiguity and frequency effects in regular verb inflection. In J. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 181–200.

Harris, T. E. 1963. *The theory of branching processes*. Berlin: Springer-Verlag.

Hasher, L., and Zacks, R. T. 1984. Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist* 39, 1372–1388.

Hastie, T., Tibshirani, R., and Friedman, J. H. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

Hay, J. B. 2000. Causes and consequences of word structure. Ph.D. thesis, Northwestern University. (Downloadable from ⟨http://www.ling.canterbury.ac.nz/jen⟩. Summary to appear in GLOT.)

Hay, J. B., and Baayen, R. H. 2002. Parsing and productivity. In G. E. Booij and J. van Marle, eds., *Yearbook of morphology 2001*. Dordrecht: Kluwer, pp. 203–235.

Hay, J. B., Jannedy, S., and Mendoza-Denton, N. 1999. Oprah and /ay/: Lexical frequency, referee design, and style. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, pp. 1389–1392.

Hay, J. B., Mendoza-Denton, N., and Jannedy, S. 2000. Style and economy: When frequency effects collide. Paper presented at the 7th New Zealand Language and Society Conference, Auckland, New Zealand.

Hay, J. B., Pierrehumbert, J., and Beckman, M. E. In press. Speech perception, wellformedness, and the statistics of the lexicon. In J. Local, R. Ogden, and R. Temple, eds., *Papers in laboratory phonology VI*. Cambridge: Cambridge University Press. (Downloadable from ⟨http://www.ling.canterbury.ac.nz/jen⟩.)

Hazen, K. 2001. An introductory investigation into bidialectalism. In *Selected papers from NWAV 29*. Philadelphia: University of Pennsylvania, Penn Linguistics Club.

Hazen, V., and Barrett, S. 2000. The development of phonemic categorization in children aged 6–12. *Journal of Phonetics* 24, 377–396.

Heemskerk, J. 1993. A probabilistic context-free grammar for disambiguation in morphological parsing. *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL 6)*, Utrecht, The Netherlands, pp. 183–192.

Heim, I., and Kratzer, A. 1998. *Semantics in generative grammar*. Oxford: Blackwell.

Hoogweg, L. 2002. Extending DOP with the insertion operation. In R. Bod, R. Scha, and K. Sima'an, eds., *Data-Oriented Parsing*. Stanford, Calif.: CSLI Publications.

Hooper, J. B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie, ed., *Current progress in historical linguistics*. Amsterdam: North Holland, pp. 96–105.

Hopper, P., and Traugott, E. C. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Hornby, A. S. 1989. *Oxford advanced learner's dictionary of current English*. 4th ed. Oxford: Oxford University Press.

Horning, J. 1969. A study of grammatical inference. Ph.D. thesis, Stanford University.

Horvath, B. 1985. *Variation in Australian English: The sociolects of Sydney*. New York: Cambridge University Press.

Howes, D. 1957. On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America* 29, 296–305.

Howes, D. H., and Solomon, R. L. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology* 41, 401–410.

Hukari, T. E., and Levine, R. D. 1995. Adjunct extraction. *Journal of Linguistics* 31, 195–226.

Hurford, J., Studdert-Kennedy, M., and Knight, C. 1998. *Approaches to the evolution of language: Social and cognitive bases*. Cambridge: Cambridge University Press.

Hymes, D. 1974. *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Ihalainen, O. 1991. On grammatical diffusion in Somerset folk speech. In P. Trudgill and J. K. Chambers, eds., *Dialects of English: Studies in grammatical variation*. London: Longman, pp. 104–119.

Ince, D. C., ed. 1992. *Mechanical intelligence*. Elsevier Science.

Itkonen, E. 1983. The use of a corpus in the study of syntax: Historical and methodological considerations. Technical report.

Jackendoff, R. S. 1975. Morphological and semantic regularities in the lexicon. *Language* 51, 639–671.

Jaeger, J. J., Lockwood, A. H., Kemmerrer, D. L., Van Valin, R. D., and Murphy, B. W. 1996. A positron emission tomographic study of regular and irregular verb morphology in English. *Language* 72, 451–497.

Jelinek, E., and Demers, R. 1983. The agent hierarchy and voice in some Coast Salish languages. *International Journal of American Linguistics* 49, 167–185.

Jelinek, E., and Demers, R. 1994. Predicates and pronominal arguments in Straits Salish. *Language* 70, 697–736.

Jelinek, F., and Lafferty, J. D. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics* 17, 315–323.

Jennings, F., Randall, B., and Tyler, L. K. 1997. Graded effects on verb subcategory preferences on parsing: Support for constraint-satisfaction models. *Language and Cognitive Processes* 12, 485–504.

Jensen, F. V., and Jensen, F. B. 2001. *Bayesian networks and decision graphs*. Berlin: Springer-Verlag.

Jescheniak, J. D., and Levelt, W. J. M. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition* 20, 824–843.

Johnson, K. 1997a. *Acoustic and auditory phonetics*. Cambridge, Mass.: Blackwell.

Johnson, K. 1997b. The auditory/perceptual basis for speech segmentation. In K. Ainsworth-Darnell and M. D'Imperio, eds., *Papers from the linguistics laboratory*. Ohio State University Working Papers in Linguistics 50. Columbus: Ohio State University, Department of Linguistics, pp. 101–113.

Johnson, K. 1997c. Speech perception without speaker normalization. In K. Johnson and J. Mullenix, eds., *Talker variability in speech processing*. San Diego, Calif.: Academic Press, pp. 146–165.

Johnson, M. 1998a. Optimality-theoretic Lexical Functional Grammar. Commentary on Joan Bresnan's presentation at the 11th Annual CUNY Conference on Human Sentence Processing 1998. (Downloadable from ⟨http://www.cog.brown.edu/~lowmj/papers/cuny98.pdf⟩.)

Johnson, M. 1998b. PCFG models of linguistic tree representations. *Computational Linguistics* 24, 613–632.

Johnson, M. 2002. The DOP1 estimation method is biased and inconsistent. *Computational Linguistics* 28, 71–76.

Johnson, M., Geman, S., Canon, S., Chi, Z., and Riezler, S. 1999. Estimators for stochastic "unification-based" grammars. *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics*, College Park, Md., pp. 535–541.

Joos, M. 1950. Description of language design. *Journal of the Acoustical Society of America* 22, 701–708.

Jordan, M. I., ed. 1999. *Learning in graphical models*. Cambridge, Mass.: MIT Press.

Joshi, A. K., and Schabes, Y. 1997. Tree-adjoining grammars. In G. Rozenberg and A. Salomaa, eds., *Handbook of formal languages*. Berlin: Springer-Verlag, pp. 69–123.

Juliano, C., and Tanenhaus, M. K. 1993. Contingent frequency effects in syntactic ambiguity resolution. *Proceedings of the 15th Annual Conference of the Cognitive Science Society (COGSCI-94)*, Boulder, Colo., pp. 593–598.

Jurafsky, D. 1992. An on-line computational model of human sentence interpretation: A theory of the representation and use of linguistic knowledge. Ph.D. thesis, University of California, Berkeley. (Available as technical report 92/676. Berkeley: University of California, Computer Science Division.)

Jurafsky, D. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20, 137–194.

Jurafsky, D. 2002. Pragmatics and computational linguistics. In L. R. Horn and G. Ward, eds., *Handbook of pragmatics*. Oxford: Blackwell.

Jurafsky, D., Bell, A., and Girand, C. 2002. The role of the lemma in form variation. In C. Gussenhoven and N. Warner, eds., *Laboratory phonology VII*. Berlin: Mouton de Gruyter. (Downloadable from ⟨http://www.colorado.edu/linguistics/jurafsky/pubs.html⟩.)

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 229–254.

Jurafsky, D., and Martin, J. H. 2000. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall.

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* 33, 630–645.

Kamp, H., and Reyle, U. 1993. *From discourse to logic*. Dordrecht: Kluwer.

Kamp, J. A. W. 1975. Two theories about adjectives. In E. L. Keenan, ed., *Formal semantics of natural language*. Cambridge: Cambridge University Press, pp. 123–155.

Kato, K. 1979. Empathy and passive resistance. *Linguistic Inquiry* 10, 149–152.

Kay, P. 1990. Even. *Linguistics and Philosophy* 13, 59–111.

Kay, P., and McDaniel, C. 1979. On the logic of variable rules. *Language and Society* 8, 151–187.

Keenan, E. L., ed. 1975. *Formal semantics of natural language*. Cambridge: Cambridge University Press.

