

UNIVERSITY OF CALIFORNIA

Los Angeles

The Evolving Lexicon

A dissertation submitted in partial satisfaction of the  
requirements of the degree Doctor of Philosophy  
in Linguistics

by

Andrew Thomas Martin

2007

The dissertation of Andrew Thomas Martin is approved.

---

Jody Kreiman

---

Bruce Hayes

---

Colin Wilson

---

Kie Zuraw, Committee Chair

University of California, Los Angeles

2007

*To Hisako*

## TABLE OF CONTENTS

1.	LEXICAL COMPETITION AND EVOLUTION.....	1
1.1.	THE PROPOSAL .....	2
1.2.	NATURAL SELECTION IN LANGUAGE .....	9
1.3.	LEXICAL COMPETITION .....	11
1.4.	DERIVING CONVERGENCE .....	15
1.5.	THE ROLE OF FEEDBACK IN LEXICAL SELECTION .....	28
1.6.	THE EVOLUTION OF FREQUENCY DISTRIBUTIONS .....	33
1.7.	SUMMARY OF THE MODEL.....	43
1.8.	EMPIRICAL CONSEQUENCES OF THE MODEL .....	44
2.	MARKEDNESS AND LEXICAL CHANGE.....	47
2.1.	INTRODUCTION .....	47
2.2.	WORD-INITIAL CLUSTERS IN ENGLISH (BERG 1998) .....	48
2.3.	VOICED STOPS AND PLACE OF ARTICULATION .....	50
2.4.	THE EVOLUTION OF THE FRENCH LEXICON.....	54
2.4.1.	<i>From Proto-Indo-European to Latin</i> .....	55
2.4.2.	<i>From Latin to French</i> .....	57
2.4.3.	<i>Sources of French vocabulary</i> .....	61
2.5.	MODIFYING THE MODEL.....	64
2.6.	CONCLUSION.....	69
3.	PHONEME COOCCURRENCE AND LEXICAL BIAS.....	70
3.1.	GRADIENT OCP EFFECTS IN ENGLISH .....	70
3.2.	LIQUID COOCCURRENCE IN ENGLISH NEOLOGISMS .....	71
3.2.1.	<i>The Monte Carlo test for significance</i> .....	73
3.2.2.	<i>The OCP in several stages of English</i> .....	76
3.3.	LIQUID COOCCURRENCE IN AMERICAN BABY NAMES.....	79
3.4.	PROCESSING AND THE OCP.....	83
3.4.1.	<i>Refractory period</i> .....	84
3.4.2.	<i>Confusability</i> .....	86
3.4.3.	<i>Repetition blindness</i> .....	87
3.4.4.	<i>Why is repetition difficult?</i> .....	88
3.5.	EXTENDING THE MODEL.....	90
4.	MORPHOLOGICALLY DRIVEN PHONOTACTIC PREFERENCES.....	95
4.1.	THE PROPOSAL.....	95
4.2.	ENGLISH CONSONANT CLUSTERS.....	97
4.2.1.	<i>Geminates in English</i> .....	100
4.2.2.	<i>Compounds with geminates</i> .....	101
4.2.3.	<i>Suffixed words with geminates</i> .....	103
4.2.4.	<i>Summary of English data</i> .....	106

4.3.	NAVAJO SIBILANT HARMONY .....	106
4.3.1.	<i>Navajo compounds</i> .....	108
4.4.	TURKISH VOWEL HARMONY .....	109
4.4.1.	<i>Turkish compounds</i> .....	110
4.5.	DISCUSSION .....	113
4.6.	THE PHONOTACTIC LEARNER .....	117
4.7.	MAXIMUM ENTROPY.....	123
4.8.	TESTING THE MAXENT LEARNER ON THE P-T LANGUAGE.....	127
4.9.	THE NATURE OF THE LEARNING BIAS .....	133
4.10.	CONSEQUENCES FOR THE MODEL .....	136
5.	CONCLUSION.....	137
5.1.	SUMMARY OF FINDINGS .....	137
5.2.	SUMMARY OF THE MODEL.....	138
5.3.	DIRECTIONS FOR FUTURE RESEARCH.....	140
5.3.1.	<i>Data collection</i> .....	140
5.3.2.	<i>Correlations between historical change and processing ease</i> .....	141
5.3.3.	<i>Experimental tests of phonotactic preferences</i> .....	142
5.3.4.	<i>The interaction of phonotactics and morphology</i> .....	143
5.3.5.	<i>Avoidance in children</i> .....	143
5.3.6.	<i>The interaction of phonotactics and sociolinguistic variables</i> .....	144
	REFERENCES .....	145

## TABLE OF EXHIBITS

(1) Consonant type frequencies in English .....	4
(2) Consonant type frequencies in Old English and Modern English .....	7
(3) <i>Shivaree</i> and its competitors .....	13
(4) Feedforward speech production network .....	15
(5) Activating synonyms .....	17
(6) Within-speaker convergence (bars represent resting activations for synonyms A, B, C).....	19
(7) Convergence simulation network.....	21
(8) Calculating activation (modified from Dell 1986).....	21
(9) Speaker weight adjustment.....	22
(10) Listener weight adjustment .....	22
(11) Set of selfish agents ( $\alpha > \beta$ ): each agent settles on a different word .....	24
(12) Set of cooperative agents ( $\beta > \alpha$ ): agents come to agree on a single word .....	25
(13) Set of cooperative agents ( $\beta > \alpha$ ) who talk only to neighbors: dialects emerge..	27
(14) Network with feedback .....	29
(15) Network initial state .....	34
(16) Network final state (after 1,000 generations).....	36
(17) Phoneme type frequencies over time .....	37
(18) Final phoneme type frequencies.....	40
(19) Decreasing entropy in distribution of name-initial consonants.....	41
(20) Word retention by cluster type .....	49
(21) Lexical counts of word-initial stops in several languages' .....	52
(22) /b/- and /d/-initial words in Latin and French' .....	57
(23) Possible fates of Latin words .....	58
(24) /b/- and /d/-initial survival rates in modern Romance languages.....	59
(25) Main sources of French vocabulary .....	62
(26) Breakdown of French words by origin (no prefixed forms) .....	62
(27) Network with gestural nodes.....	65
(28) Introduction of new phoneme: no markedness .....	67
(29) Calculating activation with weights .....	67
(30) Introduction of new phoneme: with markedness .....	68
(31) Preparing a word list for Monte Carlo test.....	73
(32) Performing the Monte Carlo test.....	74
(33) Results of Monte Carlo test on CELEX two-liquid words.....	75
(34) Comparing attested CELEX liquid pairs to Monte Carlo results.....	76
(35) Sequences of identical liquids are underrepresented in Old English .....	76
(36) Sequences of identical liquids are underrepresented in Middle English.....	77
(37) OED neologisms by decade: liquid identity rates .....	78
(38) Liquid pairs in popular names by decade.....	81
(39) Identical liquids are underrepresented in drug brand names .....	82
(40) Examples of two-liquid drug brand names ( $N = 88$ ).....	82

(41) Identical liquids are underrepresented in names for fantasy role-playing game characters .....	83
(42) Examples of two-liquid role-playing game character names ( $N = 82$ ) .....	83
(43) Identical liquids are underrepresented in unusual baby names .....	83
(44) Examples of two-liquid unusual baby names ( $N = 20$ ) .....	83
(45) Sequencing synonyms in parallel: <i>on the (couch/sofa)</i> .....	93
(46) Legal word-medial CC clusters in English monomorphemes .....	98
(47) Illegal non-geminate clusters are underrepresented in compounds .....	99
(48) Legal clusters are overrepresented in compounds .....	99
(49) Geminates are underrepresented in English compounds .....	101
(50) Geminates are underrepresented in compounds spelled closed .....	102
(51) Geminates are underrepresented in compounds in Sepp corpus .....	103
(52) Illegal consonant clusters are underrepresented in <i>-ness</i> suffixed words .....	105
(53) Geminates are underrepresented in <i>-ly</i> suffixed words .....	105
(54) Geminates are underrepresented in <i>-less</i> suffixed words .....	105
(55) Navajo sibilant classes .....	106
(56) Examples of sibilant harmony (Fountain 1998) .....	107
(57) Exceptions to sibilant harmony in compounds (Young and Morgan 1987) .....	107
(58) Examples of compounds with two sibilants in adjacent syllables (one per root) .....	108
(59) Disharmonic compounds are underrepresented in Navajo .....	109
(60) Turkish vowel system .....	110
(61) <i>Izafet</i> compounds .....	111
(62) Single-word compounds .....	111
(63) Disharmonic stems are underrepresented in Turkish single-word compounds .....	112
(64) Disharmonic stems are overrepresented in Turkish <i>izafet</i> compounds .....	113
(65) Network with consonant cluster nodes .....	115
(66) Logically possible words in the p-t language .....	119
(67) Attested words in the p-t language .....	119
(68) Biased lexicon for p-t language .....	120
(69) Probabilities assigned to biased lexicon .....	120
(70) Probabilities assigned to unbiased lexicon .....	121
(71) Structure-sensitive vs. structure-blind frequency counts: unbiased lexicon .....	122
(72) Definition of score .....	124
(73) Determining candidate probability .....	124
(74) Probability of training data .....	125
(75) Maxent learning function .....	126
(76) Training data (unbiased) .....	128
(77) Constraints .....	128
(78) Final grammar, all constraints .....	129
(79) Example outputs as evaluated by grammar, all constraints .....	130
(80) Final grammar, structure-sensitive constraints only .....	130
(81) Example outputs as evaluated by grammar, structure-sensitive constraints only .....	131

(82) Effect of $\sigma^2$ on geminate ratio in compounds .....	131
(83) The learning bias with different lexicons .....	135



## ACKNOWLEDGEMENTS

First and foremost, I would like to thank my adviser, Kie Zuraw. Kie has been everything an adviser should be: generous, patient, critical at the right times, and always supportive. She suggested the paper topic that eventually blossomed into this dissertation, and she has since shaped my linguistic career in countless other ways. I feel truly privileged to have had Kie as my adviser, and will strive in my own teaching and research to live up to the high standard she has set.

I am also grateful to have had the opportunity to work with my other committee members. I have benefited greatly from Bruce Hayes' seemingly effortless command of data and theory as well as his always sage advice on many matters, linguistic and otherwise. Colin Wilson has shared with me his uncanny ability to not only find the flaws in any analysis, but invariably offer a better one, and my work is much stronger for it. I also thank Jody Kreiman for her insightful comments and support. I cannot imagine a better environment to do linguistics than in the company of these scholars, and I will miss them all dearly.

The UCLA Linguistics Department has been a fantastic place to be a graduate student, in no small part due to the other students who have shared this journey with me. I do not have the space to list everyone who made my life better in some way while I was here, but I would especially like to thank Christina Esposito, Jeff Heinz, and Sarah VanWagenen for their friendship.

I would also like to thank those who helped set me on the path that has led me here: my first linguistics teacher, John Haviland, for planting the seed that would eventually turn into a career, and Bernard Tranel, for single-handedly turning me into a phonologist. Of course, I would not be here at all were it not for my parents, Dennis and Gloria Martin, who not only gave me my first copy of Fromkin and Rodman, but whose unflagging love and encouragement have been a constant source of inspiration in my life. No matter how many languages I study, I will never have the words to express how much their support has meant to me.

Finally, I thank my wife, Hisako, and my daughter Yuki. Hisako has made many sacrifices to allow me to do what I most wanted to do, and Yuki's infectious smile has always managed to illuminate the darkest moments of writing this dissertation. To you both I offer a heartfelt *arigato*.

## VITA

December 3, 1969	Born, Los Angeles, California
1991	B.A., English Literature Reed College Portland, Oregon
1995–1997	English as a Second Language Instructor Saint Martin's College Lacey, Washington
1997–2000	English as a Second Language Instructor Nagoya YMCA Nagoya, Japan
2000–2002	Teaching Assistant Department of Linguistics University of California, Irvine
2003–2007	Teaching Assistant, Associate, Fellow Department of Linguistics University of California, Los Angeles
2005	M.A., Linguistics University of California, Los Angeles Los Angeles, California
2006	Dissertation Year Fellowship University of California, Los Angeles

## PUBLICATIONS AND PRESENTATIONS

- Martin, Andrew. January 2004. The structural nature of locality: gradient distance effects in Navajo sibilant harmony. Paper presented at the Annual Meeting of the Linguistic Society of America, Boston, Massachusetts.
- Martin, Andrew. November 2004. The origins of lexical biases: evidence from Navajo compounds and English naming preferences. Paper presented at the Western Conference on Linguistics, Los Angeles, California.

- Martin, Andrew. 2005. Loanwords as pseudo-compounds in Malagasy. In Jeffrey Heinz and Dimitris Ntelitheos (eds.), *Proceedings of the Twelfth Annual Conference of the Austronesian Formal Linguistics Association*. 287–295.
- Martin, Andrew. April 2007. Less-than-absolute ungrammaticality: geminate avoidance in English morphology. Paper presented at Generative Linguistics in the Old World 30, Tromsø, Norway.
- Martin, Andrew. To appear. The correlation between markedness and frequency: evidence from Latin and French. In Emily Efner and Martin Walkow (eds.), *Proceedings of the 37th Meeting of the North East Linguistic Society*.

# ABSTRACT OF THE DISSERTATION

The Evolving Lexicon

by

Andrew Thomas Martin

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2007

Professor Kie Zuraw, Chair

Although gradient phonotactics, phonological generalizations that are statistical rather than categorical, are a ubiquitous feature of human languages, current models of gradient phonotactics do not address the typological or diachronic aspects of these generalizations—why some phonotactic patterns are more common than others, and how and why these patterns change over time. In this dissertation I propose that the statistical properties of the lexicon are shaped in part by unconscious *phonotactic preferences* on the part of language users—biases that affect a word’s chance of becoming established among a community of speakers, or remaining in use once established.

The dissertation is devoted to establishing two main claims: first, that phonotactic preferences are real, and second, that the mechanism that drives these

preferences consists of competitions among words during speech production. In support of the first, empirical, claim, I present three main examples of phonotactic preferences: (1) a bias in favor of /b/ over /d/ in the development of Latin into French, which is motivated by articulatory ease, (2) a gradient OCP effect in English, motivated by processing ease, and (3) a correlation between tautomorphemic and heteromorphemic phonotactics in several languages, which I argue is motivated by the structure of the human phonotactic learning algorithm.

In support of the second, theoretical claim, that phonotactic preferences can result from competitions among synonyms, I present a spreading activation model of speech production (Dell 1986) which consists of a network in which lexical items which match concepts the speaker wishes to express are activated by those concepts, and then in turn activate their constituent phonological subparts. Synonyms, words that represent the same concept, are simultaneously activated and race to reach an activation threshold—the winner of this race is selected and used by the speaker to express the concept. Properties that allow a word to be accessed more quickly thus confer an advantage to words that have those properties. Words with phonotactic patterns that facilitate lexical access will tend to be used more than words without those patterns, leading over time to a lexicon in which these “good” patterns predominate.



## 1. Lexical competition and evolution

Licuit semperque licebit  
signatum praesente nota producere nomen.  
Ut silvae foliis pronos mutantur in annos,  
prima cadunt, ita verborum vetus interit aetas,  
et iuvenum ritu florent modo nata vigentque.

—Horace, *Ars Poetica*

[Men ever had, and ever will have, leave  
To coin new words well suited to the age,  
Words are like leaves, some wither ev'ry year,  
And ev'ry year a younger race succeeds.]

[tr. Earl of Roscommon]

Perhaps the most obvious type of historical language change, readily apparent to anyone who uses language, is the birth and death of words. In every language, some words “wither ev'ry year,” a phenomenon that most people can observe happening within their lifetimes. This is perhaps an inevitable Malthusian consequence of the ubiquitous human impulse to “coin new words well suited to the age.” The vocabulary of a language, limited by the memories and lifespans of its individual speakers, cannot expand indefinitely. As Charles Darwin (1871) put it, “We see variability in every tongue, and new words are continually cropping up; but as there is a limit to the powers of the memory, single words, like whole languages, gradually become extinct” (58).