Keenan, E. L., and Stavi, J. 1986. A semantic characterization of natural language determiners. *Linguistics and Philosophy* 9, 253–326.

Kegl, J. 1995. Levels of representation and units of access relevant to agrammatism. *Brain and Language* 50, 151–200.

Keller, F. 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Ph.D. thesis, University of Edinburgh.

Kent, R., and Forner, L. 1980. Speech segment durations in sentence recitations by children and adults. *Journal of Phonetics* 8, 157–168.

Kersten, D. 1999. High-level vision as statistical inference. In M. S. Gazzaniga, ed., *The new cognitive neurosciences*. Cambridge, Mass.: MIT Press.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL 7)*, Dublin, pp. 60–66.

Kessler, B. 2001. The significance of word lists. Technical report. Stanford, Calif.: Stanford University, CSLI.

Kim, A., Srinivas, B., and Trueswell, J. 2002. The convergence of lexicalist perspectives in psycholinguistics and computational linguistics. In P. Merlo and S. Stevenson, eds., *Sentence processing and the lexicon: Formal, computational, and experimental perspectives*. Amsterdam: John Benjamins, pp. 109–136.

Kintsch, W. 1970. Models for free recall and recognition. In D. A. Normal, ed., *Models of human memory*. New York: Academic Press, pp. 334–374.

Kiparsky, P. 1973. "Elsewhere" in phonology. In S. R. Anderson, ed., *A festschrift for Morris Halle*. New York: Holt, Rinehart and Winston, pp. 93–106.

Kirby, S. 1999. *Function, selection, and innateness: The emergence of language universals*. Oxford: Oxford University Press.

Kirby, S. 2001. Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* 5, 102–110.

Kirchner, R. 1997. Contrastiveness and faithfulness. *Phonology* 14, 83–111.

Kirchner, R. In press. Preliminary thoughts on "phonologisation" within an exemplar-based speech processing model. In *UCLA working papers in linguistics 6*. Los Angeles: University of California at Los Angeles, Department of Linguistics.

Kisseberth, C. 1970. On the functional unity of phonological rules. *Linguistic Inquiry* 1, 291–306.

Knight, C., Studdert-Kennedy, M., and Hurford, J. 2000. *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge: Cambridge University Press.

Knoke, D., and Bohrnstedt, G. W. 1994. *Statistics for social data analysis*. 3rd ed. Itaska, Ill.: Peacock.

Kontra, M., and Váradi, T. 1997. The Budapest sociolinguistic interview: Version 3. In *Working papers in Hungarian sociolinguistics*.

Koontz-Garboden, A. J. 2001. A stochastic OT approach to word order variation in Korlai Portuguese. *Papers from the 37th Regional Meeting, Chicago Linguistic Society*. Vol. 1, pp. 347–361.

Koopman, H., and Sportiche, D. 1987. Variables and the bijection principle. *The Linguistic Review* 2, 139–160.

Kornai, A. 1998. Analytic models in phonology. In J. Durand and B. Laks, eds., *The organization of phonology: Constraints, levels and representations*. Oxford: Oxford University Press, pp. 395–418. (Downloadable from ⟨http://www.kornai.com/pub.html⟩.)

Kratzer, A. 1981. The notional category of modality. In H. J. Eikmeyer and H. Rieser, eds., *Words, worlds, and contexts: New approaches to word semantics*. Berlin: de Gruyter, pp. 38–74.

Krenn, B., and Samuelsson, C. 1997. The linguist's guide to statistics. Manuscript, University of Saarbrücken.

Krifka, M. 1995. Focus and the interpretation of generic sentences. In G. Carlson and F. J. Pelletier, eds., *The generic book*. Chicago: University of Chicago Press, pp. 238–264.

Kroch, A. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1, 199–244.

Kroch, A. S. 2001. Syntactic change. In M. Baltin and C. Collins, eds., *Handbook of contemporary syntactic theory*. Oxford: Blackwell, pp. 699–729.

Kroeger, P. 1993. *Phrase structure and grammatical relations in Tagalog*. Stanford, Calif.: CSLI Publications.

Krott, A., Baayen, R. H., and Schreuder, R. 2001. Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics* 39, 51–93.

Krott, A., Schreuder, R., and Baayen, R. H. 2002. Linking elements in Dutch noun-noun compounds: Constituent families as predictors for response latencies. *Brain and Language* 81, 708–722.

Krug, M. 1998. String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics* 26, 286–320.

Kruschke, J. K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* 99, 22–44.

Kučera, H., and Francis, W. N. 1967. *Computational analysis of present-day American English*. Providence, R.I.: Brown University Press.

Kuhn, J. 2001a. Formal and computational aspects of optimality-theoretic syntax. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Kuhn, J. 2001b. Sketch of a hybrid optimization architecture of grammar in context. Manuscript, Stanford University.

Kuno, S., and Kaburaki, E. 1977. Empathy and syntax. *Linguistic Inquiry* 8, 627–672.

Labov, W. 1966. *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov, W. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762.

Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. 1975. *What is a linguistic fact?* Lisse, Belgium: Peter de Ridder Press.

Labov, W. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, 205–254.

Labov, W. 1994. *Principles of linguistic change*. Vol. 1, *Internal factors*. Cambridge, Mass.: Blackwell.

Labov, W. 2001. *Principles of linguistic change*. Vol. 2, *Social factors*. Malden, Mass.: Blackwell.

Labov, W., Ash, S., and Boberg, C. In press. *The atlas of North American English*. Berlin: Mouton de Gruyter.

Labov, W., Karen, M., and Miller, C. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3, 33–74.

Ladefoged, P., and Maddieson, I. 1996. *The sounds of the world's languages*. Oxford: Blackwell.

Lance, M. 1991. Probabilistic dependence among conditionals. *Philosophical Review* 100, 269–276.

Landman, F. 1989. Groups, II. *Linguistics and Philosophy* 12, 723–744.

Lappin, S. 1988. The semantics of *many* as a weak quantifier. *Linguistics* 26, 1021–1037.

Lappin, S. 1993. ||Many|| as a two-place determiner function. In M. Cobb and Y. Yian, eds., *SOAS working papers in linguistics and phonetics 3.* London: University of London, SOAS, pp. 337–358.

Laudanna, A., and Burani, C. 1995. Distributional properties of derivational affixes: Implications for processing. In L. B. Feldman, ed., *Morphological aspects of language processing*. Hillsdale, N.J.: Erlbaum, pp. 345–364.

Laudanna, A., Burani, C., and Cermele, A. 1994. Prefixes as processing units. *Language and Cognitive Processes* 9, 295–316.

Lavandera, B. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7, 171–182.

Lebart, L. 1995. Correspondence analysis and neural networks. *Analyses multidimensionelles des données: Troisième Congrès International NGUS'95*, pp. 27–36.

Lee, S., Potamianos, A., and Narayanan, S. 1999. Acoustics of children's speech: Developmental changes in temporal and spectral parameters. *Journal of the Acoustical Society of America* 105, 1455–1468.

Legendre, G., Miyata, Y., and Smolensky, P. 1990. Can connectionism contribute to syntax? Harmonic Grammar, with an application. *Papers from the 26th Regional Meeting, Chicago Linguistic Society*, pp. 237–252.

Legendre, G., Raymond, W., and Smolensky, P. 1993. An optimality-theoretic typology of case and grammatical voice systems. *Proceedings of the 19th Annual Meeting of the Berkeley Linguistics Society*, pp. 464–478.

Lennig, M. 1978. Acoustic measurement of linguistic change: The modern Paris vowel system. Ph.D. thesis, University of Pennsylvania.

Levelt, W. J. M. 1974. *Formal grammars in linguistics and psycholinguistics*. Vol. 1. The Hague: Mouton.

Levelt, W. J. M. 1989. *Speaking: From intention to articulation*. Cambridge, Mass.: MIT Press.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22, 1–75.

Levin, B. 1993. *English verb classes and alternations*. Chicago: University of Chicago Press.

Lewis, D. 1975. Adverbs of quantification. In E. L. Keenan, ed., *Formal semantics of natural language*. Cambridge: Cambridge University Press, pp. 3–15.

Lewis, D. 1976. Probabilities of conditionals and conditional probabilities. *Philosophical Review* 85, 297–315.

Li, A. Y.-H. 1990. *Order and constituency in Mandarin Chinese*. Dordrecht: Kluwer.

Li, P., and Yip, M. C. 1996. Lexical ambiguity and context effects in spoken word recognition: Evidence from Chinese. *Proceedings of the 18th Annual Conference of the Cognitive Science Society (COGSCI-96)*, pp. 228–232.

Lightfoot, D. 1991. *How to set parameters: Evidence from language change*. Cambridge, Mass.: MIT Press.

Liljencrants, J., and Lindblom, B. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48, 839–861.

Lindblom, B. 1986. Phonetic universals in vowel systems. In J. Ohala and J. Jaeger, eds., *Experimental phonology*. Orlando, Fla.: Academic Press, pp. 1773–1781.

Lindblom, B. 1990. Models of phonetic variation and selection. *PERILUS* 11, 65–100.

Lippi-Green, R. 1997. The real trouble with Black English. In R. Lippi-Green, ed., *English with an accent: Language ideology and discrimination in the United States*. London: Routledge, pp. 176–202.

Lloyd, C. J. 1999. *Statistical analysis of categorical data*. New York: Wiley.

Lødrup, H. 1999. Linking and optimality in the Norwegian presentational focus construction. *Nordic Journal of Linguistics* 22, 205–230.

Lowe, E. 1991. Noun phrases, quantifiers and generic names. *The Philosophical Quarterly* 41(164), 287–300.

Luce, P. A., and Pisoni, D. B. 1990. Recognizing spoken words: The neighborhood activation model. In G. T. M. Altmann, ed., *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge, Mass.: MIT Press, pp. 122–147.

Luce, P. A., and Pisoni, D. B. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19, 1–36.

Luce, R. D. 1959. *Individual choice behavior*. New York: Wiley.

Luce, R. D., Bush, R. R., and Galanter, E., eds. 1963a. *Handbook of mathematical psychology*. Vol. I. New York: Wiley.

Luce, R. D., Bush, R. R., and Galanter, E., eds. 1963b. *Handbook of mathematical psychology*. Vol. II. New York: Wiley.

Lycan, W. 1991. *Even* and *even if*. *Linguistics and Philosophy* 14, 115–150.

MacDonald, M. C. 1993. The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language* 32, 692–715.

MacDonald, M. C. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes* 9, 157–201.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101, 676–703.

MacWhinney, B., and Bates, E., eds. 1989. *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press.

MacWhinney, B., Bates, E., and Kliegl, R. 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior* 23, 127–150.

MacWhinney, B., and Leinbach, J. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition* 40, 121–157.

Maling, J. 1989. Adverbials and structural case in Korean. In S. Kuno, I. Lee, J. Whitman, S.-Y. Bak, Y.-S. Kang, and Y.-J. Kim, eds., *Harvard studies in Korean linguistics*. Vol. 3. Cambridge, Mass.: Harvard University, Department of Linguistics, pp. 297–308.

Maling, J. 1993. Of nominative and accusative: The hierarchical assignment of grammatical case in Finnish. In A. Holmberg and U. Nikanne, eds., *Case and other functional categories in Finnish syntax*. Berlin: Mouton de Gruyter, pp. 51–76.

Malinowski, B. 1937. The dilemma of contemporary linguistics. In D. Hymes, ed., *Language in culture and society*. New York: Harper & Row, pp. 63–68.

Malouf, R. P. 2000. *Mixed categories in the hierarchical lexicon*. Stanford, Calif.: CSLI Publications.

Manaster Ramer, A., and Hitchcock, C. 1996. Glass houses: Greenberg, Ringe, and the mathematics of comparative linguistics. *Anthropological Linguistics* 38, 601–619.

Manning, C. D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 235–242.

Manning, C. D., and Schütze, H. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Marchman, V., and Bates, E. 1994. Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language* 21, 339–366.

Marcus, G. F. 2001. *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, Mass.: MIT Press.

Marcus, G. F., Brinkman, U., Clahsen, H., Wiese, R., and Pinker, S. 1995. German inflection: The exception that proves the rule. *Cognitive Psychology* 29, 189–256.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19, 313–330.

Marr, D. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Masciarotte, G. 1991. C'mon girl: Oprah Winfrey and the discourse of feminine talk. *Genders* 11, 81–110.

Matsuda, K. 1993. Dissecting analogical leveling quantitatively: The case of the innovative potential suffix in Tokyo Japanese. *Language Variation and Change* 5, 1–34.

Mattys, S., Jusczyk, P. W., Luce, P. A., and Morgan, J. L. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology* 38, 465–494.

Maye, J., and Gerken, L. 2000. Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, and T. Keith-Lucas, eds., *Proceedings of the 24th Annual Boston University Conference on Language Development*. Somerville, Mass.: Cascadilla Press, pp. 522–533.

Maye, J., Werker, J. F., and Gerken, L. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82 (3), B101–B111.

McCarthy, J. 1988. Feature geometry and dependency: A review. *Phonetica* 45, 88–108.

McCarthy, J. 1999. Sympathy and phonological opacity. *Phonology* 16, 331–399.

McCarthy, J., and Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, eds., *Machine intelligence*. Vol. 4. Edinburgh: Edinburgh University Press, pp. 463–502.

McCarthy, J., and Prince, A. 1995. Faithfulness and reduplicative identity. In J. Beckman, S. Urbanczyk, and L. Walsh, eds., *Papers in Optimality Theory*. University of Massachusetts Occasional Papers in Linguistics 18. Amherst: University of Massachusetts, GLSA, pp. 249–384.

McClelland, J. L. 1998. Connectionist models and Bayesian inference. In M. Oaksford and N. Chater, eds., *Rational models of cognition*. Oxford: Oxford University Press, pp. 21–53.

McClelland, J. L., Rumelhart, D. E., and the PDP Research Group. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2, *Psychological and biological models*. Cambridge, Mass.: MIT Press.

McCullagh, P., and Nelder, J. A. 1989. *Generalized linear models*. 2nd ed. London: Chapman and Hall.

McCulloch, W., and Pitts, W. 1943. A logical calculus of ideas immanent in neural activity. *Bulletin of Mathematical Biophysics* 5, 115–133.

McDonald, J. L. 1986. The development of sentence comprehension strategies in English and Dutch. *Journal of Experimental Child Psychology* 41, 317–335.

McDonald, J. L., and MacWhinney, B. 1989. Maximum likelihood models for sentence processing. In B. MacWhinney and E. Bates, eds., *The crosslinguistic study of sentence processing*. Cambridge: Cambridge University Press, pp. 397–421.

McDonald, S., Shillcock, R., and Brew, C. 2001. Low-level predictive inference in reading: Using distributional statistics to predict eye movements. Poster presented at AMLaP-2001, Saarbrücken, Germany, September 20–22, 2001. (Downloadable from ⟨http://www.cogsci.ed.ac.uk/~scottm/AMLaP-2001_poster.id.pdf⟩.)

McEnery, T., and Wilson, A. 2001. *Corpus linguistics*. 2nd ed. Edinburgh: Edinburgh University Press.

McLachlan, G. J., and Krishnan, T. 1996. *The EM algorithm and extensions*. New York: Wiley.

McLeod, P., Plunkett, K., and Rolls, E. 1998. *Introduction to connectionist modelling of cognitive processes*. Oxford: Oxford University Press.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language* 38, 283–312.

Mehta, C., and Patel, N. 1986. Algorithm 643: FEXACT: A Fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Transactions on Mathematical Software* 12, 154–161.

Mendoza-Denton, N. 1997. Chicana/Mexicana identity and linguistic variation: An ethnographic and sociolinguistic study of gang affiliation in an urban high school. Ph.D. thesis, Stanford University.

Mendoza-Denton, N. 1999. Style. In A. Duranti, ed., *Language matters: A lexicon for the millennium. Journal of Linguistic Anthropology* 9(1–2), 238–240.

Mendoza-Denton, N. 2001. Language and identity. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes, eds., *The handbook of language variation and change*. Oxford: Blackwell, pp. 475–495.

Merlo, P. 1994. A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research* 23, 435–457.

Merlo, P., and Leybold, M. 2001. Automatic distinction of arguments and modifiers: The case of prepositional phrases. *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001)*, pp. 121–128.

Mermelstein, P. 1975. Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America* 58, 880–883.

Miller-Ockhuizen, A., and Sands, B. E. 2000. Contrastive lateral clicks and variation in click types. *Proceedings of the Eighth International Conference on Spoken Language Processing (ICSLP-00)*. Beijing, Vol. 2, pp. 499–500.

Milne, R. W. 1982. Predicting garden path sentences. *Cognitive Science* 6, 349–374.

Milroy, L. 1980. *Language and social networks*. Baltimore, Md.: University Park Press.