As the quote from Horace demonstrates, interest in the life cycles of words has a long history. Despite this, however, the subject has played little role in modern linguistic theory, remaining the domain of lexicographers and amateur observers of language.<sup>1</sup> This dissertation represents an attempt to rectify this situation. In what follows I hope to show that studying the creation and survival of words in a speech

---

<sup>1</sup> Recent popular treatments of the subject include McFedries 2004, Kelz Sperling 2005, Crystal 2006, and Steinmetz and Kipfer 2006.



community offers important insights into historical language change, the nature of speech production, and the biases that humans bring to the task of learning a language.

### **1.1. The proposal**

Modern linguistic theory concerns itself with two fundamental questions: the *learning problem*—how do humans learn language?—and the *typology problem*—why do the world’s languages exhibit some properties and not others? These two questions represent two different, albeit complementary, approaches to understanding how the language faculty is structured. Although there is disagreement regarding how related learning and typology are, there is a consensus that a complete theory of human language should account for the facts in both domains.

Until recently, most research on these two problems has been confined to the categorical generalizations present in linguistic data. In recent years, however, a growing body of research has focused on *gradient phonotactics*, statistical rather than categorical sound patterns that hold over the lexicon. Although this work has been a welcome corrective to the categorical bias in the field, it has been largely restricted to answering the learning problem—determining the extent to which humans can learn statistical phonological patterns (e.g., Coleman and Pierrehumbert 1997, Dankovičová et al. 1998, Frisch, Pisoni, and Large, 2000, Treiman et al. 2000, Bailey and Hahn 2001, Frisch and Zawaydeh 2001, Hay et al. 2003). Much less work has been devoted to the typology problem—understanding why languages exhibit the patterns they do, or indeed why they have gradient phonotactics at all. This is the

question I will attempt to answer in this dissertation—what forces shape the statistical properties of a language’s lexicon?

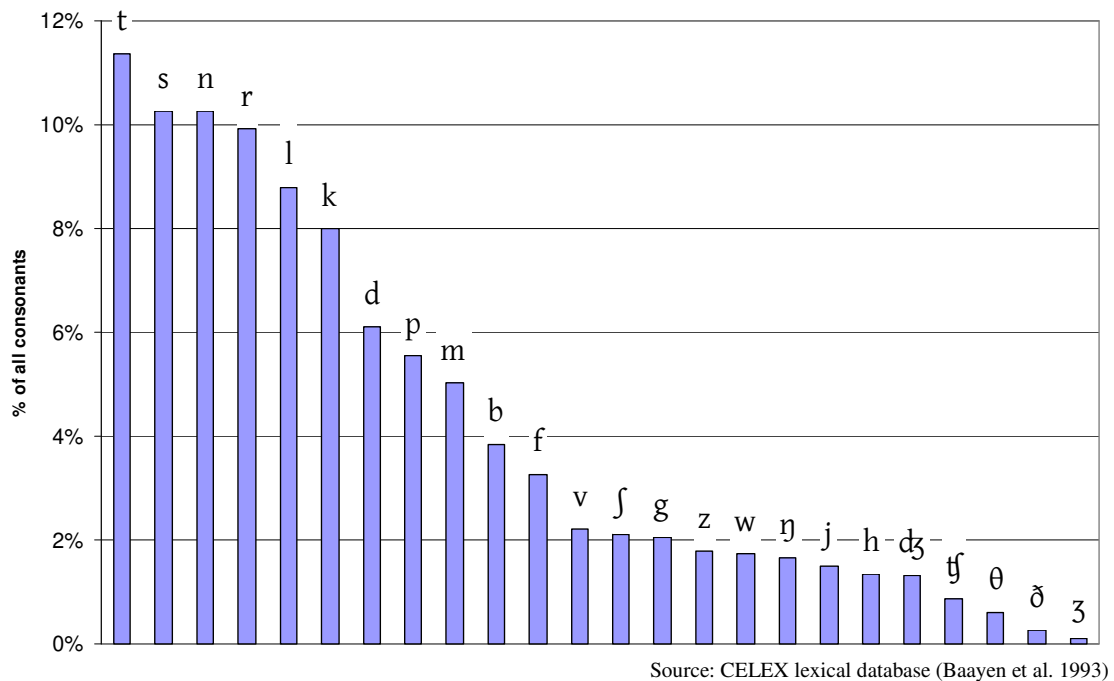
In every language that has been studied quantitatively, sounds exhibit a wide range of frequencies.<sup>2</sup> The distribution of consonants in English, shown in (1), has a typical shape—although English has 25 consonant phonemes, the five most frequent sounds, /t/, /s/, /n/, /r/, and /l/, together account for more than half of all the consonants occurring in English words (the bottom five, in contrast, make up just over 3% of all consonants). The distribution is thus not only non-uniform, but heavily skewed.<sup>3</sup>

---

<sup>2</sup> The frequencies reported throughout this section are type frequencies, i.e., the frequencies of segments in the lexicon, here represented by a lemmatized word list.

<sup>3</sup> Tambovtsev and Martindale (2007) argue, using data from 95 languages, that phoneme frequencies are best modeled with the Yule distribution (Yule 1924), of which the more famous Zipf distribution is a special case. The Yule distribution, they claim, is characteristic of discrete distributions over relatively small numbers of elements, such as the frequencies of phonemes in a language, while Zipfian distributions are more likely to occur with large numbers of elements, such as the frequencies of words in a language.

### (1) Consonant type frequencies in English



Similarly shaped distributions can be found in every language, not only for individual phonemes, but for cooccurrence frequencies among phonemes.

The fact that not all phonemes are created equal, although a striking linguistic universal, is by no means a logical necessity. Huffman (1952) showed that the average information content of a set of phonemes is maximized when they are all of equal probability, meaning that favoring some sounds over others results in a less efficient coding system. Why, then, are phoneme frequencies so skewed?

The answer proposed in this chapter is that lexicons with uniform phoneme distributions are diachronically unstable. I will argue that words that contain more frequent phonemes are more likely to enter the lexicon, thereby making these frequent sounds even more frequent, in a “rich get richer” feedback loop. Any slight

deviations from uniformity will thus become exaggerated over time, ultimately resulting in a skewed distribution.

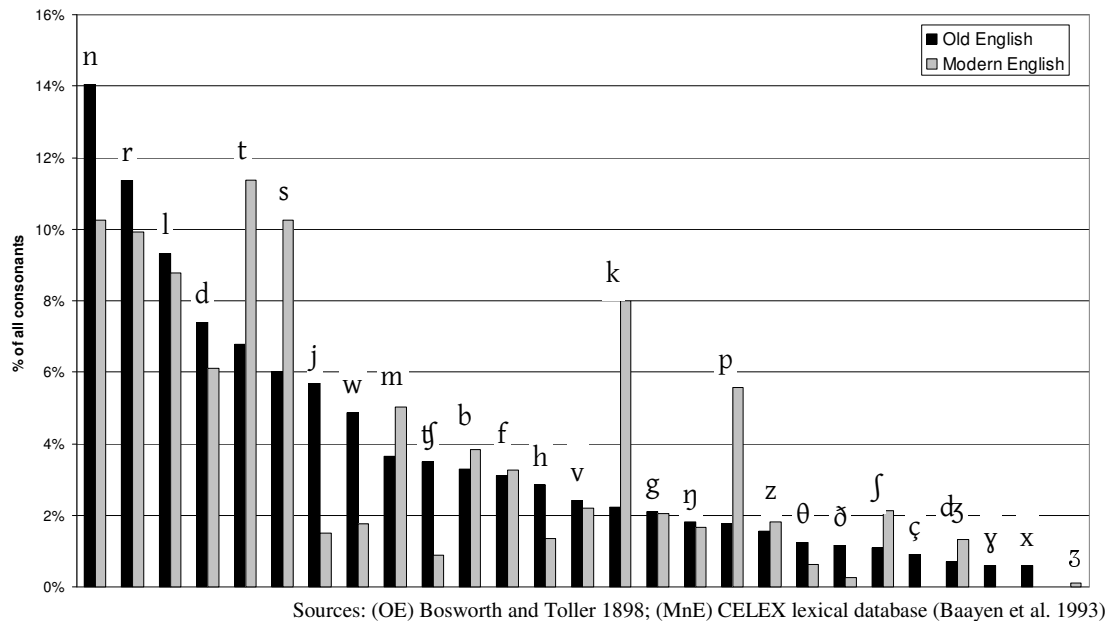
This mechanism, however, does not explain why certain sounds are consistently more frequent across languages, or across different historical stages of the same language. Markedness theory (Trubetzkoy 1931, Andrews 1990, Battistella 1990, de Lacy 2006) has generated a large literature devoted to establishing why languages seem to “prefer” some sounds over others—I will draw on this work, demonstrating the effects of several types of markedness on lexical frequency by means of a series of case studies. In chapter 2, for example, I present examples in which the articulatory ease of individual sounds is correlated with type frequency. In chapter 3 I argue that processing constraints on sound sequencing affects the frequency of certain sequences. In chapter 4 I focus on morphological word formation, and show that language-specific phonotactic generalizations also play a role in shaping the lexicon.

The dissertation will also present a theory of how these differences among sounds are translated into lexical frequencies. I will consider two possible mechanisms. The first involves historical sound change. If individual sounds change over time, and this change is biased towards creating certain sounds over others, the result would be a lexicon skewed as in (1) (Zipf 1935). If sound changes in English tend to create /t/ more often than /d/, for example, it would explain why /t/ is more frequent than /d/ in the current lexicon.

Although sound change undoubtedly has a hand in shaping the lexicon, I will argue that *phonotactic preferences* on the part of language users, which act as biases on the creation or retention of words, also play a role. On this theory, speakers of English prefer to retain (or borrow, or coin) words that contain /t/ over words that contain /d/, and these preferences over time have resulted in a lexicon skewed towards /t/. In other words, lexical items are afforded a better chance of spreading throughout a speech community, and of surviving across subsequent generations of speakers, if they contain certain sounds or combinations of sounds. These phonotactic preferences shape the statistical properties of the lexicon even in the absence of sound change. This idea is not new—it has been mentioned by Sevald and Dell (1994), Berg (1998), Boersma (1998), Frisch et al. (2004), Coetzee and Pater (2005), McClelland and Vander Wyck (2006), and Hansson (2007), and is the subject of recent work by Boersma (2007). What is lacking, however, and what I hope to provide, is a concrete theory of the mechanism underlying phonotactic preferences and its connection to language processing and phonotactic learning.

Understanding this mechanism will allow us to predict not only which kinds of statistical change are more likely than others, but also how statistical patterns can remain stable over time. For an illustration of the issues involved, it is instructive to compare the frequencies of consonants in Old English (spoken from the fifth to the twelfth century AD) to those in Modern English. The two distributions are shown below in (2), with black bars representing Old English frequencies and gray bars representing Modern English.

## (2) Consonant type frequencies in Old English and Modern English



There are several differences between the two distributions, reflecting the roughly 1,000 years between the two stages of English. The phoneme /j/, for example, is much more frequent relative to other consonants in Old English than Modern English. This is largely due to the common Old English verbal prefix *ge-*, pronounced /je/, which was later lost, depriving the language of many instances of /j/. Another difference can be seen in the ratios of voiced to voiceless obstruents. The obstruents /p/, /t/, /k/, and /s/ all increase greatly in frequency between the two stages of English. This is largely due to the disappearance of a rule of intervocalic voicing—in Old English, these obstruents could only occur in a limited set of contexts, which kept their frequency down. In Modern English, which has lost this rule, voiceless obstruents occur in a greater range of environments, and have concomitantly greater frequencies.

What is perhaps more surprising than the differences between the distributions in Old and Modern English, however, is how similar they are. It is well-known that between these two stages of English, most of the lexicon was replaced—roughly 85% of the Old English vocabulary is no longer in use (Baugh and Cable 1993), and more than 80% of the Modern English vocabulary consists of words borrowed from other languages (Stockwell and Minkova 2001). Given this large turnover in lexical items, the frequency distribution of consonants in both languages is surprisingly similar in a number of ways—for example, alveolar consonants are clustered at the top, while interdental fricatives are near the bottom.

This similarity can be quantified by means of the Kullback-Leibler (KL) divergence, which is a measure of the difference between two probability distributions over the same space.<sup>4</sup> Identical distributions have a KL number of zero, while higher numbers indicate larger differences between the distributions. If we restrict the set of consonants to those shared by both languages (a total of 23), the resulting KL divergence between Old English and Modern English is 0.26. This is significantly lower ( $p < .001$  by Monte Carlo) than the average KL divergence of 1.01 obtained when the rank order of the Modern English consonants is randomly shuffled 10,000 times.<sup>5</sup> The degree of similarity between the two distributions is thus unlikely to be due to chance.

---

<sup>4</sup> The Kullback-Leibler divergence for two probability distributions  $p$  and  $q$  is defined as  $D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$ . Note that KL divergence is not a symmetric property;  $D(p \parallel q)$  does not necessarily equal  $D(q \parallel p)$ . The values reported here are calculated from Old English ( $p$ ) to Modern English ( $q$ ), although similar results are obtained if the calculation is done the other way.

<sup>5</sup> This Monte Carlo test of significance is described in detail in §3.2.1.

A frequency distribution can thus remain relatively stable over time even as the specific contents of the lexicon instantiating that distribution are replaced, a property of languages that Lahiri (2002) has called *pertinacity*. Given the high rate at which words are created and become obsolete, a lack of statistical change requires explanation just as much as cases of change do. The theory advanced in this dissertation, therefore, is as much a theory of language stasis as of language change.

My strategy throughout will be to model these phenomena with a minimum of new machinery. That is, I will show that the existence and nature of phonotactic preferences is largely predicted by cognitive models that have been argued for on independent grounds. At every step I hope to show that the model I develop is justified not just because it accounts for the data I present here, but because it also accords with evidence from other sources, such as speech errors or data collected in experimental settings.

## **1.2. Natural selection in language**

The theory of lexicon change I propose has striking parallels with evolutionary theory. Just as organisms compete to survive and reproduce, words compete to be used by speakers of a language. This research thus joins a long list of attempts to make sense of linguistic data within the framework of Darwinian natural selection.<sup>6</sup> Recent examples of this approach include Lass 1990, McMahon 1994,

---

<sup>6</sup> There is a long debate over the question of whether the theory of natural selection can be properly applied to anything other than biological organisms (e.g., Dawkins 1976, Hull 1988, Dennett 1995,



Niyogi and Berwick 1997, Haspelmath 1999, Croft 2000, Pulleyblank and Turkel 2000, Kirby and Hurford 2001, Nowak and Komarova 2001, Redford and Miikkulainen 2001, the papers in Briscoe 2002, Blevins 2004, Ritt 2004, Wedel 2006, and Niyogi 2006.

Although these accounts differ widely in their theoretical background and assumptions, they share the postulate that linguistic units or patterns are Darwinian replicators—their survival depends on their ability to be copied, that is, internalized by other speakers of the language. “A linguistic regularity survives because it has properties that make its faithful replication easy,” as Brighton et al. (2005) put it. Darwin himself subscribed to this view—in *Descent of Man* (1871) he writes that “[t]he survival or preservation of certain favoured words in the struggle for existence is natural selection” (58-59).

For a replicator to evolve, two mechanisms are necessary: *variation* and *selection* (Dennett 1995). The properties of a population of replicators must vary, and some properties must be correlated with a higher chance of replication than others. Within a population of replicators for which these two conditions hold, over time the more successful variants will come to outnumber, or completely replace, the less successful variants.

The accounts of linguistic evolution cited above differ in the types of replicator, and types of variation among replicators, they consider; for example,

---

Plotkin 1995, Blackmore 2000, Richardson and Boyd 2004, Croft 2006, Andersen 2006, Mesoudi et al. 2006). My point in this section is not that the parallels between language and biology are exact, but simply that applying the tools of evolutionary biology to language change can produce insight unobtainable through traditional linguistic theory.

phonetic variation among outputs for a single lexical item (Blevins 2004), or variation in parameter settings among competing grammars (Niyogi 2006). My account of phonotactic preferences will consider lexical items to be replicators, and variation to mean variation across lexical items—in short, words compete with other words. In the next section, I examine the forces that drive the competition among these replicators.

### **1.3. Lexical competition**

The operation of natural selection is predicated on the limited availability of resources. If there are not enough resources to support every organism in an environment, some will die without reproducing; those that are better able to obtain resources are more likely to be among those that survive and reproduce. In the case of competing lexical items, I will argue that the resource in question is the concepts to which words refer.