Mitchell, D. C. 1994. Sentence parsing. In M. A. Gernsbacher, ed., *Handbook of psycholinguistics*. San Diego, Calif.: Academic Press, pp. 375–409.

Mitchell, D. C., and Brysbaert, M. 1998. Challenges to recent theories of cross-linguistic variation in parsing: Evidence from Dutch. In D. Hillert, ed., *Sentence processing: A crosslinguistic perspective*. Syntax and Semantics 31. San Diego, Calif.: Academic Press, pp. 313–344.

Mitchell, T. M., ed. 1997. *Machine learning*. New York: McGraw-Hill.

Moore, M., and MaCabe, G. 1989. *Introduction to the practice of statistics*. New York: Freeman.

Moore, R., Appelt, D., Dowding, J., Gawron, J. M., and Moran, D. 1995. Combining linguistic and statistical knowledge sources in natural-language processing for ATIS. *Proceedings of the January 1995 ARPA Spoken Language Systems Technology Workshop*, Austin, Tex., pp. 261–264.

Moreton, E. 1997. Phonotactic rules in speech perception. Abstract 2aSC4 from the 134th Meeting of the Acoustical Society of America, San Diego, Calif.

Morgan, J. N., and Messenger, R. C. 1973. THAID: A sequential analysis program for the analysis of nominal scale dependent variables. Technical report. Ann Arbor: University of Michigan, Institute for Social Research.

Morgan, J. N., and Sonquist, J. A. 1963. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* 58, 415–435.

Morreau, M. P. 1992. Conditionals in philosophy and artificial intelligence. Ph.D. thesis, University of Amsterdam.

Morrill, G. 1994. *Type-logical grammar*. Dordrecht: Kluwer.

Morton, J. 1969. Interaction of information in word recognition. *Psychological Review* 76, 165–178.

Müller, G. 1999. Optimality, markedness, and word order in German. *Linguistics* 37, 777–818.

Mumford, D. 1999. The dawning of the age of stochasticity. Based on a lecture at the Accademia Nazionale dei Lincei. (Downloadable from ⟨http://www.dam.brown.edu/people/mumford/Papers/Dawning.ps⟩.)

Munson, B. 2000. Phonological pattern frequency and speech production in children and adults. Ph.D. thesis, Ohio State University.

Munson, B. 2001. Phonological pattern frequency and speech production in children and adults. *Journal of Speech, Language and Hearing Research* 44, 778–792.

Nagy, N., and Reynolds, B. 1997. Optimality Theory and variable word-final deletion in Faetar. *Language Variation and Change* 9, 37–56.

Nagy, W. E., and Anderson, R. C. 1984. How many words are there in printed school English? *Reading Research Quarterly* 19, 304–330.

Napoli, D. J. 1981. Semantic interpretation vs. lexical governance. *Language* 57, 841–887.

Narayanan, S., and Jurafsky, D. 1998. Bayesian models of human sentence processing. *Proceedings of the 20th Annual Conference of the Cognitive Science Society (COGSCI-98)*, Madison, Wisc., pp. 752–757.

Narayanan, S., and Jurafsky, D. 2002. A Bayesian model predicts human parse preference and reading times in sentence processing. *Proceedings of Neural Information Processing Systems 2001 (NIPS 2001)*, Vancouver.

Nelder, J., and Mead, R. 1965. A simplex method for function minimization. *Computer Journal* 7, 308–313.

Nelson, G., Wallis, S., and Aarts, B. In press. *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins.

Nerbonne, J., and Heeringa, W. 2001. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica* 9, 69–83.

Neumann, G. 1998. Automatic extraction of stochastic lexicalized tree grammars from treebanks. *Proceedings of the Fourth Workshop on Tree-Adjoining Grammars and Related Frameworks*, Philadelphia, Pa., pp. 17–23.

Niedzielski, N. A. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Social Psychology* 18, 62–85.

Nittrouer, S. 1992. Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics* 20, 351–382.

Nittrouer, S. 1993. The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. *Journal of Speech and Hearing Research* 30, 319–329.

Niyogi, P., and Berwick, R. 1995. A dynamical system model for language change. Technical report 1515. Cambridge, Mass.: MIT, Artificial Intelligence Laboratory.

Noble, S. 1985. To have and have got. Paper presented at New Ways of Analyzing Variation (NWAV 14), Georgetown University.

Norris, D. G. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52, 189–234.

Norris, D. G., McQueen, J., Cutler, A., and Butterfield, S. 1997. The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology* 34, 191–243.

Oakes, M. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oja, E. 1982. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology* 15, 267–273.

Oldfield, R. C., and Wingfield, A. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17, 273–281.

Oliveira e Silva, G. 1982. Estudo da regularidade na variagco dos possessivos no portugujs do Rio de Janeiro. Ph.D. thesis, Universidade Federal do Rio de Janeiro.

Olofsson, A. 1990. A participle caught in the act: On the prepositional use of *following. Studia Neophilologica* 62, 23–35.

Osgood, C., and Sebeok, T. 1954. Psycholinguistics: A survey of theory and research problems. *Journal of Abnormal and Social Psychology* 49, 1–203.

Oswalt, R. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3, 117–129.

Pan, S., and Hirschberg, J. 2000. Modeling local context for pitch accent prediction. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, pp. 233–240.

Partee, B. H. 1988. Many quantifiers. *ESCOL '88: Proceedings of the Fifth Eastern States Conference on Linguistics*, Philadelphia, Pa., pp. 383–402.

Partee, B. H., ter Meulen, A., and Wall, R. E. 1990. *Mathematical methods in linguistics*. Dordrecht: Kluwer.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, Calif.: Morgan Kaufmann.

Pearlmutter, N., Daugherty, K., MacDonald, M., and Seidenberg, M. 1994. Modeling the use of frequency and contextual biases in sentence processing. *Proceedings of the 16th Annual Conference of the Cognitive Science Society (COGSCI-94)*, Atlanta, Ga., pp. 699–704.

Peck, J. 1994. Talk about racism: Framing a popular discourse of race on Oprah Winfrey. *Cultural Critique*, Spring, 89–126.

Pelletier, F. J., and Asher, N. 1997. Generics and defaults. In J. van Benthem and A. ter Meulen, eds., *Handbook of logic and language*. Amsterdam: Elsevier, pp. 1125–1177.

Pereira, F. 2000. Formal grammar and information theory: Together again. *Philosophical Transactions of the Royal Society* 358, 1239–1253. (Downloadable from ⟨http://citeseer.nj.nec.com/pereira00formal.html⟩.)

Pesetsky, D. 1998. Principles of sentence pronunciation. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, and D. Pesetsky, eds., *Is the best good enough?* Cambridge, Mass.: MIT Press, pp. 337–383.

Peterson, G. E., and Barney, H. L. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24, 175–184. (Data downloadable from ⟨http://www-2.cs.cmu.edu/afs/cs/project/airepository/ai/areas/speech/database/pb⟩.)

Phillips, B. 1998. Word frequency and lexical diffusion in English stress shifts. In R. Hog and L. van Bergen, eds., *Historical linguistics 1995*. Vol. 2, *Germanic linguistics*. Amsterdam: John Benjamins, pp. 223–232.

Phillips, B. 2001. Lexical diffusion, lexical frequency and lexical analysis. In J. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 123–136.

Pickering M. J., Traxler, M. J., and Crocker, M. W. 2000. Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language* 43, 447–475.

Pierrehumbert, J. 1992. Dissimilarity in the Arabic verbal roots. *Proceedings of the 23rd Meeting of the North East Linguistic Society (NELS 23)*, Ottawa, pp. 367–381.

Pierrehumbert, J. 1993. Prosody, intonation, and speech technology. In M. Bates and R. Weischedel, eds., *Challenges in natural language processing*. Cambridge: Cambridge University Press, pp. 257–282.

Pierrehumbert, J. 1994. Syllable structure and word structure. In P. Keating, ed., *Papers in laboratory phonology III*. Cambridge: Cambridge University Press, pp. 168–188.

Pierrehumbert, J. 2000. What people know about sounds of language. *Studies in the Linguistic Sciences* 29, 111–120. (Downloadable from ⟨http://www.ling.nwu.edu/~jbp⟩.)

Pierrehumbert, J. 2001a. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 137–157. (Downloadable from ⟨http://www.ling.nwu.edu/~jbp⟩.)

Pierrehumbert, J. 2001b. Why phonological constraints are so coarse grained. *SWAP (Spoken Word Access Processes)* 16, 691–698.