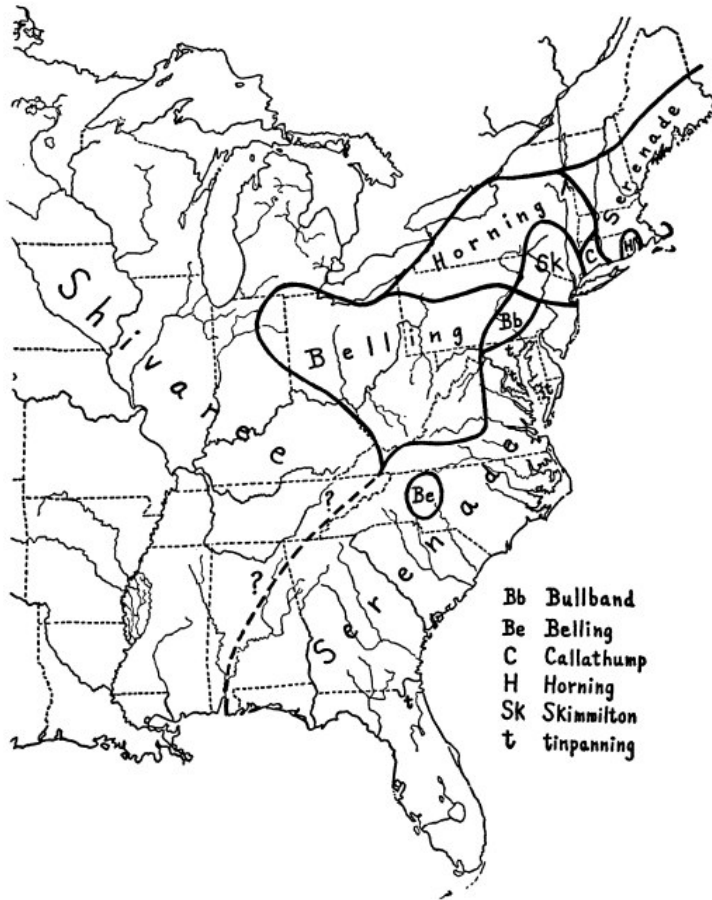
Competition between words is driven by a pressure, both within speech communities and within individuals, to express each concept with a single word; that is, a pressure to avoid exact synonymy. As Baronchelli et al. (2006) put it, lexical innovation is typically marked by “a period in which novelty spreads and different words compete, followed by a dramatic transition after which almost everyone uses the same word.” This tendency, also noted by Lass (1997), has been called the “First Law of Propagation” by Croft (2000), which he describes as the “natural human tendency for a community to select one alternative as the conventional signal for a

recurrent coordination problem” (176). I will refer to this tendency as *vocabulary convergence*.

An example of vocabulary convergence comes from Davis and McDavid’s (1949) work on English dialect geography. Their paper focuses on the word *shivaree*, a word referring to “a noisy burlesque serenade used chiefly as a means of teasing newly married couples” (249), which was borrowed into English from French *charivari* some time in the eighteenth century (the earliest citation in the *OED* occurs in 1805; presumably the word came into common use at some point prior to that).

Davis and McDavid find that although *shivaree* is used throughout Canada and most of the United States, there are areas on the American eastern seaboard in which the same folk ritual is referred to by a host of other terms, including *bellringing*, *tinpanning*, *serenade*, and *callathump*; they summarize the distribution of the competing words in a map, reproduced here in (3).

(3) *Shivaree* and its competitors<sup>7</sup>



Davis and McDavid point out that the greatest variety is found in the areas of the U.S. that have been longest settled by English speakers. In these regions, relatively stable populations allowed multiple words to each acquire and retain a foothold in various communities. West of the Mississippi, however, where populations were made up of settlers from many different areas, all with different vocabularies, no one term had the advantage of tradition or entrenchment among a large number of users. One word, *shivaree*, was able to dominate the usage of this large and heterogeneous group of speakers.

<sup>7</sup> From Davis and McDavid 1949.

Although this seems a plausible explanation of how one word came to dominate its potential competitors, it does not explain why it was *shivaree* in particular that emerged as the ultimate winner in this competition. Of course, many factors determine whether a word will successfully outcompete its rivals. The existing literature on this topic has primarily addressed the social factors—how the properties of a word’s users (social status, age, gender, etc.) affect the word’s success (Weinreich et al. 1968, Milroy 1987, Labov 1973, 2001). Croft (2000, 2006), in fact, argues that these are the *only* factors that play a role in the selection of linguistic variants. The purpose of this dissertation is to examine the extent to which the inherent properties of the word itself—more specifically, its phonotactic properties—contribute to its entrenchment in a speech community.

In a comment on the Davis and McDavid paper, Bolinger (1950) addresses the question of *shivaree*’s success, suggesting that the word’s phonological properties gave it an advantage. The final stressed [i:] in *shivaree*, he argues, although unusual for English, is characteristic of words like *jamboree*, *jubilee*, *spree*, and *whoopee*, which have in common a festive connotation that is shared by the meaning that *shivaree* was competing to express.

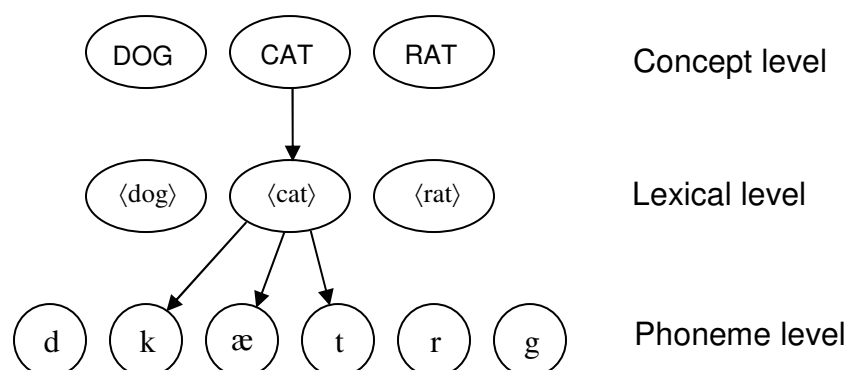
I will have more to say about exactly how the final vowel in *shivaree* could have increased its likelihood of success in §1.7. For present purposes, this example is intended to demonstrate both that there is pressure towards the use of a single word for each concept within a community, and also that the competition fostered by this pressure can be biased by the phonological properties of the word.

#### 1.4. Deriving convergence

In the previous section I claimed that a speech community will tend to converge on a single term for a single concept, and that this tendency drives competition among words. In this section I show that vocabulary convergence is a predicted outcome if we assume that lexical selection—the choice of a word from among a set of synonyms—is governed by the structure of the speech production system.

Most current models of language production (e.g., Stemmerger 1985, Dell 1986, Levelt et al. 1999) assume that the process of producing an utterance involves the selection of lexical items that correspond to the concepts the speaker wishes to express. These models typically consist of an associative network of linked nodes, each of which has an activation level, represented by a numerical value. Nodes become more active when the nodes they are connected to are active, allowing activation to spread through the network. A schematic example of such a network is shown in (4).

(4) Feedforward speech production network



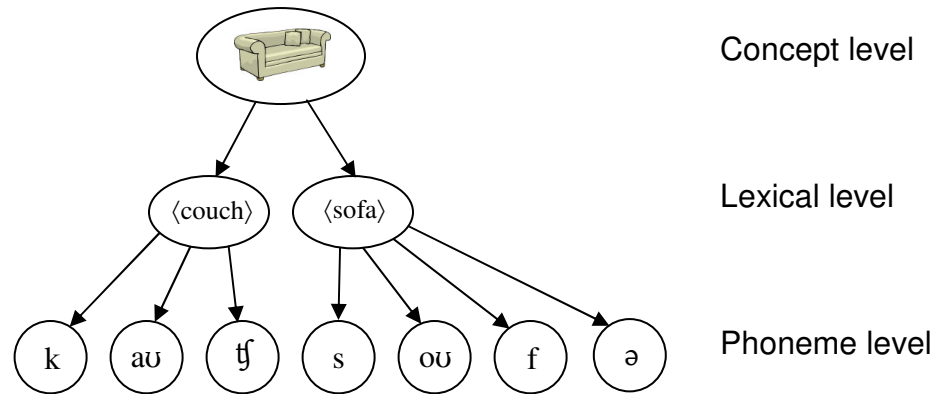
In this example, the concept CAT is first activated, representing the speaker's desire to express that concept. Activation spreads from the concept to the lexical entry ⟨cat⟩, whose semantic features match those of the concept—this is the stage at which a lexical item appropriate to the concept is selected. The lexical entry is in turn linked to the phonemes that make up its phonological representation, and so activation spreads to /k/, /æ/, and /t/—the selected lexical item is thereby phonologically encoded in preparation for speech.

Most such models simplify by assuming that there is a single lexical entry corresponding to the active concept. The point of these models is to determine how the correct item is selected out of the entire set of lexical entries, or in the case of speech errors, how an incorrect lexical item is selected instead. What happens, though, when there is more than one correct lexical entry—i.e., a set of synonyms—for the intended concept?<sup>8</sup> In such a case, the network would resemble the one in (5).

---

<sup>8</sup> I have not defined precisely what qualifies two words to be synonyms. This will depend on the nature of semantic representations. If they consist of sets of semantic features, then it is likely that synonymy is a gradient property; the more features shared by two lexical items, the greater the chance that they will be activated simultaneously, and thus compete. Atomic lexical concepts, on the other hand, should result in categorical synonymy—two lexical items would be either synonymous or not. The theory of lexical competition I advance here does not require a commitment to the exact nature of synonymy, however.

(5) Activating synonyms



Because both synonyms presumably match the semantic features of the active concept, the spreading activation model predicts that both are simultaneously activated. Evidence for this multiple activation comes from speech errors known as *blends*, in which two words are conflated into one (e.g., *frowl* from *frown* and *scowl*). These errors, which are typically formed from synonyms or near-synonyms (Wells 1951, Fromkin 1971, Poulisse 1999), could occur when two lexemes are equally activated—their respective phonological plans also become equally activated, and if one does not emerge as a clear winner, the production system may conflate them.

In the absence of a speech error, however, only one word will eventually be selected and pronounced. What decides which synonym is chosen? One way to model the selection process is as a kind of race, in which the first lexical entry whose activation reaches a certain threshold is selected. This is the model assumed in most accounts of speech errors—the intended lexical entry is usually activated first, but occasionally, because of noise in the system, an extraneous lexical entry may win the race and be pronounced instead (Dell 1986, Levelt 1989). In the case of synonyms,



choosing one over the other is considered a lexical choice rather than an error, but, I am suggesting, the mechanism is the same. This claim is at the center of the dissertation—in a set of synonyms, those words that can be accessed most quickly will be used most often. Where there is independent evidence that some property of words contributes to faster lexical access, the theory thus predicts that words with those properties should be more frequent than synonyms without them.<sup>9</sup>

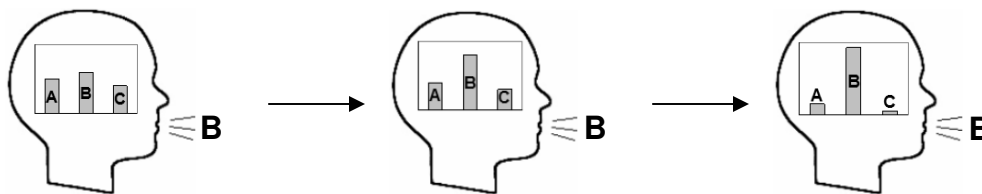
Looking at one such property, word frequency, will provide an answer to the question asked at the beginning of this section—why speech communities converge on a single lexical item for a given concept. Reaction time data from picture naming and lexical decision tasks suggests that more frequent words are accessed more quickly (Oldfield and Wingfield 1965, Jescheniak and Levelt 1994). Evidence from word-substitution speech errors also supports the view that higher frequency results in faster lexical access: errors are more likely to occur on low-frequency items (Stemberger 1984, Harley and MacAndrew 1992, 1995, Vitevitch 1997, 2002), and intended targets are more likely to be replaced with higher-frequency than lower-frequency words (Del Viso et al. 1991, Vitevitch 1997). In the network model of speech production, this can be modeled by assigning each lexical node a resting activation, and assuming that this resting activation increases each time that node is selected during production. Words with higher resting activations will reach the activation threshold more quickly, and thus have an advantage over synonyms with lower resting activations.

---

<sup>9</sup> Note that the theory predicts only that speed of access will correlate with the relative frequency of synonyms—it says nothing about the relative frequencies of unrelated words.

In this model, having a high frequency will increase a word's resting activation, which will further increase its frequency, in turn increasing its resting activation, and so on. This feedback loop results in a “rich get richer” effect in which any word with a slightly higher resting activation than other synonyms will eventually come to be exclusively used by the speaker to express the concept. A state in which a speaker knows several synonyms for a single concept and uses all of them with equal frequency is thus inherently unstable—any difference in the resting activations of the synonyms will be magnified until only one is used (or alternatively, some of the synonyms could have their meaning altered so that they are not exact synonyms). The process, which is schematically illustrated in (6), can be thought of as vocabulary convergence within a single speaker.<sup>10</sup>

(6) Within-speaker convergence (bars represent resting activations for synonyms A, B, C)



This model can explain why exact synonyms do not tend to occur in the lexicons of individuals, but does not account for how an entire speech community comes to converge on a single lexical item. Why does this tendency towards

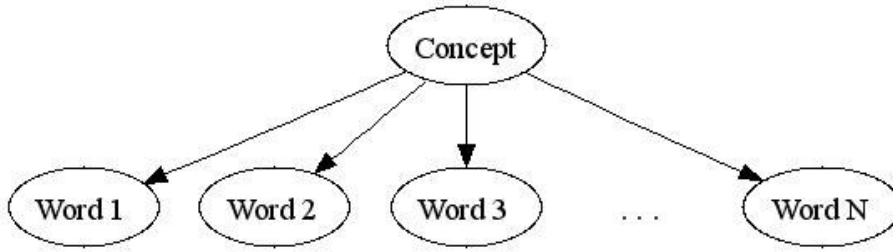
<sup>10</sup> This within-speaker convergence could alternatively be driven by an explicit cognitive bias against exact synonyms. It is well established that when learning new words, children typically assume that a new word will not have exactly the same meaning as a word they already know (Clark 1993, Pinker 1984, Chomsky and Lasnik 1977, Markman 1984). This bias could be responsible for convergence, but it could also be seen as an evolved response to convergence. If the rarity of exact synonyms is a ubiquitous emergent property of languages, a learning strategy which assumes that no words are exact synonyms would reduce the semantic hypothesis space at a relatively low cost in terms of errors. Which came first—the absence of synonyms or a bias against synonyms—is a chicken-and-egg problem that I will not attempt to resolve here.

vocabulary convergence exist? Croft (2000) argues that it results from the desire of speakers to identify with a given social group. But recent research suggests that vocabulary convergence is a naturally emergent property even in very simple populations of communicating agents in which social identity plays no role.

Steels et al. (2000), for example, describe a simulation in which large numbers of agents interact with each other by playing a “naming game” in which agents take turn “pointing to” objects in the simulated world and describing them to other agents using words from their lexicons. Although there is no explicit pressure against synonymy built into the simulation, for each object one word eventually comes to dominate the population. More recent work has explored the precise mathematical conditions under which this convergence occurs (Baronchelli et al. 2005, Wang et al. 2006).

The convergence process is best illustrated with a simple simulation of a speech community consisting of  $N$  agents. Each agent is equipped with a “concepticon” consisting of a single lexical concept that it must communicate to the other agents, and a lexicon consisting of  $N$  words that each express this same concept. All agents have the same concept and same set of synonyms, instantiated in a spreading activation speech production network as in (7) (the phoneme level is not implemented in this simulation).

(7) Convergence simulation network



Each agent's network is implemented using the model proposed in Dell 1986: lexical selection (i.e., the selection by a speaker of an appropriate lexical item for the intended concept) is simulated by giving the intended concept an arbitrary activation level of 100, and then allowing activation to spread through the network in a series of discrete time steps. After a specified number of time steps, the lexical node with the highest activation is selected. During each time step  $t_i$ , the activation level of node  $j$ ,  $A(j, t_i)$ , is calculated with the equation in (8).