Pierrehumbert, J. In press. Word specific phonetics. In C. Gussenhoven and N. Warner, eds., *Laboratory phonology VII*. Berlin: Mouton de Gruyter. (Downloadable from ⟨http://www.ling.nwu.edu/~jbp⟩.)

Pierrehumbert, J., and Beckman, M. E. 1988. *Japanese tone structure*. Cambridge, Mass.: MIT Press.

Pierrehumbert, J., Beckman, M. E., and Ladd, D. R. 2001. Conceptual foundations of phonology as a laboratory science. In N. Burton-Roberts, P. Carr, and G. Docherty, eds., *Phonological knowledge*. Oxford: Oxford University Press, pp. 273–304. (Downloadable from ⟨http://www.ling.nwu.edu/~jbp⟩.)

Pierrehumbert, J., and Talkin, D. 1992. Lenition of /h/ and glottal stop. In G. Docherty and D. R. Ladd, eds., *Papers in laboratory phonology II*. Cambridge: Cambridge University Press, pp. 90–117.

Pinker, S. 1991. Rules of language. *Science* 153, 530–535.

Pinker, S. 1997. Words and rules in the human brain. *Nature* 387, 547–548.

Pinker, S. 1999. *Words and rules: The ingredients of language*. New York: Basic Books. Weidenfeld and Nicolson.

Pinker, S., and Prince, A. 1988. On language and connectionism. *Cognition* 28, 73–193.

Pinker, S., and Prince, A. 1994. Regular and irregular morphology and the psychological status of rules of grammar. In R. C. S. Lima and G. Iverson, eds., *The reality of linguistic rules*. Amsterdam: John Benjamins, pp. 321–351.

Pintzuk, S. 1995. Variation and change in Old English clause structure. *Language Variation and Change* 7, 229–260.

Plag, I., Dalton-Puffer, C., and Baayen, R. H. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3, 209–228.

Plaut, D. C., and Gonnerman, L. M. 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes* 15, 445–485.

Plunkett, K., and Juola, P. 1999. A connectionist model of English past tense and plural morphology. *Cognitive Science* 23, 463–490.

Poisson, S. D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile précédées des règles du calcul des probabilités*. Paris: Bachelier.

Pollard, C., and Sag, I. A. 1987. *Information-based syntax and semantics*. Vol. 1. Stanford, Calif.: CSLI Publications.

Pollard, C., and Sag, I. A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Pollock, J.-Y. 1989. Verb movement, Universal Grammar, and the structure of IP. *Linguistic Inquiry* 20, 365–424.

Poplack, S. 1979. Function and process in a variable phonology. Ph.D. thesis, University of Pennsylvania.

Port, R. F., Mitleb, F., and O'Dell, M. 1981. Neutralization of obstruent voicing in German is incomplete. *Journal of the Acoustical Society of America* 70, S10.

Powers, D. A., and Xie, Y. 1999. *Statistical methods for categorical data analysis*. San Diego, Calif.: Academic Press.

Preston, D. 1991. Sorting out the variables in sociolinguistic theory. *American Speech* 66, 33–56.

Prince, A., and Smolensky, P. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical report TR-2. New Brunswick, N.J.: Rutgers University, Center for Cognitive Science.

Pritchett, B. 1988. Garden path phenomena and the grammatical basis of language processing. *Language* 64, 539–576.

Przepiórkowski, A. 1999a. On case assignment and "adjuncts as complements." In G. Webelhuth, J.-P. Koenig, and A. Kathol, eds., *Lexical and constructional aspects of linguistic explanation*. Stanford, Calif.: CSLI Publications, pp. 231–245.

Przepiórkowski, A. 1999b. On complements and adjuncts in Polish. In R. D. Borsley and A. Przepiórkowski, eds., *Slavic in Head-Driven Phrase Structure Grammar*. Stanford, Calif.: CSLI Publications, pp. 183–210.

Pullum, G. K. 1996. Learnability, hyperlearning, and the poverty of the stimulus. *Proceedings of the 22nd Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on the Role of Learnability in Grammatical Theory*, pp. 498–513.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. 1985. *A comprehensive grammar of the English language*. London: Longman.

Raaijmakers, J. G. W., and Shiffrin, R. M. 1981. Search of associative memory. *Psychological Review* 88, 93–134.

Radford, A. 1988. *Transformational grammar*. Cambridge: Cambridge University Press.

Raimy, E., and Vogel, I. 2000. Compound and phrasal stress: A case of late acquisition. Paper presented at the Annual Meeting of the Linguistic Society of America, Chicago.

Ramscar, M. 2002. The role of meaning in inflection: Why the past tense doesn't require a rule. *Cognitive Psychology*.

Ramsey, F. L., and Schafer, D. W. 1997. *The statistical sleuth: A course in methods of data analysis*. Belmont, Calif.: Duxbury Press.

Rand, D., and Sankoff, D. 1990. Goldvarb version 2: A variable rule application for the Macintosh. Montréal: Université de Montréal, Centre de Recherches Mathématiques.

Rao, R. P. N., Olshausen, B. A., and Lewicki, M. S. 2002. *Probabilistic models of the brain: Perception and neural function*. Cambridge, Mass.: MIT Press.

Ratnaparkhi, A. 1998. Maximum entropy models for natural language ambiguity resolution. Ph.D. thesis, University of Pennsylvania.

Ratnaparkhi, A. 1999. Learning to parse natural language with maximum entropy models. *Machine Learning* 34, 151–175.

Rehder, B. 1999. A causal model theory of categorization. *Proceedings of the 21st Annual Conference of the Cognitive Science Society (COGSCI-99)*, Vancouver, British Columbia, pp. 595–600.

Reichenbach, H. 1949. *A theory of probability*. Berkeley and Los Angeles: University of California Press. Original German edition published 1935.

Resnik, P. 1992. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. *Proceedings of COLING 1992*, Nantes, France, pp. 418–424.

Resnik, P. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61, 127–159.

Richards, N. 1997. Leerdil yuujmen bana yanangarr (Old and new Lardil). In R. Pensalfini and N. Richards, eds., *Papers on Australian languages*. MIT Working Papers on Endangered and Less Familiar Languages 2. Cambridge, Mass.: MIT, Department of Linguistics and Philosophy, MIT WPL, pp. 147–163.

Rickford, J. R. 1986. The need for new approaches to social class analysis in sociolinguistics. *Language and Communication* 6, 215–221.

Rickford, J. R. 2001. Implicational scales. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes, eds., *The handbook of language variation and change*. Oxford: Blackwell, pp. 142–168.

Rickford, J. R., and McNair-Knox, F. 1994. Addressee and topic-influenced style shift: A quantitative and sociolinguistic study. In D. Biber and E. Finegan, eds., *Sociolinguistic perspectives on register*. Oxford: Oxford University Press, pp. 235–276.

Rickford, J. R., Wasow, T., Mendoza-Denton, N., and Espinoza, J. 1995. Syntactic variation and change in progress: Loss of the verbal coda in topic-restricting *as far as* constructions. *Language* 71, 102–131.

Riezler, S., Prescher, D., Kuhn, J., and Johnson, M. 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, pp. 480–487.

Ringe, D. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82.

Ripley, B. D. 1996. *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Rizzi, L. 1990. *Relativized minimality*. Cambridge, Mass.: MIT Press.

Roberts, I. 1993. Agreement parameters and the development of the English modal auxiliaries. *Natural Language and Linguistic Theory* 3, 21–58.

Roland, D. 2001. Verb sense and verb subcategorization probabilities. Ph.D. thesis, University of Colorado, Boulder.

Roland, D., and Jurafsky, D. 1998. How verb subcategorization frequencies are affected by corpus choice. *Proceedings of COLING 1998*, Montreal, pp. 1122–1128.

Roland, D., and Jurafsky, D. 2002. Verb sense and verb subcategorization probabilities. In P. Merlo and S. Stevenson, eds., *Sentence processing and the lexicon: Formal, computational, and experimental perspectives*. Amsterdam: John Benjamins, pp. 325–346.

Roland, D., Jurafsky, D., Menn, L., Gahl, S., Elder, E., and Riddoch, C. 2000. Verb subcategorization frequency differences between business-news and balanced corpora: The role of verb sense. *Proceedings of the Association for Computational Linguistics (ACL-2000) Workshop on Comparing Corpora*, Hong Kong, pp. 28–34.

Romaine, S. 1985. Variable rules, o.k.? or can there be sociolinguistic grammars? *Language Communication* 5, 53–67.

Rosenfeld, R. 1994. Adaptive statistical language modeling: A maximum entropy approach. Ph.D. thesis, Carnegie Mellon University. (Available as technical report CMU-CS-94-138.)

Ross, J. R. 1972. The category squish: Endstation Hauptwort. *Papers from the Eighth Regional Meeting, Chicago Linguistic Society*, pp. 316–328.

Ross, J. R. 1973a. Clausematiness. In E. L. Keenan, ed., *Formal semantics of natural language*. Cambridge: Cambridge University Press, pp. 422–475.

Ross, J. R. 1973b. A fake NP squish. In C.-J. N. Bailey and R. W. Shuy, eds., *New ways of analyzing variation in English*. Washington, D.C.: Georgetown University Press, pp. 96–140.

Ross, J. R. 1973c. Nouniness. In O. Fujimura, ed., *Three dimensions of linguistic theory*. Tokyo: Tokyo English Corporation, pp. 137–257.

Ross, S. 2000. *Introduction to probability models*. San Diego, Calif.: Academic Press.

Rousseau, P. 1978. Analyse des données binaires. Ph.D. thesis, Université de Montréal.

Rousseau, P. 1989. A versatile program for the analysis of sociolinguistic data. In R. Fasold and D. Schiffrin, eds., *Language change and variation*. Amsterdam: John Benjamins, pp. 395–409.

Rousseau, P., and Sankoff, D. 1978a. Advance in variable rule methodology. In D. Sankoff, ed., *Linguistic variation: Models and methods*. New York: Academic Press, pp. 57–69.

Rousseau, P., and Sankoff, D. 1978b. A solution to the problem of grouping speakers. In D. Sankoff, ed., *Linguistic variation: Models and methods*. New York: Academic Press, pp. 97–117.

Rubenstein, H., Garfield, L., and Millikan, J. A. 1970. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior* 9, 487–494.

Rubenstein, H., and Pollack, I. 1963. Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior* 2, 147–158.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2, *Psychological and biological models*. Cambridge, Mass.: MIT Press, pp. 318–362.

Rumelhart, D. E., and McClelland, J. L. 1986. On learning the past tenses of English verbs. In J. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2, *Psychological and biological models*. Cambridge, Mass.: MIT Press, pp. 216–271.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1, *Foundations*. Cambridge, Mass.: MIT Press.

Russo, R. 2001. *Empire Falls*. New York: Knopf.

Saffran, J. R. 2001. The use of predictive dependencies in language learning. *Journal of Memory and Language* 44, 493–515.

Saffran, J. R., Aslin, R. N., and Newport, E. L. 1996. Statistical cues in language acquisition: Word segmentation by infants. *Proceedings of the 18th Annual Conference of the Cognitive Science Society (COGSCI-96)*, San Diego, Calif., pp. 376–380.

Saffran, J. R., Newport, E. L., and Aslin, R. N. 1996a. Statistical learning by 8-month-old infants. *Science* 274, 1926–1928.

Saffran, J. R., Newport, E. L., and Aslin, R. N. 1996b. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35, 606–621.

Salmon, W. C. 1977. Objectively homogeneous reference classes. *Synthese* 36, 399–414.

Sampson, G. 1987. Evidence against the "grammatical"/"ungrammatical" distinction. In W. Meijs, ed., *Corpus linguistics and beyond*. Amsterdam: Rodopi, pp. 219–226.

Sampson, G. 2001. *Empirical linguistics*. London: Continuum International.

Sankoff, D. 1978a. Probability and linguistic variation. *Synthese* 37, 217–238.

Sankoff, D. 1978b. Statistical dependence among successive occurrences of a variable in discourse. In D. Sankoff, ed., *Linguistic variation: Models and methods*. New York: Academic Press, pp. 227–238.

Sankoff, D. 1985. Statistics in linguistics. In *Encyclopedia of the social sciences 5*. New York: Wiley, pp. 74–81.

Sankoff, D. 1988. Variable rules. In U. Ammon, N. Dittmar, and K. J. Mattheier, eds., *Sociolinguistics: An international handbook of the science of language and society*. Vol. 2. Berlin: Walter de Gruyter, pp. 984–997.

Sankoff, D., and Labov, W. 1979. On the uses of variable rules. *Language in Society* 8, 189–222.

Sankoff, D., and Rousseau, P. 1974. A method for assessing variable rule and implicational scale analyses in language variation. In J. Mitchell, ed., *Computers in humanities*. Edinburgh: Edinburgh University Press, pp. 3–15.

Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank project. 3rd rev., 2nd printing, Feb. 1995. Philadelphia: University of Pennsylvania.

Santorini, B. 1993. The rate of phrase structure change in the history of Yiddish. *Language Variation and Change* 5, 257–283.

Sapir, E. 1921. *Language: An introduction to the study of speech*. New York: Harcourt Brace.

Savin, H. B. 1963. Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America* 35, 200–206.

Scarborough, D. L., Cortese, C., and Scarborough, H. S. 1977. Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance* 3, 1–17.

Scarborough, R. B. 2002. Effects of lexical confusability on the production of coarticulation. Poster presented at the Eighth Conference on Laboratory Phonology, New Haven, Conn., June 27–29.

Schabes, Y. 1992. Stochastic lexicalized tree-adjoining grammars. *Proceedings of COLING 1992*, Nantes, France, pp. 426–433.

Schilling-Estes, N. 1996. The linguistic and sociolinguistic status of /ay/ in Outer Banks English. Ph.D. thesis, University of North Carolina at Chapel Hill.

Schilling-Estes, N. 2001. Investigating stylistic variation. In J. K. Chambers, P. Trudgill, and N. Schilling-Estes, eds., *The handbook of language variation and change*. Oxford: Blackwell, pp. 375–402.

Schilling-Estes, N., and Wolfram, W. 1994. Convergent explanation and alternative regularization patterns: *Were/weren't* leveling in a vernacular English variety. *Language Variation and Change* 6, 273–302.

Schreuder, R., and Baayen, R. H. 1997. How complex simplex words can be. *Journal of Memory and Language* 37, 118–139.

Schubert, L. K., and Pelletier, F. J. 1989. Generically speaking, or using discourse representation theory to interpret generics. In G. Chierchia, B. H. Partee, and R. Turner, eds., *Properties, types and meaning*. Dordrecht: Kluwer, pp. 193–268.

Schuchardt, H. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Robert Oppenheim. Excerpted with English translation in T. Vennemann and T. H. Wilbur, eds., *Schuchardt, the Neogrammarians, and the transformational theory of phonological change*. Frankfurt: Athenaum Verlag (1972).

Schulman, R. 1983. Vowel categorization by the bilingual listener. *PERILUS* 3, 81–100.

Schultink, H. 1961. Produktiviteit als morfologisch fenomeen. *Forum der Letteren* 2, 110–125.

Schütze, C. T. 1995. PP attachment and argumenthood. In C. Schütze, J. Ganger, and K. Broihier, eds., *Papers on language processing and acquisition*. MIT Working Papers in Linguistics 26. Cambridge, Mass.: MIT, Department of Linguistics and Philosophy, MITWPL, pp. 95–151.

Schütze, C. T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Seidenberg, M. S., and Gonnerman, L. M. 2000. Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences* 4, 353–361.

Seidenberg, M. S., and Hoeffner, J. H. 1998. Evaluating behavioral and neuroimaging data on past tense processing. *Language* 74, 104–122.

Seidenberg, M. S., and MacDonald, M. C. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23, 569–588.

Selkirk, E. O. 1980. *On prosodic structure and its relation to syntactic structure*. Bloomington: Indiana University Linguistics Club.

Selkirk, E. O. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, Mass.: MIT Press.

Sereno, J., and Jongman, A. 1997. Processing of English inflectional morphology. *Memory and Cognition* 25, 425–437.

Shanklin, T. 1990. The grammar of negation in Middle English. Ph.D. thesis, University of Southern California.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423, 623–656.

Shao, J. 1999. *Mathematical statistics*. New York: Springer-Verlag.

Shattuc, J. 1997. *The talking cure: TV talk shows and women*. London: Routledge.

Shi, Z. 1989. The grammaticalization of the particle *le* in Mandarin Chinese. *Language Variation and Change* 1, 99–114.

Sigley, R. 2001. The importance of interaction effects. Paper presented at the 14th New Zealand Linguistic Society Conference.

Silva-Corvalán, C. 1989. *Sociolinguística: Teoría y análysis*. Madrid: Alhambra.