(8) Calculating activation (modified from Dell 1986)

$$A(j, t_i) = w_j [A(j, t_{i-1}) + \sum_{k=1}^n p_k A(c_k, t_{i-1})] (1 - q) + noise$$

where  $c_1, c_2, \dots, c_n$  are all of the nodes connected to  $j$ ,  $p_1, p_2, \dots, p_n$  are the weights of the connections, and  $q$  is the decay rate.<sup>11</sup> The noise added to the activation level is a value sampled from a Gaussian distribution with a mean of 0 and a standard deviation equal to 0.05 times the node's previous activation level  $A(j, t_{i-1})$ .

Where the implementation of the network differs from that of Dell 1986 is in the addition of weights for each node, which is intended to simulate differing resting activations. The weight of node  $j$  (indicated by  $w_j$  in (8)) acts as a multiplier on the

<sup>11</sup> In all of the simulations reported in this chapter,  $p$  was set to 0.3 for all connections, and  $q$  was set to 0.6.

activation  $A_j$  at each time step; nodes with higher weights will have a higher probability of being selected. Because of the noise in the calculation of activation, however, even nodes with lower weights may be selected—the set of lexical node weights thus determines a probability distribution over the set of synonyms. The simulation begins with all weights set to 1.0 for all agents.

The simulation proceeds as follows: during each round of the simulation, the agents are paired randomly, and each agent tells its partner a word from its lexicon, using its current network to determine which synonym to choose. Once an agent has produced a given word, the node for that word has its weight  $W_s$  increased according to the formula in (9). The agent's partner, who has just heard the word, has the node's weight  $W_L$  increased according to the formula in (10).

(9) Speaker weight adjustment

$$W_s \leftarrow W_s + \alpha(e^{-W_s})$$

(10) Listener weight adjustment

$$W_L \leftarrow W_L + \beta(e^{-W_L})$$

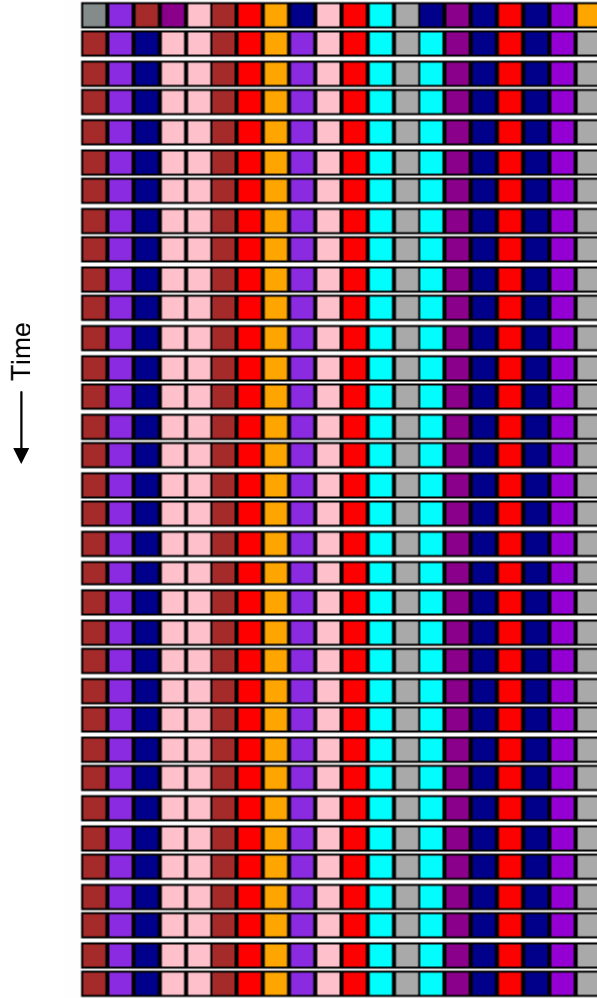
The amount added to each weight is a function of the current weight; as the weight increases, the amount added each time decreases. At the end of each round of the simulation, when all the agents have interacted with their partners, all weights are multiplied by 0.99—weights of nodes that do not get used therefore steadily decay over the course of the simulation.

Words which an agent either uses or hears being used are strengthened, and thus have a higher probability of being used in the future. The values of the

parameters  $\alpha$  and  $\beta$  determine the relative contributions of speaking and hearing to this strengthening. When these parameters have the same value, agents modify their activations equally regardless of whether they have used the word or heard it from another agent. If  $\alpha$  is equal to or greater than  $\beta$ , agents will value their own usage over that of others—I will call these selfish agents. If  $\beta$  is sufficiently greater than  $\alpha$ , agents will be more affected by others' usage than their own—I will refer to these as cooperative agents.

The diagram in (11) shows what happens when 20 selfish agents interact for 34 rounds. Each row in the diagram represents a single round, while each square represents one of the agents. The color of each square indicates which word the agent used during that round.

(11) Set of selfish agents ( $\alpha > \beta$ ): each agent settles on a different word

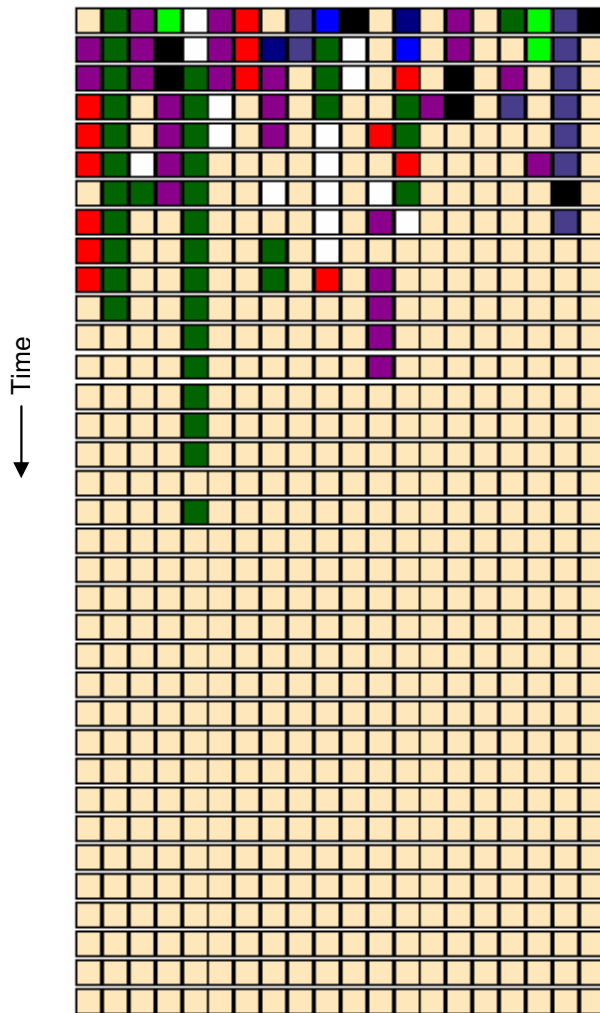


It is clear from the diagram that although each agent converges on its own single word from among the 20 synonyms, there is no systematic agreement across agents. This is a consequence of the fact that once an agent converges on its own word, it will use that word every round, further strengthening its activation. Because the agent is paired with a randomly chosen agent each round, it will hear a variety of other words, but each only a fraction of the time it uses its own favored word. The result is

“selfish” behavior in which each agent settles on its own preferred synonym independently of the rest of the community.

Selfish behavior also results when  $\alpha$  is equal to  $\beta$ , for the same reasons—an agent will always use its own word more often than it hears any other individual word. If  $\beta$  is set to a value sufficiently higher than  $\alpha$ , however, a different behavior emerges. The diagram in (12) shows the results of a simulation with 20 such cooperative agents (in this simulation,  $\alpha$  is 1.0 and  $\beta$  is 3.0).

(12) Set of cooperative agents ( $\beta > \alpha$ ): agents come to agree on a single word





When  $\beta$  is greater than  $\alpha$ , agents pay more attention to what they hear than what they say. When this is the case, the entire population eventually converges on one word, although which word is ultimately successful is a matter of chance. Convergence is thus predicted to emerge as long as the probability of a speaker using a given word is a function of how often the speaker hears others using the word.<sup>12</sup> Evidence from the sociolinguistic literature suggests that speakers do in fact accommodate many features of their speech, including lexical choice, to match their interlocutors (Giles and Smith 1979).

If vocabulary convergence is a near-inevitability, we might wonder how dialectal difference is maintained at all. Why didn't *shivaree*, for example, come to dominate the entire United States? Stable dialects can be modeled by slightly modifying the simulation—if a spatial dimension is added, and agents' interactions are limited to “nearby” agents, multiple dialect areas can emerge, each converging on a different word (Livingstone and Fyfe 1999, Livingstone 2002).

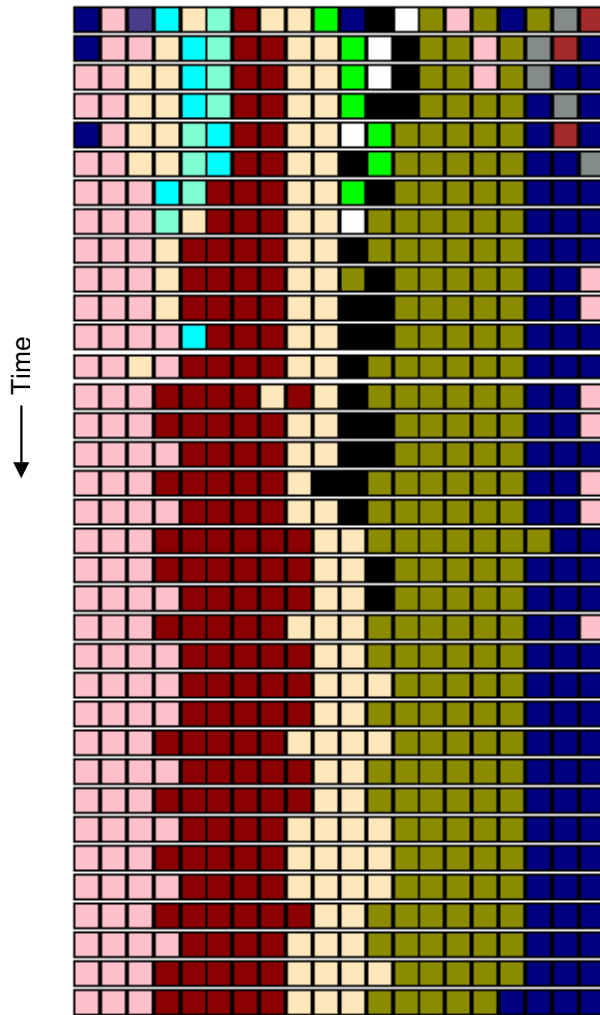
In this version of the simulation, agents are organized in a circle, with each agent having two neighbors, one on either side. Each round, every agent chooses at random to interact with either the neighbor on its left or right; the simulation otherwise proceeds as described above. When agent interaction is limited in this way, convergence to a single word is not guaranteed. Rather, different subgroups of agents can converge to different words, as shown in (13) (note that the leftmost agent is

---

<sup>12</sup> Note that in my simulation communication between agents is assumed to be perfect—there is no chance that one agent will misunderstand another. These results therefore suggest that convergence is not necessarily the result of a pressure to communicate effectively.

considered to be adjacent to the rightmost, so that each agent has exactly two neighbors).

(13) Set of cooperative agents ( $\beta > \alpha$ ) who talk only to neighbors: dialects emerge



These results suggest that dialect areas are stable when speaker interaction is relatively local, whether spatially or socioeconomically. In such a case, a lexical item may be protected from a competitor by the network of speakers that use the item, and who thus continually reinforce each others' lexical entries for the established word. When speakers from many different dialect areas mingle, however, as happened in

the settlement of the western United States, the stage is set for one lexical item to eliminate its synonymous competitors. Just such a scenario seems plausible in the *shivaree* case.

The simulations described in this section show that vocabulary convergence can emerge in a community of speakers provided that two conditions hold. First, hearing a word must have some effect on the probability of using that word in the future—speakers cannot, for example, have entirely separate perception and production lexicons. Second, the probability that a speaker will use a word must depend more strongly on the usage of others than on the speaker’s own usage. Given these conditions, and a speech production system that is structured as I have described it here, vocabulary convergence is an expected consequence.

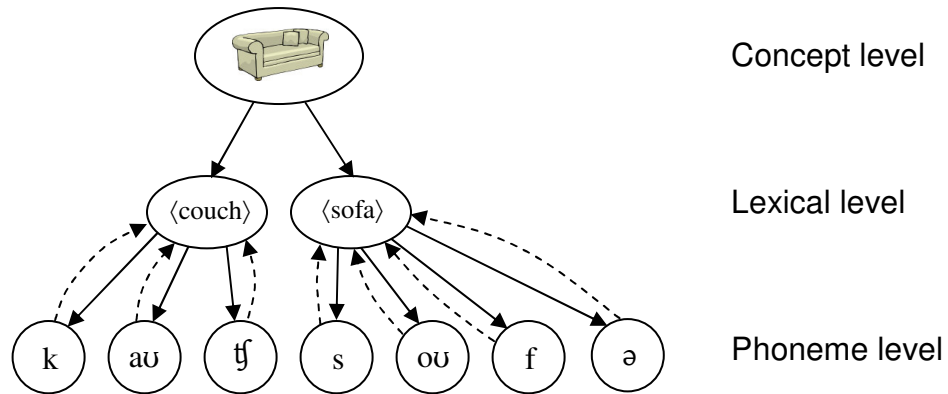
### **1.5. The role of feedback in lexical selection**

Nothing in the model so far predicts that the sounds that make up a word contribute to its being selected over other synonyms. Such a prediction is only made if the strictly serial character of the model is relaxed, and feedback is permitted from later levels to earlier levels. In *interactive* models of speech production, activation spreads not only “downstream,” from concept to lexical entry to phonemes, but also spreads back from the phoneme level to the lexical level, as depicted in (14).<sup>13</sup>

---

<sup>13</sup> The extent to which feedback is a necessary part of the model is a subject of intense debate in the speech production literature. For a concise overview of the range of possible positions, see Rapp and Goldrick (2000).

(14) Network with feedback



If we assume that phoneme nodes, like lexical entries, have differing resting activations, then some phonemes will reach threshold more quickly than others. These “better” phonemes will feed activation back to their lexical entries, giving those words an advantage in reaching threshold more quickly than others.<sup>14</sup>

There is substantial evidence that feedback of this kind is a characteristic of the speech production system. In speech errors involving word substitution, for example, the word actually produced tends to be phonologically similar to the intended target (e.g., saying *button* for *butter*; Fromkin 1971, Martin et al. 1996). This suggests that the lexical entry for the intended target activates the phonemes contained in the word, and that those phonemes in turn activate all of the lexical entries that contain them. Thus, a word which contains many of the same phonemes as the target and matches some of its semantic features (thus receiving activation

---

<sup>14</sup> Throughout this section, I will assume that phoneme nodes each refer to a single segment, but of course in an actual speech production network they may refer to more complex units such as syllables (Levelt and Wheeldon 1994) or simpler units such as phonological features (Dell 1986).

from the concept node) has the highest chance of being mistakenly selected (although see Levelt et al. 1999 for an alternative, feedforward account of these effects).

Peterson and Savoy (1998) also report on an experiment in which subjects are shown a picture, and then presented with a word which they are asked to pronounce. When shown a picture of a couch, subjects are faster at producing both *couch* and its synonym *sofa*. However, production of both *count* and *soda* following the same picture is also speeded, demonstrating not only that both synonyms are selected, but that both activate phonologically similar lexical entries. This is explained in the feedback model as lexical entries activating each other via the phonemes they have in common. Jescheniak and Schriefers (1998) achieved results consistent with this in an experiment in which subjects heard a distractor word, and then named a picture—the distractor *soda* was shown to interfere with the naming of a picture of a couch, even when the subjects used the word *couch* to describe the picture.

In another experiment, Ferreira and Griffin (2003) found that subjects in a picture-naming task, when shown a picture of a priest, would sometimes mistakenly produce *nun* as a description of the picture when the picture was preceded with presentation of a sentence containing the semantically unrelated but homophonous word *none*. On the basis of these results they argue that “semantic factors and phonological factors can affect lexical selection jointly” (90).

Several studies by Vitevitch (1997, 2002) and Vitevitch and Sommers (2003) have shown, using evidence from speech errors, picture naming, and tip-of-the-tongue states, that words are produced more quickly and accurately by English

speakers if they have many phonological neighbors. Dell and Gordon (2003) show that these effects are predicted in a network equipped with feedback from the phoneme to the lexical level. When the concept CAT activates the lexical entry ⟨cat⟩, for example, neighbors such as ⟨hat⟩ and ⟨cap⟩ are activated via the phonemes they share with the target. These neighbors are unlikely to be selected themselves, as they receive no activation from the concept level, but they do send activation back to their constituent phonemes, which in turn send more activation back to the target entry ⟨cat⟩. In other words, phonemes that are connected to many lexical entries exert more influence, because of the activation they receive from these entries, than phonemes connected to few lexical entries.<sup>15</sup>

These and other similar studies are part of a growing body of evidence that speakers access phonological characteristics of candidate words at an early stage of lexical selection (*contra* the strictly serial model of Levelt et al. 1999, in which phonological characteristics are only accessed once a single lexical entry has been selected).

Given this model, it is clear how *shivaree* could have benefited from the existence of words that were both phonologically similar (in having a final stressed [i:]) and semantically similar (in referring to festive, boisterous events). Once the SHIVAREE concept is activated, words like *jamboree* and *whoopee* would also be

---

<sup>15</sup> In recent work, Vitevitch and Stamer (2006) have found the opposite effect in Spanish: words with *fewer* neighbors are produced more quickly. They speculate that this is due to the highly inflectional nature of Spanish as opposed to English, one consequence of which is the fact that the neighbors of a Spanish word are more likely to be morphologically related to it than would be true for an English word. As Vitevitch and Stamer point out, it is unclear whether these effects can be accounted for by a spreading activation model of speech production. More research on this issue, involving a wider range of languages, is clearly called for.

somewhat activated due to the semantic features they have in common with the concept. They in turn would activate their constituent phonemes, which would then boost the activation of *shivaree*, giving it an advantage relative to other words that lack such allies.

This process could be responsible for the existence of phonesthemes (Firth 1930, Magnus 2001), seemingly arbitrary sound-meaning pairings across groups of words, of which the final stressed [i:] in *shivaree* is an example.<sup>16</sup> An initial, accidental correspondence between sound and meaning among a few words might serve as an attractor that could bias future lexical competitions in favor of words that share the correspondence. Hock and Joseph (1996) call this *phonesthematic attraction*, and cite as an example the case of Early Modern English *sacke* changing to Modern English *sag* due to attraction from *drag*, *flag*, and *lag*.

This processing-based theory of phonesthemes makes the prediction that among words with the same number of phonological neighbors, those words that have a greater number of semantically related neighbors will be accessed more quickly. Further research would be required to test this hypothesis, and distinguish it from other possible theories of phonesthemes, for example, one that views phonesthemes as pseudo-morphemes.

---

<sup>16</sup> Other phonesthemes in English include initial *gl-*, in words like *glisten*, *glow*, *glimmer*, *glitter*, and *glance*, all having to do with light or vision, and initial *sn-*, in words like *snout*, *snarl*, *snot*, *snort*, and *snore*, having meanings related to the nose (Bloomfield 1933).

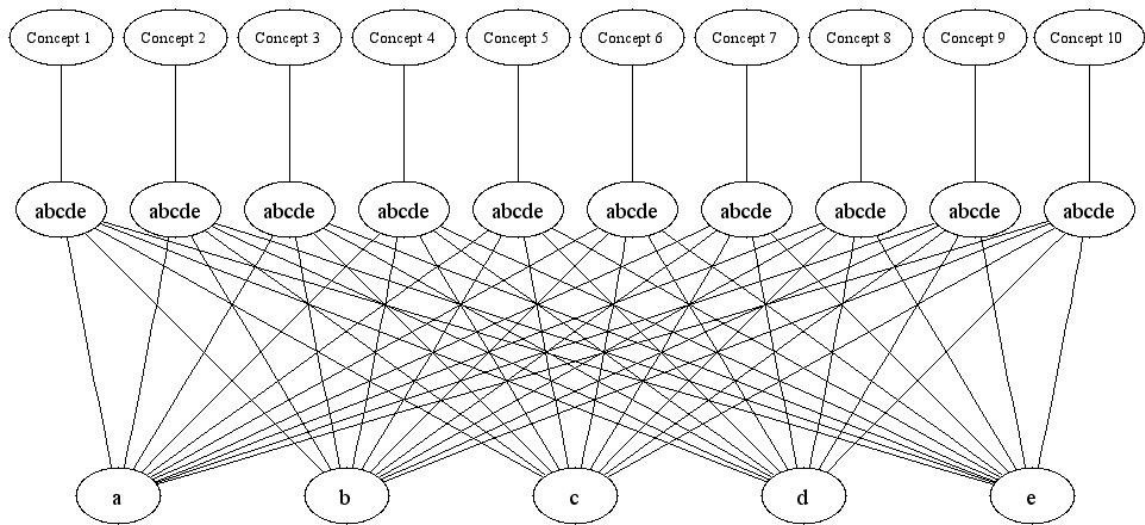
## 1.6. The evolution of frequency distributions

At the beginning of this chapter, I asked why it is that phoneme frequencies in a language tend to exhibit a skewed frequency distribution, with a small handful of phonemes accounting for most of the segments in the lexicon. We can now answer that question using the speech production model described above, by considering what would happen if a hypothetical language had its phonemes uniformly distributed. The remainder of this section describes the results of a simulation which demonstrates that given this model of speech production, a uniform distribution is unstable; over time, it will evolve into a more skewed distribution.

Having established in the previous section that a community of agents equipped with speech production networks will converge on a single word for each concept, this simulation will model a single speaker in lieu of a group of interacting agents, on the assumption that the results will scale up to an entire community of speakers. The simulation involves a hypothetical language with five phonemes (arbitrarily labeled *a*, *b*, *c*, *d*, *e*) in which all words are five segments long. The lexicon of this language consists of 10 words, each representing a different concept—in the initial state, all words have the same form, *abcde*, guaranteeing that all of the phonemes are equally frequent. Each concept, lexical entry, and phoneme is represented by a node in a spreading activation network, organized as in (15).



(15) Network initial state



Each line in the graph represents a bidirectional connection between two nodes along which activation may spread; for each phoneme in a word, there is a connection from the word's lexical node to the node for that phoneme.<sup>17</sup> Each lexical node is also connected to the node for its corresponding concept.

The simulation is intended to model the evolution of a lexicon over time. In each “generation,” each of the words in the current lexicon in turn is confronted with a randomly generated synonym (made up of five phonemes drawn from a uniform distribution over the phoneme inventory). Both the existing word and the synonym are connected to the same concept node, and the network is used to determine which word is selected—activation is allowed to spread for a fixed number of time steps, after which the lexical node with the highest activation is selected. The activation for

---

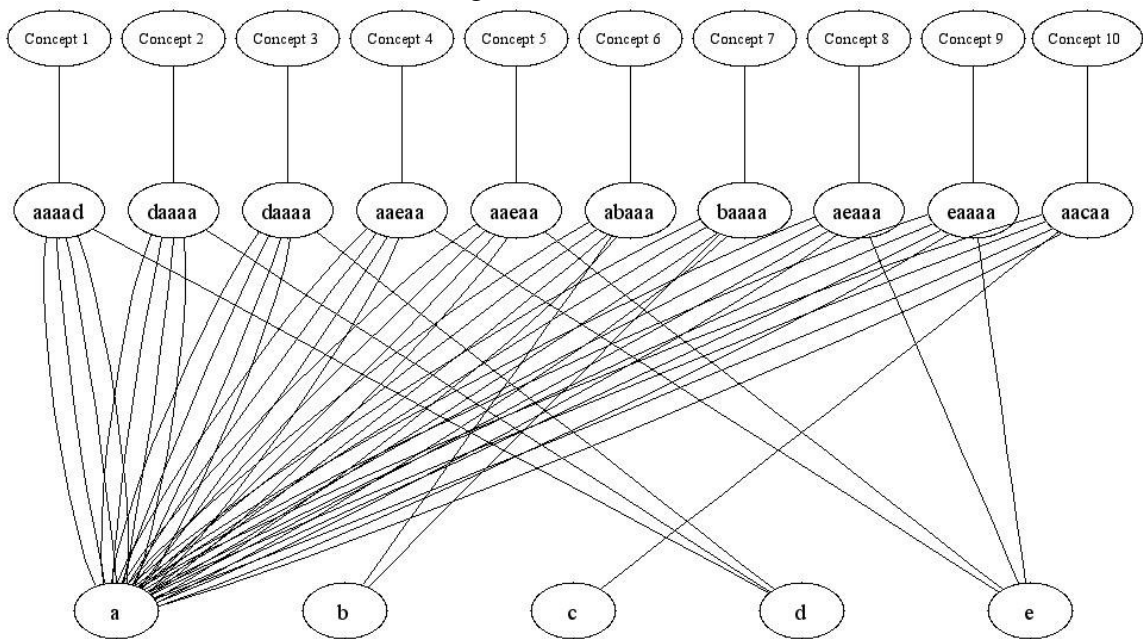
<sup>17</sup> Note that in this simulation, phoneme sequence is irrelevant; each word is essentially treated as an unordered set of phonemes. See chapter 3 for a discussion of the role of sequencing in the model.

each node at each time step is calculated as in (8), with the difference that weights for all nodes are set at 1.0 (resting activations are not modeled).

If the existing word wins, the newcomer is discarded, but if the newcomer wins, it replaces the existing word in the lexicon. The lexicon thus remains the same size, but after the first generation it is populated only by words that have won every competition in previous generations. Any property that gives words an advantage in this competition will thus come to dominate the lexicon.

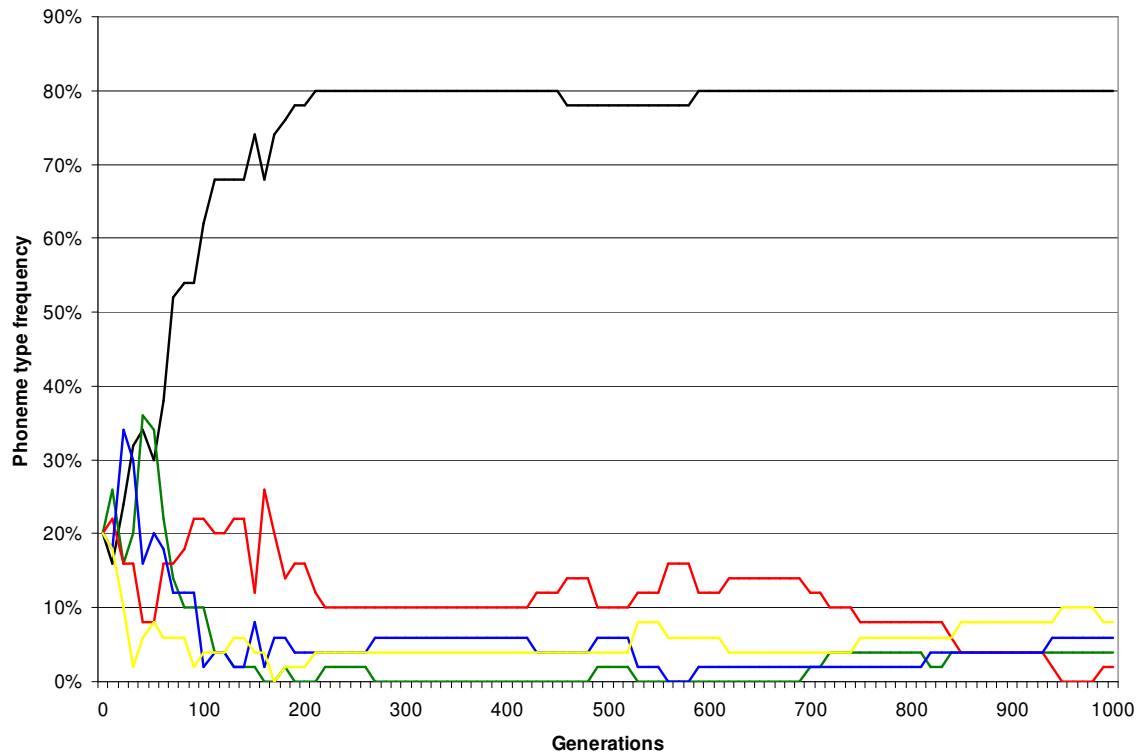
I let the simulation run for 1,000 generations, and allowed activation to spread for five time steps, which is enough time for the type frequency effects discussed by Dell and Gordon (2003) to affect the lexical nodes. Words with phonemes that are themselves contained in many words thus have an advantage. If any phoneme becomes sufficiently more frequent than the others (a virtual inevitability given the stochastic nature of the simulation), words containing that phoneme have a higher probability of entering the lexicon, which in turn increases the phoneme's frequency, and so on, in a feedback loop that magnifies the inequality over time. This can be seen in the final state of the network for one sample run, shown in (16), in which *a* represents 80% of all phonemes.

(16) Network final state (after 1,000 generations)



The evolution of the frequency distribution over time is depicted in (17), with each line representing a different phoneme.

(17) Phoneme type frequencies over time



The uniform distribution that the simulation starts out with is thus unstable, and tends to evolve into a highly skewed but stable distribution. Barabási and Albert (1999), following work by Simon (1955), have shown that this is a general property of networks: the distribution of connections in any graph will tend to become skewed towards a small number of highly connected nodes<sup>18</sup> as long as two conditions are met: (1) new nodes are continually added, and (2) connections from new nodes to existing nodes exhibit *preferential attachment*—that is, new nodes prefer to attach to nodes that already have many connections. Both of these conditions hold in the simulation: several new nodes are introduced each generation, and new words that are

---

<sup>18</sup> More technically: the number of connections per node will be distributed according to a power law.

connected to phoneme nodes that themselves have many connections are more likely to be retained, due to the extra activation received from these nodes.

This model also suggests how statistical patterns could persist across time even as the lexicon changes. If we compare the lexicon in the simulation after 200 generations to the lexicon after 1,000 generations, we find only 2 of the 10 words in common; despite this, the frequency distributions at the two stages are quite similar, with *a* representing roughly 80% of the phonemes at both points. This is reminiscent of the differences between the Old English and Modern English lexicons discussed in §1.1, which exhibit similar phoneme frequency distributions despite sharing less than 20% of their lexicons. In this model this is a natural consequence of the pressure for new words to contain phonemes that are frequent in the existing lexicon.

One way in which the simulation results differ from the distributions seen in natural languages, however, is the degree of skew—the final state in the simulation is heavily dominated by a single phoneme. One reason for this is simply the extreme simplicity of the model—a word’s ability to enter the lexicon is solely a function of its phonotactic properties. The model does not take into account social factors, for example, which might benefit words that contain low-frequency phonemes. Incorporating these factors, which I will assume are largely independent of phonotactics, into the model allows us to fine tune the extent of the skew seen in the stable distribution, as I show below.

I modified the simulation to account for non-phonotactic factors by simply assigning the lexical nodes random weights. The weight for each word represents that

word's "goodness" according to all factors other than phonotactics. A word's weight is determined by choosing a random number from a Gaussian distribution with mean 1 and standard deviation  $\delta$  (the minimum weight is 0; the weight of a given word does not change over the course of the simulation).<sup>19</sup> Each word's weight is determined once, and remains the same for that word for the duration of the simulation; the weights of new words are drawn from the same distribution as existing words, so that new words have neither an advantage or disadvantage over words already in the lexicon. Words with higher weights will have a correspondingly higher probability of being selected, independent of what phonemes they contain.

The charts in (18) show the resulting frequency distributions after 1,000 generations for different values of  $\delta$ . As a way of quantifying the degree of skewness, the Shannon entropy is also given for each distribution.<sup>20</sup> The entropy can be thought of as a measure of how close a probability distribution is to uniform; entropy is maximized when the distribution is uniform, and gets smaller as the distribution becomes skewed towards a subset of its values.