Silverstein, M. 1976. Hierarchy of features and ergativity. In R. M. W. Dixon, ed., *Grammatical categories in Australian languages*. Canberra: Australian Institute of Aboriginal Studies, pp. 112–171.

Simpson, G. B., and Burgess, C. 1985. Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance* 11, 28–39.

Sinclair, J. M. 1997. Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles, eds., *Teaching and language corpora*. London: Longman, 27–39.

Skousen, R. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.

Skousen, R. 1992. *Analogy and structure*. Dordrecht: Kluwer.

Skousen, R. 2000. Analogical modeling and quantum computing. Los Alamos, N.M.: Los Alamos National Laboratory. (Downloadable from ⟨http://arXiv.org⟩.)

Smolensky, P. 1986. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1, *Foundations*. Cambridge, Mass.: MIT Press, pp. 194–281.

Smolensky, P. 1993. Harmony, markedness, and phonological activity. Handout of keynote address, Rutgers Optimality Workshop-1. (Downloadable from the Rutgers Optimality Archive, ⟨http://roa.rutgers.edu/index.php.3⟩.)

Smolensky, P. 1997. Constraint interaction in generative grammar II: Local conjunction, or random rules in Universal Grammar. Presented at the Hopkins Optimality Theory Workshop/University of Maryland Mayfest. (Downloadable from ⟨http://hebb.cog.jhu.edu/pdf/constraint_interaction_generative_grammar.pdf⟩.)

Smolensky, P. 1999. Principles of Dave's philosophy. Contribution to the David Rumelhart Celebration at Carnegie Mellon University. (Downloadable from ⟨http://hebb.cog.jhu.edu/pdf/what_i_learned_from_dave_rumelhart.pdf⟩.)

Smolensky, P., and Legendre, G. 2000. Architecture of the mind/brain: Neural computation, optimality, and Universal Grammar in cognitive science. Manuscript, Johns Hopkins University.

Sokolov, J. L., and Snow, C. E. 1994. The changing role of negative evidence in theories of language development. In C. Gallaway and B. J. Richards, eds., *Input and interaction in language acquisition*. New York: Cambridge University Press, pp. 38–55.

Sonnenstuhl-Henning, I., and Huth, A. 2002. Processing and representation of German *-n* plurals: A dual mechanism approach. *Brain and Language* 81, 276–290.

Sorace, A. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76, 859–890.

Spivey, M. J., and Tanenhaus, M. K. 1998. Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24, 1521–1543.

Spivey-Knowlton, M. J. 1996. Integration of visual and linguistic information: Human data and model simulations. Ph.D. thesis, University of Rochester.

Spivey-Knowlton, M. J., and Sedivy, J. 1995. Resolving attachment ambiguities with multiple constraints. *Cognition* 55, 227–267.

Spivey-Knowlton, M. J., Trueswell, J., and Tanenhaus, M. K. 1993. Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology* 47, 276–309.

Stabler, E. P. 2001. *Computational minimalism: Acquiring and parsing languages with movement*. Oxford: Blackwell.

Stallings, L. M., MacDonald, M. C., and O'Seaghdha, P. G. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language* 39, 392–417.

Stalnaker, R. 1970. Probability and conditionals. *Philosophy of Science* 37, 64–80.

Stalnaker, R., and Jeffrey, R. 1994. Conditionals as random variables. In E. Eells and B. Skyrms, eds., *Probability and conditionals: Belief revision and rational decision*. Cambridge: Cambridge University Press, pp. 75–111.

Steels, L. 2000. Language as a complex adaptive system. In M. Schoenauer, ed., *Proceedings of PPSN VI*. Berlin: Springer-Verlag, pp. 17–26.

Sterelny, K. 1983. Linguistic theory and variable rules. *Language and Communication* 3, 47–69.

Steriade, D. 1993. Closure, release, and nasal contours. In M. Huffman and R. Krakow, eds., *Nasals, nasalization, and the velum*. San Diego, Calif.: Academic Press, pp. 401–470.

Steriade, D. 2000. Paradigm uniformity and the phonetics-phonology interface. In M. Broe and J. Pierrehumbert, eds., *Papers in laboratory phonology V: Acquisition and the lexicon*. Cambridge: Cambridge University Press, pp. 313–335.

Stevens, K. N. 1998. *Acoustic phonetics*. Cambridge, Mass.: MIT Press.

Stevenson, S. 1994. Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research* 23, 295–322.

Stevenson, S., and Merlo, P. 1997. Lexical structure and parsing complexity. *Language and Cognitive Processes* 12, 349–399.

Stolcke, A. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics* 21, 165–202.

Strand, E. A., and Johnson, K. 1996. Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon, ed., *Natural language processing and speech technology: Results of the 3rd KONVENS Conference in Bielefeld*. Berlin: Mouton, pp. 14–26.

Sudbury, A., and Hay, J. In press. The fall and rise of /r/: Rhoticity and /r/-sandhi in early New Zealand English. In *Selected papers from NWAV 30*.

Suppes, P. 1970. Probabilistic grammars for natural languages. *Synthese* 22, 95–116.

Suzuki, K., Maye, J., and Ohno, K. 2000. On the productivity of the lexical stratification of Japanese. Paper presented at the Annual Meeting of the Linguistic Society of America, Los Angeles, Calif.

Svartvik, J. 1966. *On voice in the English verb*. The Hague: Mouton.

Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96, 452–463.

Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21, 121–137.

Tabor, W. 1994. Syntactic innovation: A connectionist model. Ph.D. thesis, Stanford University.

Tabor, W. 2000. Lexical categories as basins of curvature. Manuscript, University of Connecticut.

Tabor, W., Juliano, C., and Tanenhaus, M. K. 1997. Parsing in a dynamical system. *Language and Cognitive Processes* 12, 211–272.

Tabossi, P., Spivey-Knowlton, M. J., McRae, K., and Tanenhaus, M. K. 1994. Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. In C. Umilta and M. Moscovitch, eds., *Attention and performance XV*. Hillsdale, N.J.: Erlbaum, pp. 589–615.

Taft, M. 1979. Recognition of affixed words and the word frequency effect. *Memory and Cognition* 7, 263–272.

Tagliamonte, S. 1999. *Was-were* variation across the generations: View from the city of New York. *Language Variation and Change* 10, 153–191.

Tanenhaus, M. K., Spivey-Knowlton, M. J., and Hanna, J. E. 2000. Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In M. W. Crocker, M. Pickering, and C. Clifton, eds., *Architectures and mechanisms for language processing*. Cambridge: Cambridge University Press, pp. 90–118.

Tanenhaus, M. K., Stowe, L. A., and Carlson, G. 1985. The interaction of lexical expectation and pragmatics in parsing filler-gap constructions. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society (COGSCI-85)*, Irvine, Calif., pp. 361–365.

Taylor, L., Grover, C., and Briscoe, E. 1989. The syntactic regularity of English noun phrases. *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL 4)*, Manchester, U.K., pp. 256–263.

Tenenbaum, J. B. 1999. Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, and D. A. Cohn, eds., *Advances in neural information processing systems*. Vol. 11. Cambridge, Mass.: MIT Press, pp. 59–65.

Tenenbaum, J. B., and Griffiths, T. L. 2001a. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences* 24, 579–616.

Tenenbaum, J. B., and Griffiths, T. L. 2001b. The rational basis of representativeness. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society (COGSCI-01)*, Edinburgh.

Tenenbaum, J. B., and Xu, F. 2000. Word learning as Bayesian inference. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society (COGSCI-00)*, Philadelphia, Pa., pp. 517–522.

Tesnière, L. 1959. *Eléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.

Teuscher, C. 2001. *Turing's connectionism: An investigation of neural network architectures*. London: Springer-Verlag.

Thibault, P., and Daveluy, M. 1989. Quelques traces du passage du temps dans le parler des Montréalais, 1971–1984. *Language Variation and Change* 1, 19–45.

Thibault, P., and Sankoff, G. 1993. Diverses facettes de l'insécurité linguistique: Vers une analyse comparative des attitudes et du français parlé par des franco- et des anglomontréalais. *Cahiers de l'Institut de Linguistique de Louvain* 19, 209–218.

Thomas, E. 1995. Phonetic factors and perceptual reanalysis in sound change. Ph.D. thesis, University of Texas at Austin.

Thomason, R. 1970. Indeterminist time and truth-value gaps. *Theoria* 36, 264–281.

Thomason, R. 1988. Theories of nonmonotonicity and natural language generics. In M. Krifka, ed., *Genericity in natural language: Proceedings of the 1988 Tübingen Conference*. SNS-Bericht 88-42. Tübingen, Germany: Tübingen University, Seminar für natürlich-sprachliche Systeme, pp. 395–406.