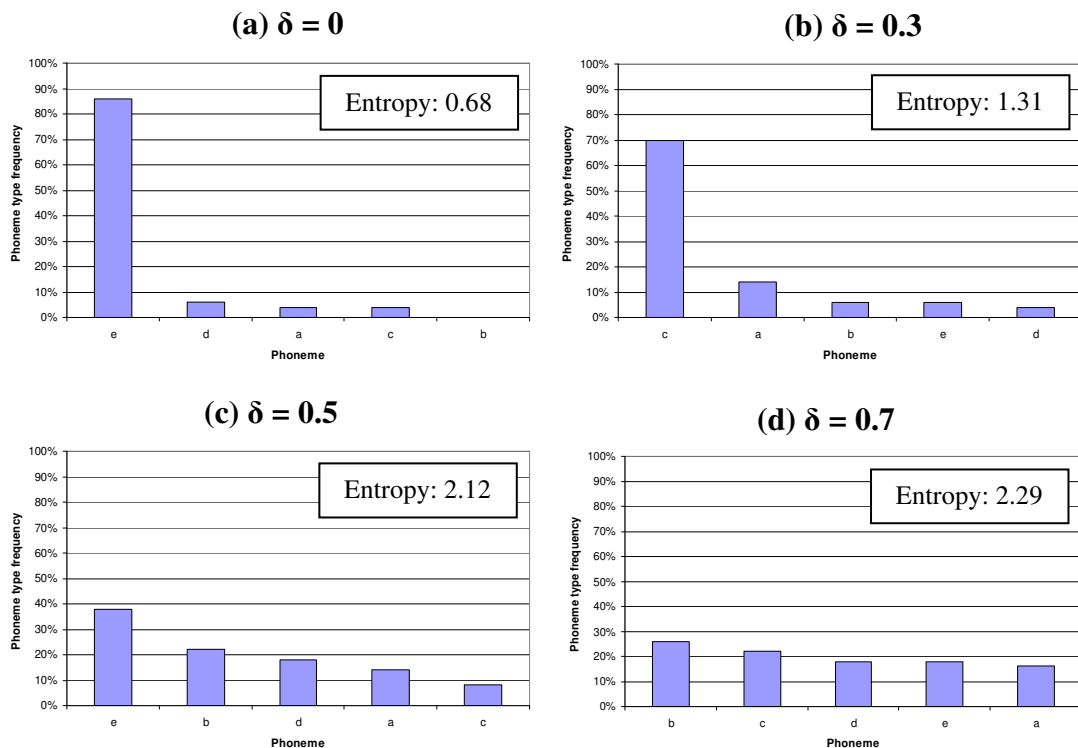
---

<sup>19</sup> Only weights for lexical nodes are given random weights. All other nodes (those for concepts and phonemes) have weights of 1.0.

<sup>20</sup> The Shannon entropy  $H$  of a probability distribution  $X$  over the values  $\{x_1 \dots x_n\}$  is calculated as

follows:  $H(X) = -\sum_{i=1}^n p(x_i) \log_2 p(x_i)$ . It is measured in bits.

(18) Final phoneme type frequencies

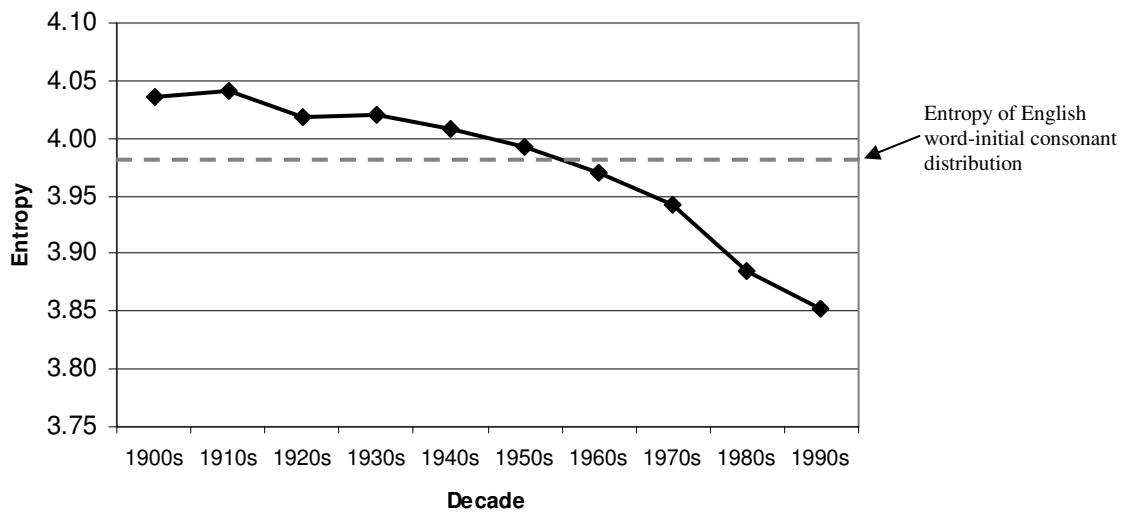


The chart in (18a) represents the case where  $\delta = 0$ , and thus  $w$  for all words is 1.0. This is identical to the earlier version of the model without lexical node weights, and as before results in a highly skewed, low-entropy distribution. As the standard deviation for weights increases, the final distribution approaches uniformity, nearing the maximum possible entropy of 2.32 bits. In other words, as the standard deviation and thus potential size of  $w$  increases, the importance of non-phonotactic factors increases with respect to phonotactic factors, and the distribution begins to resemble the flat distribution we would expect if the phonemes played no role at all in a word's success. It is likely that phoneme frequencies in natural languages occupy a middle ground between the extremely skewed distribution in (18a) and the near-uniform

distribution in (18d), suggesting that both phonotactic and non-phonotactic factors play non-negligible roles in synonym competition.<sup>21</sup>

An illustration of this interplay between phonotactics and other factors can be seen in the popularity of American given names over time. Using data on the most common names for Americans born in each decade of the twentieth century,<sup>22</sup> we can observe how the frequency distribution of phonemes in names changes over time. As shown in (19), the frequency distribution over the initial consonants in these names underwent a gradual decrease in entropy, corresponding to an increase in skew, over the course of the century.<sup>23</sup>

(19) Decreasing entropy in distribution of name-initial consonants



<sup>21</sup> Of course, other phonotactic factors, such as constraints on sequencing, surely play a role in shaping phoneme frequencies. If identical phonemes are prevented from occurring next to each other, for example, no single phoneme will be able to achieve a frequency as high as 80%, which will contribute to de-skewing the final distribution.

<sup>22</sup> The data comes from the United States Social Security Administration, and consists of the 1,000 most popular names each for boys and girls in each decade. It is available at <http://www.ssa.gov/OACT/babynames/>.

<sup>23</sup> Because the data consists only of orthographic representations, I used the Carnegie Mellon Pronouncing Dictionary (Weide 1998) to supply phonetic transcriptions for most names (roughly 90%). For those names not listed in the dictionary, I decided what the most likely initial sound was.



The increase in skew seen here corresponds in my model to a decrease in  $\delta$ , which itself represents a decreased role played by non-phonotactic factors.

Comparison with the horizontal dashed line, which indicates the entropy of the distribution of initial consonants in the entire English lexicon,<sup>24</sup> shows that the distribution of name-initial consonants began the century less skewed than English, and ended up more skewed, with more of the probability mass concentrated in fewer consonants. This suggests that American culture underwent a shift in naming practices during the twentieth century in which influences that might have competed with phonotactic forces became less important. It is plausible, for example, that in the early part of the century tradition played a greater role in naming—children were often named after relatives or famous figures, or given biblical names. Today, parents are more free to simply invent a completely novel name. Pharr (1993), for example, found that the percentage of African-American high school students with “coined or freely invented” names increased steadily from 1.3% for those born in the 1920s to 46.6% for those born in the early 1980s (401). The result of these gradually loosening cultural restrictions is a greater role for phonotactics—choosing a name based almost entirely on its sound.

This is somewhat counterintuitive—as cultural restrictions on naming were relaxed, to the point that today many complain that “anything goes” when it comes to naming children, the phonotactic properties of names became *more* regular and predictable. This makes sense, however, if the process of name (or word) selection is

---

<sup>24</sup> The distribution described here is based on the initial consonants in all non-prefixed words found in CELEX (Baayen et al. 1993).

seen as a competition between rival forces; weakening one such force can strengthen another.

A similar effect can be seen in Grant Smith's (1996, 1998) work on the influence of the phonotactics of names on the electability of candidates for political office. Smith found that the phonotactics of a candidate's last name can to a certain extent predict his or her success in an election,<sup>25</sup> but that phonotactics is a better predictor when voters know little about the candidates' positions on issues (Smith 2007). As in the *shivaree* case discussed earlier, phonotactic preferences make themselves felt most strongly when other factors are absent or weakened.

### **1.7. Summary of the model**

We are now in a position to make explicit what it means for two lexical items to compete. When multiple words are completely synonymous, all of them will always be simultaneously activated by the same concept. In general, a word with a phonotactic advantage, such as having phonemes with high resting activations or phonemes that belong to many other words, will tend to win the race to be selected more often due to the feedback it receives from the phoneme level, and consequently end up being used more than the other synonyms. This greater usage will have the effect of raising the resting activation of the lexical entry throughout the speech community. As shown in §1.4, over time the community will converge on a single synonym, while usage of the other synonyms drops to a point at which new learners

---

<sup>25</sup> Examples of phonotactic properties that correlate with success include having two syllables with trochaic stress, ending in a nasal, and beginning with a liquid (Smith 2007).

of the language are unlikely to acquire them. The ultimate result of many such competitions will be a lexicon biased towards those phonemes (or other phonological units that are encoded by the system) that confer a processing advantage.

My goal in this chapter has been to show that current models of speech production are compatible with the existence of phonotactic preferences, which we can now understand as biases in the speech production network towards the selection of lexical entries that contain certain phonemes over others. This is not to suggest that this is the only way that phonotactic preferences could be manifested. For example, word learning could also be biased by phonotactics, such that words containing certain sounds are easier to learn or recall (Gathercole et al 1999, Storkel and Rogers 2000). My point is merely to show that it would be at least unsurprising, given what we know about language production, for lexical selection to be biased by phonotactics. It remains to be shown that such biases in fact exist. The next section discusses how the theory can be tested empirically.

## **1.8. Empirical consequences of the model**

The model I have outlined here makes a number of predictions regarding how sounds can influence the lexical selection process. On the assumption that repeated activation of a node raises its resting activation (which appears to be the case for lexical nodes), phonemes with a high token frequency should give words a selection advantage. In addition, phonemes with a high type frequency (i.e., those that occur in

many words) should also confer an advantage, for the reasons given by Dell and Gordon (2003).

Other biases will result depending on the content of the phoneme nodes. If, for example, entire syllables have their own nodes, as Levelt and Wheeldon (1994) argue, then words with high-frequency syllables, and not just high-frequency sounds, should be preferred. Dell's (1986) model of speech production incorporates units encoding syllables, rimes, clusters of phonemes, and features, meaning that all of these elements can be accompanied by frequency effects.

The intrinsic properties of the sounds encoded by the phoneme nodes could also play a role. I will argue in chapter 2 that phonemes corresponding to sounds that are easy to articulate have higher resting activations regardless of their frequency. Another possible factor is the relative difficulty of sequencing successive phonemes due to their content. Extensive psycholinguistic evidence suggests that sequences of highly similar phonemes are difficult to sequence (see Frisch 2004 for a summary)—in chapter 3 I show that these kinds of sequences are indeed avoided by English speakers.

The case of *shivaree* versus *callathump* described in this chapter is merely a single (albeit colorful) anecdote, and by itself is evidence of nothing. In the case of a single example of lexical competition it is perhaps impossible to determine the exact contributions of phonotactic and sociolinguistic factors. My strategy will therefore be to look at the patterns that emerge in a large number of such cases. If, for example, we compare all of the neologisms that enter a language over a period of time—all

presumably winners of their respective competitions—to the words that fall out of use over the same time period—all losers—a consistent, significant phonotactic difference between the two groups is evidence of selection pressure.

The remainder of the dissertation is devoted to case studies of this type. In chapter 2 I present evidence confirming Boersma’s (1998) proposal that a sound change resulted in an unnatural statistical pattern in the lexicon of Latin, and that phonotactic preferences operated over time to “repair” this gap in Latin’s modern descendent languages. From this case I conclude that phonotactic biases have a component which is derived from the physical facts of articulation, and is thus not solely an effect of frequency.

In chapter 3 I turn to sequences of sounds, and show that sequences of identical liquids in English are statistically underrepresented. I present evidence from American baby names and neologisms that illustrate how a lexical bias may be maintained even as the contents of the lexicon are replaced. Finally, in chapter 4 I use evidence from English, Turkish, and Navajo to argue that learning biases can also create phonotactic preferences, and discuss the role of the grammar in the processing model outlined above.

## 2. Markedness and lexical change

### 2.1. Introduction

In the previous chapter I argued for a model of speech production in which synonyms expressing a single concept compete with each other to be used to express that concept. As I described it, the network that instantiates the speech production system does not contain any *a priori* biases in favor of any type of phonological node—in fact, the actual content of the nodes is irrelevant. Those nodes that are connected to many words, or to high-frequency words, have more influence on lexical selection, regardless of what sounds they happen to represent. This means that wholly arbitrary phonotactic patterns may become entrenched in the lexicon and persist despite changes in the vocabulary. In this chapter I will present evidence that this model is inadequate, and that some phonotactic preferences are triggered by universal biases. A word containing a rare but unmarked sound may be preferred over a word containing a frequent but marked sound.<sup>26</sup>

The evidence I will present to support this view comes from Boersma's (1998) observation of the historical consequences of a sound change in Proto-Indo-European, the ancestor language of all Indo-European languages. By tracing the development of several Romance languages following this change, I will show that Boersma's proposed analysis of the historical facts is essentially correct: certain sounds can be

---

<sup>26</sup> The term *markedness* is problematic, as Haspelmath (2006) has pointed out, as it has been used to describe many different things: a sound's formal representation, phonetic properties, or frequency, to take just a few examples. In discussing the case study described in this chapter, I will use the word as shorthand for ease of articulation—marked sounds are more effortful than unmarked sounds. It is certainly possible that phonotactic preferences are also driven by other forms of markedness, such as ease of perception, but establishing this would require evidence beyond the single case study in this chapter.

preferred despite their low frequency, and these frequency-insensitive preferences are responsible for some typological tendencies as well as the ability of languages to “repair” accidental gaps in their lexicons.

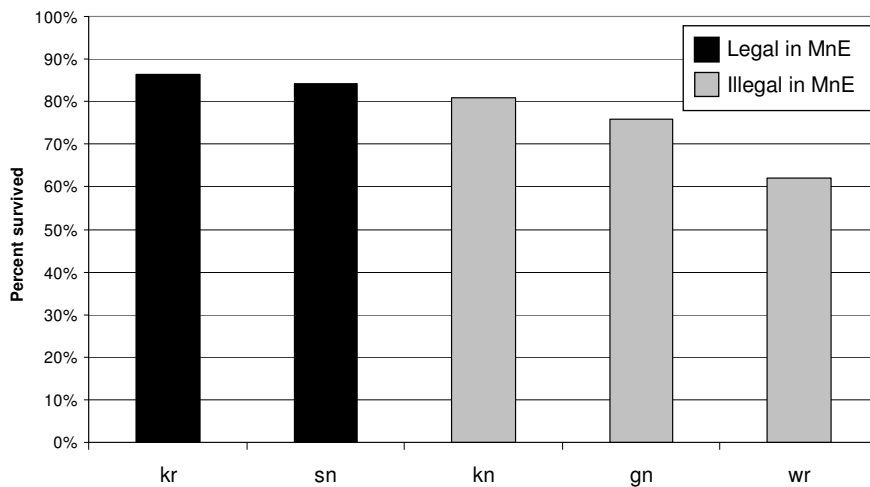
## 2.2. Word-initial clusters in English (Berg 1998)

Berg (1998) gives an example of phonotactic preferences that appear to be driven by markedness from the history of English. The consonant clusters /kn/, /gn/, and /wr/ were legal as word onsets in Old and Middle English, but are no longer legal in Modern English. The clusters were each simplified by deleting the first consonant, as can be seen by the pronunciations of modern words like *knight*, *gnaw*, and *wren*. Berg uses this simplification as evidence that these clusters are marked when compared to unsimplified clusters like /kr/ and /sn/. He then looks for evidence that fewer words with the problematic clusters have survived into the modern language than words with less marked clusters.<sup>27</sup> His results, taken from the *Oxford English Dictionary*, are reproduced here in (20). The height of each bar indicates what percentage of words beginning with each cluster have survived into Modern English.