Tiersma, P. M. 1982. Local and general markedness. *Language* 58, 832–849.

Toutanova, K., and Manning, C. D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIG DAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC 2000)*, pp. 63–70.

Trask, R. L. 1979. On the origins of ergativity. In F. Plank, ed., *Ergativity: Towards a theory of grammatical relations*. London: Academic Press, pp. 385–404.

Traugott, E. C., and Heine, B. 1991. *Approaches to grammaticalization*. Amsterdam: John Benjamins.

Treiman, R., Kessler, B., Kneewasser, S., Tincoff, R., and Bowman, M. 2000. English speakers' sensitivity to phonotactic patterns. In M. Broe and J. Pierrehumbert, eds., *Papers in laboratory phonology V: Acquisition and the lexicon*. Cambridge: Cambridge University Press, pp. 269–283.

Trudgill, P. 1974. *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Trudgill, P. 1983. Acts of conflicting identity: The sociolinguistics of British pop-song pronunciation. In P. Trudgill, ed., *On dialect: Social and geographical perspectives*. Oxford: Blackwell, and New York: NYU Press, pp. 141–160.

Trudgill, P. 1988. Norwich revisited: Recent linguistic changes in an English urban dialect. *English World-Wide* 9, 33–49.

Trueswell, J. C. 1996. The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language* 35, 566–585.

Trueswell, J. C., and Tanenhaus, M. K. 1994. Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, Jr., L. Frazier, and K. Rayner, eds., *Perspectives on sentence processing*. Hillsdale, N.J.: Erlbaum, pp. 155–179.

Trueswell, J. C., Tanenhaus, M. K., and Garnsey, S. M. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33, 285–318.

Trueswell, J. C., Tanenhaus, M. K., and Kello, C. 1993. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition* 19, 528–553.

Tyler, L. K. 1984. The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics* 36, 417–427.

van den Bosch, A., Daelemans, W., and Weijters, T. 1996. Morphological analysis as classification: An inductive learning approach. *Proceedings of New Methods in Language Processing (NEMLAP 1996)*, Ankara, pp. 59–72.

van der Does, J. M., and van Lambalgen, M. 2000. A logic of vision. *Linguistics and Philosophy* 23, 1–92.

van Fraassen, B. C. 1969. Presuppositions, supervaluations, and free logic. In K. Lambert, ed., *The logical way of doing things*. New Haven, Conn.: Yale University Press, pp. 67–91.

van Fraassen, B. C. 1976. Probabilities of conditionals. In W. L. Harper and C. A. Hooker, eds., *Foundations of probability theory, statistical inference, and statistical theories of science*. Vol. 1. Dordrecht: Reidel, pp. 261–301.

van Fraassen, B. C. 1980. *The scientific image*. Oxford: Oxford University Press.

van Hout, R. 1984. The need for theory of choice in sociolinguistics. *Linguistische Berichte* 90, 39–57.

Vater, H. 1978. On the possibility of distinguishing between complements and adjuncts. In W. Abraham, ed., *Valence, semantic case and grammatical relations*. Amsterdam: John Benjamins, pp. 21–45.

Venables, W. N., and Ripley, B. D. 1994. *Modern applied statistics with S-Plus*. New York: Springer-Verlag.

Verspoor, C. M. 1997. Contextually-dependent lexical semantics. Ph.D. thesis, University of Edinburgh.

Vihman, M. M. 1996. *Phonological development: The origins of language in the child*. Cambridge, Mass.: Blackwell.

Vincent, D. 1991. Quelques études sociolinguistiques de particules du discours. *Revue Québecoise de Linguistique Théorique et Appliquée* 10, 41–60.

Vitevich, M., and Luce, P. 1998. When words compete: Levels of processing in perception of spoken words. *Psychological Science* 9, 325–329.

Vitevich, M., Luce, P., Pisoni, D., and Auer, E. T. 1999. Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language* 68, 306–311.

von Mises, R. 1957. *Probability, statistics and truth*. New York: Macmillan. Original German edition published 1928.

Wallis, S. A., Aarts, B., and Nelson, G. 1999. Parsing in reverse: Exploring ICE-GB with fuzzy tree fragments and ICECUP. In J. M. Kirk, ed., *Corpora galore: Papers from the 19th International Conference on English Language Research on Computerised Corpora, ICAME-98*. Amsterdam: Rodopi, pp. 335–344.

Wasow, T. 1997. Remarks on grammatical weight. *Language Variation and Change* 9, 81–105.

Way, A. 1999. A hybrid architecture for robust MT using LFG-DOP. *Journal of Experimental and Theoretical Artificial Intelligence* 11, 201–233.

Wechsler, S., and Lee, Y.-S. 1996. The domain of direct case assignment. *Natural Language and Linguistic Theory* 14, 629–664.

Weiner, E. J., and Labov, W. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19, 29–58.

Weinreich, U., Labov, W., and Herzog, M. 1968. Empirical foundations for a theory of language change. In W. Lehmann and Y. Malkiel, eds., *Directions for historical linguistics*. Austin: University of Texas Press, pp. 95–188.

Werker, J., and Tees, R. C. 1994. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7, 49–63.

Westerståhl, D. 1985. Logical constants in quantifier languages. *Linguistics and Philosophy* 8, 387–413.

Wetherell, C. 1980. Probabilistic languages: A review and some open questions. *Computing Surveys* 12, 361–379.

Whalen, D. H., and Sheffert, S. 1997. Normalization of vowels by breath sounds. In K. Johnson and J. Mullenix, eds., *Talker variability in speech processing*. San Diego, Calif.: Academic Press, pp. 133–144.

Whaley, C. P. 1978. Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior* 17, 143–154.

Wingfield, A. 1968. Effects of frequency on identification and naming of objects. *American Journal of Psychology* 81, 226–234.

Withgott, M. 1983. Segmental evidence for phonological constituents. Ph.D. thesis, University of Texas at Austin.

Wolfram, W. 1974. *Sociolinguistic aspects of assimilation: Puerto Rican English in New York City*. Arlington, Va.: Center for Applied Linguistics.

Wolfram, W. 1991. The linguistic variable: Fact and fantasy. *American Speech* 66, 22–31.

Wolfram, W., Hazen, K., and Schilling-Estes, N. 1999. Dialect change and maintenance on the Outer Banks. Technical report 81. Publication of the American Dialect Society. Tuscaloosa: University of Alabama Press.

Wolfson, N. 1976. Speech events and natural speech: Some implications for sociolinguistic methodology. *Language in Society* 5, 189–209.

Wright, R. 1997. Lexical competition and reduction in speech: A preliminary report. Research on Spoken Language Processing: Progress report 21. Bloomington: Indiana University, Speech Research Laboratory.

Yaeger-Dror, M. 1996. Phonetic evidence for the evolution of lexical classes: The case of a Montreal French vowel shift. In G. Guy, C. Feagin, J. Baugh, and D. Schiffrin, eds., *Towards a social science of language*. Philadelphia: John Benjamins, pp. 263–287.

Yaeger-Dror, M., and Kemp, W. 1992. Lexical classes in Montreal French. *Language and Speech* 35, 251–293.

Yang, C. 2000. Internal and external forces in language change. *Language Variation and Change* 12, 231–250.

Young, R., and Bayley, R. 1996. VARBRUL analysis for second language acquisition research. In R. Bayley and D. Preston, eds., *Second language acquisition and linguistic variation*. Amsterdam: John Benjamins, pp. 253–306.

Young, R., and Yandell, B. 1999. Top-down versus bottom-up analyses of interlanguage data: A reply to Saito. *Studies in Second Language Acquisition* 21, 477–488.

Zaenen, A. 1993. Unaccusativity in Dutch: Integrating syntax and lexical semantics. In J. Pustejovsky, ed., *Semantics and the lexicon*. London: Kluwer, pp. 129–161.

Zamuner, T., Gerken, L., and Hammond, M. 2001. Phonotactic probability in the acquisition of English. Poster presented at Journées d'Études sur l'Acquisition Phonologique Précoce, Carry-le-Rouet (Marseille). (Abstract downloadable from ⟨http://www.ehess.fr/centres/lscp/persons/peperkamp/Zamuner.htm⟩.)

Zuraw, K. 2000. Patterned exceptions in phonology. Ph.D. thesis, University of California, Los Angeles. (Downloadable from ⟨http://www-rcf.usc.edu/~zuraw⟩.)

Name Index

Subject Index