---

<sup>27</sup> As Berg points out, there is not necessarily a single markedness principle that is responsible for the undesirability of all three clusters. The point of his analysis is not to pinpoint the precise factors that cause these clusters in particular to be bad, but simply to show that whatever these factors are, they are reflected both in historical sound change (i.e., cluster simplification) and in the survival rates of entire words.

(20) Word retention by cluster type<sup>28</sup>



Words with illegal clusters are significantly less likely to have survived than words with legal clusters. These results confirm that there is a correlation between a categorical phonological process by which certain clusters are simplified and the fitness of words containing those clusters. Berg attributes this correlation to what I am calling phonotactic preferences: “speakers may resort to a radical means of solving phonological problems: they circumvent these problems by simply not using the words in which they crop up” (233).

For present purposes, what is important about the English onset clusters case is that the relative fitness of the clusters in (20) is not arbitrary, but follows from well-established sonority-based constraints on syllable structure—the tendency for complex onsets to rise in sonority, which militates against /wr/, and the cross-linguistic preference for complex onsets in which the consonants differ maximally in

<sup>28</sup> This chart was constructed from the data in Berg’s (1998) Table 25 (232).



sonority, which makes /kn/ and /gn/ less than ideal<sup>29</sup> (Clements 1990). In the remainder of this chapter, I will examine a similar case from the history of French, in which I will argue not only that phonotactic preferences have acted to shape the lexicon, but that these preferences are rooted in differences in articulatory difficulty.

### **2.3. Voiced stops and place of articulation**

Voicing is difficult to maintain during a stop closure owing to the increase in pressure in the oral cavity, which reduces the pressure differential across the glottis needed for voicing (Westbury and Keating 1986, Ohala 1997). Ohala and Riordan (1979) demonstrate that, although this is true in general of voiced stops, place of articulation also plays a role in determining how long voicing can be maintained. Voicing can be maintained longer in stops in which the closure is farther forward—labial stops, for example, are easier to voice than coronal stops, which are easier to voice than dorsal stops. As Ohala and Riordan point out, this ease of articulation hierarchy correlates with the crosslinguistic distribution of phoneme inventories. There are many more languages with /b/ but not /d/ than languages with /d/ but not /b/ (Sherman 1975, Maddieson 1984).<sup>30</sup>

The place hierarchy among voiced stops also correlates with lexical type frequencies within a single language—/b/-initial words outnumber /d/-initial words in the lexicons of many languages. The two charts in (21) show that this is the case in

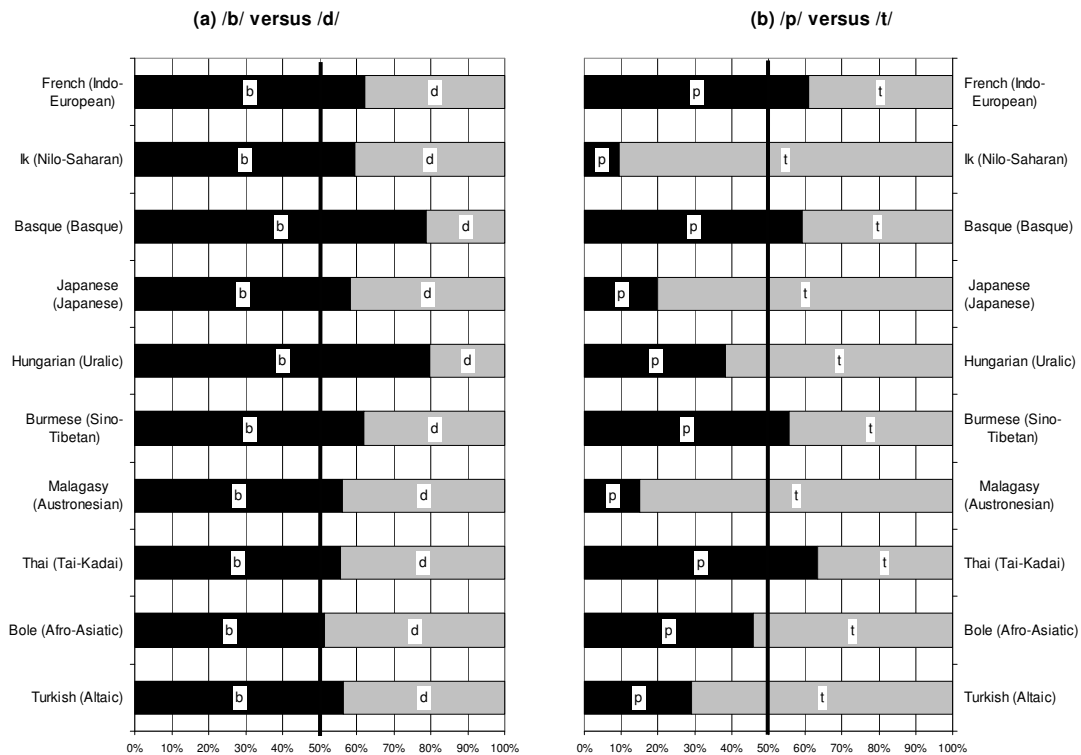
---

<sup>29</sup> The legal cluster /sn/ would seem to be worse than /kn/ along the dimension of sonority distance, but /s/-initial clusters are well-known for their ability to violate otherwise robust sonority-based generalizations (Selkirk 1982).

<sup>30</sup> I will deal only with the asymmetry between /b/ and /d/, ignoring /g/.

several unrelated languages. Each bar in the chart on the left, (21a), represents a ratio between the numbers of words that begin with /b/ or /d/ in each language (the darker vertical line indicates 50%). Data on /p/- and /t/-initial words are given in (21b) for comparison.

(21) Lexical counts of word-initial stops in several languages<sup>31,32</sup>



The figures in (21a) suggest that there is a crosslinguistic preference for /b/ over /d/ as a word-initial consonant. The /p/-/t/ ratios in (21b) demonstrate that the bias towards /b/-initial words is independent of any labial-coronal bias in the voiceless stops; in other words, regardless of whether /p/ outnumbers /t/ or /t/ outnumbers /p/ in a given language, /b/ always outnumbers /d/. This is consistent with the articulatory facts discussed earlier.

<sup>31</sup> Data sources: French: TLF (Imbs 1994); Ik: Heine 1999; Basque: Aulestia 1989; Japanese: Breen 2000; Hungarian: Magay and Országh 1990; Burmese: Stevenson and Eveleth 1953; Malagasy: Richardson 1967; Thai: McFarland 1944; Bole: wordlist compiled by Russell Schuh (p.c.); Turkish: TELL (Inkelas et al. 2000).

<sup>32</sup> Languages were included in this sample according to the following criteria: each language must allow /b/, /d/, /p/, and /t/ word-initially, must not have prefixes that begin with /b/, /d/, /p/, or /t/ (all prefixed words have been omitted from the French data), must have voicing during the closure in voiced stops (i.e., voiced stops must typically have a negative VOT), and must not be in the same family as another language in the sample. The languages listed in (21) represent all of the languages that fit these criteria and for which I have so far been able to obtain information on the lexical statistics.

How and why do languages maintain this lexical bias? In chapter 1, I mentioned two possible theories. One is that individual sounds change over time, and this change is biased towards creating sounds that are easy to produce (Zipf 1935). If, for example, initial /d/ changes to /t/ more often than initial /b/ changes to /p/, over time the result would be a lexicon skewed towards /b/. This could happen, as suggested by Ohala (1981) and Blevins (2004), through misperception on the part of learners, abetted by the smaller amount of voicing produced by speakers during /d/.

Another possibility is that phonotactic preferences are responsible: speakers of a language prefer to retain (or borrow, or coin) words that start with /b/ over words that start with /d/. Over time, this gradient preference leads to a statistical bias towards /b/ even from an initial state in which the sounds were equally frequent. I will argue that there is such a preference and that it can produce the frequency asymmetries seen in (21a). Furthermore, I will show that this preference must be innate, in that it cannot be derived from the properties of the current lexicon.

In this chapter I focus on one of the languages depicted in (21)—French. Boersma (1998) points out that Proto-Indo-European (PIE), from which French is ultimately descended, had very few if any words containing the labial voiced stop *\*b* (Grimm 1819/1837, Pederson 1951, Matasović 1994). The reasons for this gap are controversial, but some have argued that what have traditionally been reconstructed as voiced stops in PIE were originally glottalized stops (Gamkrelidze and Ivanov 1972, Hopper 1973)—this would explain the rarity of *\*b* (actually *\*p'* on this theory),

since languages with glottalized stops are often missing the labial member of the series. A sound change converted these sounds to voiced stops in many daughter languages, including Latin; these languages thus inherited a voiced stop series, but also a lexicon that reflected the earlier state of the system.

Boersma (1998) uses dictionary counts to show that the lexical gap inherited from PIE has been repaired in modern French. He argues that one way this could have happened is that “French borrowed /b/ to a larger extent than the other voiced plosives; this active de-skewing would presumably involve phonologically-determined choices between synonyms in the lexicon” (Boersma 1998: 382); in other words, through the action of what I am calling phonotactic preferences. As he admits, dictionary counts alone cannot establish the truth of this claim; an increase in the frequency of /b/-initial words could have come about through sound change as well. The remainder of this chapter represents an attempt to test the validity of Boersma’s hypothesis and further explore its consequences—is there evidence that phonotactic preferences have shaped the French lexicon over time, and if so, how are these preferences manifested?

## **2.4. The evolution of the French lexicon**

The diachronic analysis of French will proceed in three stages. In §2.4.1, I look at classical Latin, a language intermediate between PIE and modern French, and show that Latin speakers borrowed /b/-initial words heavily despite the rarity of the sound in native words. Next, in §2.4.2, I turn to how Latin evolved into French, and

the roles played by sound change and phonotactic preference. Finally, in §2.4.3, I examine other sources of French vocabulary such as borrowing and morphological derivation.

#### 2.4.1. *From Proto-Indo-European to Latin*

Glancing at any Latin dictionary reveals that Latin, unlike Proto-Indo-European, had a fair number of /b/-initial words. Where did all the instances of /b/ in Latin come from? There were two main sources. Some were the result of a sound change in early Latin in which PIE \*dw became /b/ (e.g., early Latin *duellum* > classical *bellum* ‘war’). Most of the remaining /b/-initial words in Latin were borrowed from other languages, Greek being the largest donor. Looking at the numbers of these words, however, reveals an asymmetry. In Ernout and Meillet (1959), an etymological dictionary of Latin, 35 of the /b/-initial words are listed as having been borrowed from Greek (e.g., Gr. βάρβαρος > L. *barbarus* ‘barbarous, foreign’), compared to only 15 /d/-initial words (e.g., Gr. δελφίς > L. *delphinus* ‘dolphin’). If these numbers are representative, it suggests that Latin speakers preferred to borrow words with /b/ over those with /d/.

This bias in favor of borrowing /b/-initial words is surprising for two reasons. First, Latin had fewer /b/-initial words than /d/-initial words overall (see §3.2 below for details). We might thus expect that speakers of Latin would, if anything, tend to avoid borrowing words containing the rarer sound. Second, Greek also had fewer instances of /b/ than /d/ (due, as with Latin, to the gap in the PIE stop system). One large Greek dictionary (Liddell et al. 1940) lists 3,090 /b/-initial words and 8,860 /d/-

initial words. Assuming that this ratio is representative of the Greek words that a typical Latin speaker would have been exposed to, and thus would have been available for borrowing, this means that Latin speakers heard more /d/-initial than /b/-initial Greek words, yet chose to borrow more /b/-initial words.

If neither the lexical statistics of Latin nor those of Greek can explain the influx of /b/-initial words into Latin, what can? One possibility is the articulatory facts discussed in §2. If people are biased in favor of accepting words that contain sounds that are easier to articulate, then the place asymmetry in the ease with which stops are voiced can be used to explain the borrowing asymmetry in Latin. This preference must be innate, since it contravenes the frequency asymmetry (i.e., /d/ is more common than /b/) in both the donor and recipient languages.<sup>33</sup>

Note that this theory says nothing about sound change. Over time, a lexicon may come to be skewed in this way without any change in individual sounds. Of course, sound change may also contribute to the frequency of a given sound, as in the *\*dw > b* change in early Latin, but it is not the only relevant process. This supports Boersma's (1998) proposal that correlations between frequency and phonetic ease are driven by a combination of change in individual sounds and the preferential selection of entire words by the members of a speech community. The next section presents further evidence for this claim from the evolution of Latin into French.

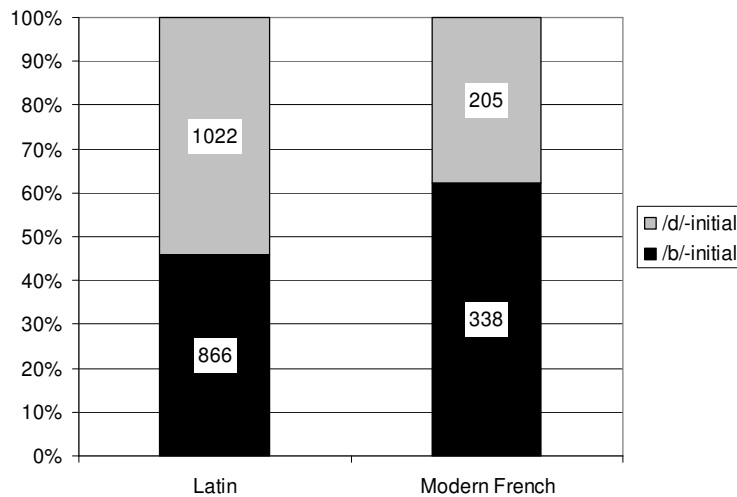
---

<sup>33</sup> The possible senses in which this preference could be innate are discussed in §2.5.

### 2.4.2. From Latin to French

We have seen that by the time of classical Latin, the gap inherited from Proto-Indo-European had been partially filled, but despite this, coronal voiced stops still outnumbered labial voiced stops. As noted above, this bias was reversed in favor of the labials by the time Latin had evolved into modern French. The two languages are compared in (22) below (labels indicate the number of words in each category).

(22) /b/- and /d/-initial words in Latin and French<sup>34,35</sup>



<sup>34</sup> French data source: *Le Trésor de la Langue Française* (Imbs 1994). The data represents every tenth word from the *b* and *d* sections of the dictionary (not including prefixed words, as explained in footnote 35). Latin data source: Lewis and Short 1879. The data represents all words from the *b* and *d* sections of the dictionary, not including proper names or prefixed words.

<sup>35</sup> Boersma (1998) points out that the presence of the productive prefixes *dē-* and *dis-* in Latin and *dé-* and *dés-* in French skews the distribution towards /d/ in both languages. It seems likely that morphological productivity may interfere with phonotactics (e.g., a highly productive /d/-initial prefix may override a phonotactic preference that disfavors /d/), so in what follows I follow him in considering only non-prefixed words in both languages. For the Latin data, I simply removed from the data all words that began with the strings *dē-* and *dis-*, as well as *dī-* and *dif-* (allomorphs of *dis-*). This method removes any words that begin with these strings, even those that are not prefixed. This is the most conservative way of eliminating prefixed forms, since a more accurate method would increase the count of /d/-initial words, and thus increase the magnitude of the asymmetry I discuss here. For the French data, I rely on the morphological parse supplied in the entry for each word in the dictionary (Imbs 1994); each word that was indicated as having a prefix was removed.



How did this happen? One way to answer this question involves examining the Latin vocabulary to determine which words survived into the modern language.

Each Latin word could have had one of three fates, summarized in (23):

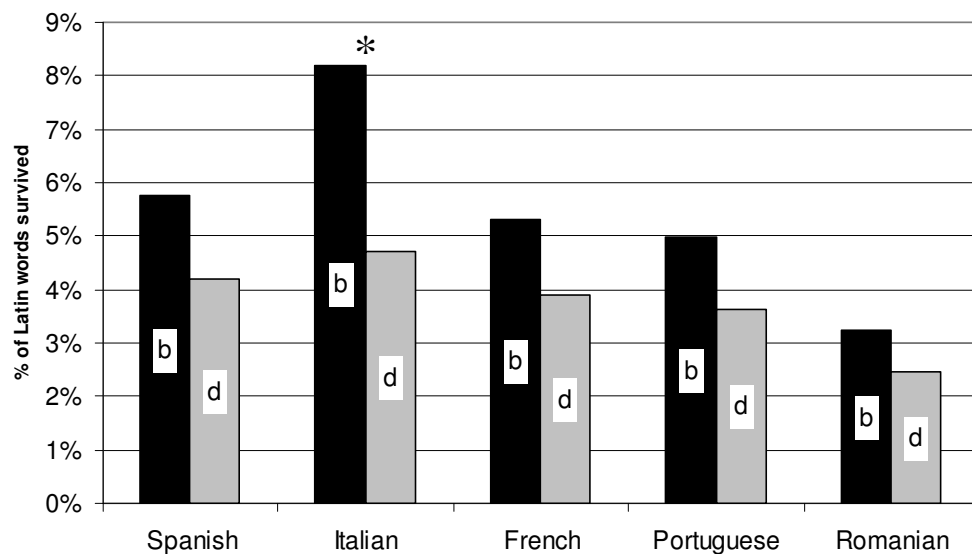
(23) Possible fates of Latin words

- |   |       |  |
|---|-------|--|
| (a) survived into French with initial sound changed   | e.g., | L. <i>diurnum</i> > Fr. <i>jour</i> ‘day’  |
| (b) survived into French with initial sound unchanged | e.g., | L. <i>basiare</i> > Fr. <i>baiser</i> ‘kiss’<br>L. <i>dubitare</i> > Fr. <i>douter</i> ‘doubt’ |
| (c) failed to survive                                 | e.g., | L. <i>balanus</i> ‘acorn’<br>L. <i>docere</i> ‘teach’  |

Of the 1,888 /b/- and /d/-initial words listed in Lewis and Short 1879, only two fall into category (a): L. *diurnum* > Fr. *jour* ‘day’, and L. *deorsum* ‘below’ > Fr. *jusant* ‘ebb tide,’ both cases of /d/ palatalizing before a front vowel. Sound change is thus responsible for only a tiny fraction of the statistical shift from Latin to French.

A larger role is played by the differing rates at which the two types of word survived. Although surprisingly few Latin words left direct descendents in French, /b/-initial words had a greater chance of surviving (5.3%) than /d/-initial words (3.9%). This was true not only for French, but for several Romance languages, as shown in (24).

(24) /b/- and /d/-initial survival rates in modern Romance languages<sup>36</sup>



It is likely that the similar behavior exhibited by these languages is partly due to the fact that they are closely related. Many of the words that failed to survive into Spanish and Portuguese, for example, may have died before West Iberian Romance, their common ancestor, split into its daughter languages. Indeed, there is a high degree of overlap between the sets of words that survived into these two languages: of the 123 words that survived into either Spanish or Portuguese, 93 survived into both.

<sup>36</sup> Etymological data for all Romance languages comes from Meyer-Lübke 1935. An asterisk indicates that the difference between survival rates of /b/- and /d/-initial words in that language is significant by Fisher's Exact Test (two-tailed  $p < .05$ ).

One might thus object that the similarities across languages depicted in (24) are all simply reflections of something that happened early in the development of Romance. But the degree of overlap is much smaller for other language pairs: 74 common words out of 157 for French and Spanish, and only 34 out of 144 for French and Romanian, for example. In other words, different languages retained largely different sets of words; what was common to all was the tendency to retain /b/-initial words more than /d/-initial words.

Another possible explanation for the biases in (24) involves token frequencies. If for some accidental reason /b/-initial words tended to be more frequent than /d/-initial words, and more frequent words had a higher chance of surviving (as is surely true), this would account for the asymmetry without making reference to the properties of the sounds themselves. This, however, is not the case. The average frequency of /b/-initial words in Latin was not significantly different than that of /d/-initial words ( $p > .5$  by Wilcoxon rank sum test).<sup>37</sup> Furthermore, the median log frequency in Latin of those /b/-initial words that survived into French (0.79) was significantly lower than the median log frequency of /d/-initial survivors (2.47) ( $p < .01$  by Wilcoxon rank sum test). It appears that /b/-initial words were preferentially retained despite, and not because of, their frequency.

Latin words beginning with /b/ were both fewer in number and less frequent in running text than /d/. Despite this, /b/-initial words tended to outlive /d/-initial

---

<sup>37</sup> The frequencies cited in this section are taken from a 3.4 million token corpus of classical texts stored at the Perseus Digital Library (<http://www.perseus.tufts.edu/>). Each figure represents the log of the raw number of instances of each word in the corpus. To avoid taking the log of zero, 0.5 was added to each number.

words; some property of the labial stop allows words containing it a better chance at remaining in use. This therefore constitutes another piece of evidence for a phonotactic preference motivated by articulatory ease. In the next section I examine how this preference could have led to biases in the rest of the modern French vocabulary.

#### 2.4.3. *Sources of French vocabulary*

In the previous section I examined how the initial segment of a Latin word influenced its ability to survive into Latin's modern descendents—to do this, I examined individual Latin words and followed their progress forward in time, as Latin evolved into French and the other Romance languages. In this section I do the reverse—look at a sample of the Modern French vocabulary and go backwards in time to determine how each word entered the French lexicon. Doing so sheds light on other ways besides survival from Latin that the bias for /b/ over /d/ has shaped the vocabulary of French.

Although French is a Romance language, most of the vocabulary of the modern language comes from other sources besides direct descent from Latin. Several of these sources are described in (25).<sup>38</sup>

---

<sup>38</sup> Etymological categories not included in (25) include onomatopoeia, acronyms, and words whose etymology is listed as unknown in Imbs 1994. These excluded words account for less than 5% of the data (22 out of 543 words).

(25) Main sources of French vocabulary

Source	Examples
(a) inherited from Latin	<i>L. basiare</i> > Fr. <i>baiser</i> ‘kiss’ <i>L. dubitare</i> > Fr. <i>douter</i> ‘doubt’
(b) re-borrowed from Latin	<i>L. baptisterium</i> > Fr. <i>baptistère</i> ‘baptistry’ <i>L. destinatio</i> > Fr. <i>destination</i> ‘destination’
(c) borrowed from another language	It. <i>bravura</i> > Fr. <i>bravoure</i> ‘bravery’ Trk. <i>dervis</i> > Fr. <i>derviche</i> ‘dervish’
(d) derived from existing French word <sup>39</sup>	<i>bêcher</i> ‘dig’ > <i>bêche</i> ‘spade’ <i>dur</i> ‘hard’ + <i>-eté</i> > <i>dureté</i> ‘hardness’

As with the inherited words discussed in the previous section, sound change played little if any role in affecting the /b/-to-/d/ ratio. In my French dictionary sample, only one word showed a historical change in the initial segment: *L. unde* > Fr. *dont* ‘of which.’

The table in (26) lists how many /b/- and /d/-initial words each of the sources in (25) contributed to the French vocabulary (the numbers listed represent every tenth word from Imbs 1994).

(26) Breakdown of French words by origin (no prefixed forms)

			/b/-initial words	/d/-initial words
(a)	inherited from Latin	<b>B ≈ D</b>	14	15
(b)	re-borrowed from Latin	<b>D &gt; B</b>	31	77
(c)	borrowed from another language	<b>B &gt; D</b>	64	31
(d)	derived from existing French word	<b>B &gt; D</b>	212	77

<sup>39</sup> Recall that the data excludes prefixed words, so that the derived words category is made up mostly of compounds, suffixed words, and deverbal nouns.

The fact that roughly the same number of /b/- and /d/-initial words were inherited from Latin (category (a)) is consistent with what we saw in the previous section—there were more /d/- than /b/-initial words in Latin, but /b/-initial words were more likely to survive, meaning that the absolute number of both word types surviving into French is about the same.

Although few words in French were directly inherited from Latin, many words were re-borrowed from Latin after the two languages had become distinct (category (b)). Most of these were used as technical religious or scientific vocabulary. As the table in (26) shows, many more /d/-initial words were borrowed in this way than /b/-initial words, by a factor of about 2.5 to 1. This might appear to be evidence that a phonotactic preference for /b/ was not active in these cases, but it must be remembered that /d/-initial words outnumbered /b/-initial words by 3 to 1 in Latin (if prefixed forms are included). These numbers are thus consistent with a slight preference for /b/. Without a larger sample of words from French, however, we cannot determine whether this difference is significant.

Likewise, although the words borrowed from other languages (category (c)) appear to be biased in favor of /b/, we cannot conclude anything from this without knowing the lexical statistics of the donor languages. This bias may simply be the result of the fact that French speakers have been exposed to more /b/-initial than /d/-initial words, and so had more chances to borrow the former.

The largest difference between /b/ and /d/ is to be found in the words derived via French morphological processes (category (d)). Words starting with /b/ are

derived in French over two and a half times more often than words starting with /d/. In non-derived words, on the other hand, /d/-initial words are slightly more common (123 words) than /b/-initial words (109 words). Thus, it appears that morphological processes themselves are biased by a preference for /b/. The large size of this difference may be the result of a feedback loop due to the recursive nature of morphological operations—a slight preference for forming /b/-initial words will lead to more /b/-initial words, which will thereby increase the number of /b/-initial bases for further derivation, and so on.

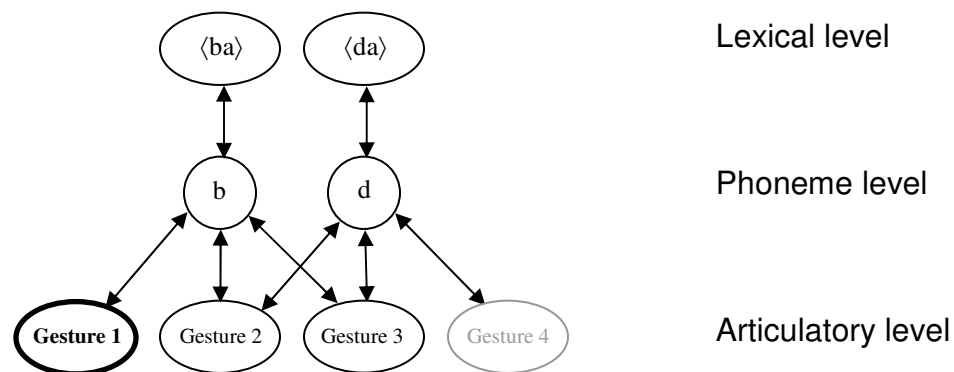
## **2.5. Modifying the model**

In this chapter I have presented evidence that when one sound is easier to articulate than another, words containing the easier sound are given an advantage in the lexical selection process. How can articulatory ease be integrated into the model of speech production I outlined in chapter 1? The simplest way would be to extend the production network to include not only nodes that correspond to phonological elements such as phonemes and features, but also nodes corresponding to articulatory elements such as individual gestures or gestural scores.

I will further assume that at the articulatory level, a principle of least effort applies, serving to strengthen nodes that represent articulations that require the least amount of energy expenditure (cf. the LAZY constraint in Kirchner 2001). Thanks to this principle, once a person has had a certain amount of experience producing sounds, the nodes responsible for producing a /b/, for example, will overall have higher

resting activations (weights, in my model) than those responsible for /d/. Assuming that articulatory nodes feed activation back to phonemic nodes, such a network would predict that phonemes which are easier to articulate, and therefore words that contain such phonemes, have an advantage over their competitors in lexical selection. The structure of such a network is shown in (27). Nodes with higher weights are represented with darker symbols.

(27) Network with gestural nodes



Some support for this architecture comes from Goldrick (2003), who found in an experiment designed to elicit speech errors that English-speaking subjects made /s/→/t/ errors more often than /t/→/s/ errors, despite the fact that /s/ is more common than /t/ in English. This finding is at odds with extensive evidence that in general speech errors tend to result in more frequent sounds replacing less frequent ones—for example, /g/ → /k/ errors are more frequent than /k/ → /g/ errors, reflecting the greater frequency of /k/ in English (Motley and Baars 1975, Levitt and Healy 1985,



Dell et al. 2000). Goldrick argues that the /t/-/s/ anomaly is due to the fact that fricatives are more difficult to articulate than stops.<sup>40,41</sup>

In my model, this result would be explained as the effects of the stronger (i.e., higher resting activation) articulatory nodes associated with /t/, which makes the phoneme node /t/ more likely to be accidentally selected during speech planning. These markedness effects can be modeled with the simulation described in §1.6. First, let us consider what the simulation without markedness predicts will happen if a new phoneme is added to the inventory. To do this, I started the simulation with a lexicon of ten words and five phonemes and set it to run for 1,000 generations, as described in §1.6.<sup>42</sup> After the first 500 generations, I added a sixth phoneme, *f*, to the inventory, simulating a sound change like that postulated by the Glottalic Theory in PIE. This new phoneme was not instantiated in any existing words, and so was unable to increase in frequency, as the graph in (28) shows. Each gray line represents the type frequency of one of the original phonemes over time; the thick black line represents the newly introduced phoneme.

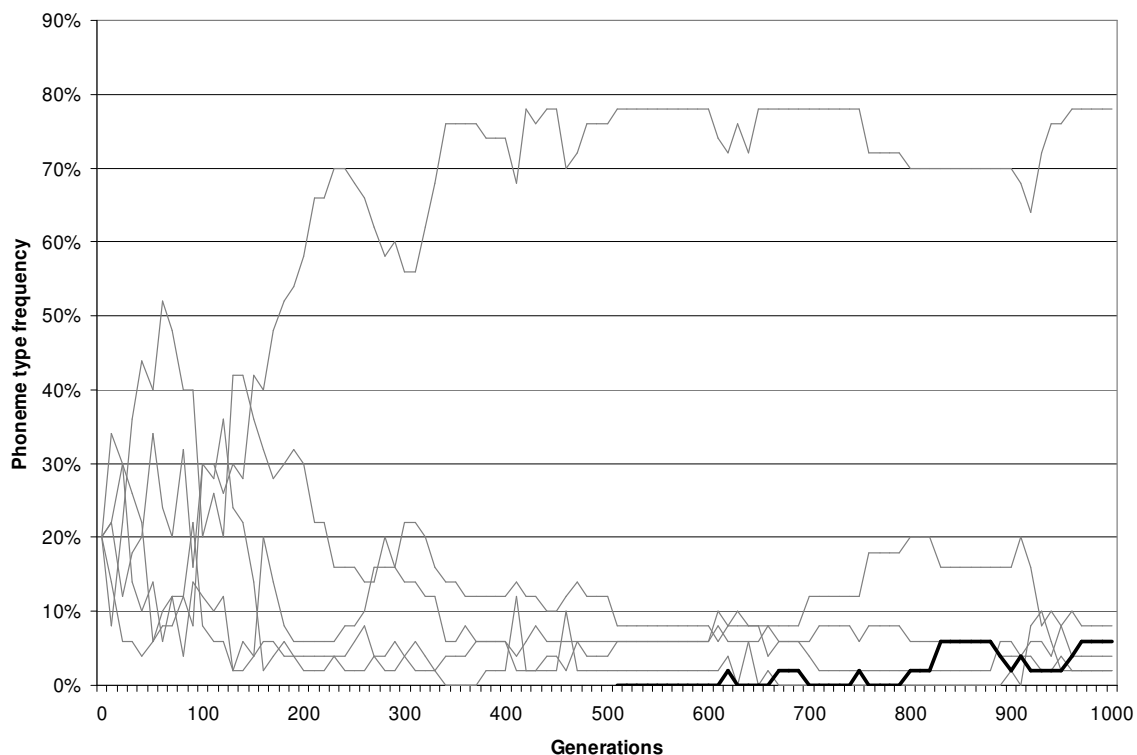
---

<sup>40</sup> Fricatives are generally held to require more effort to produce than stops due to the greater precision of the gesture involved in making a fricative (e.g., Kirchner 2001, 41-42); as Boersma (1997) puts it, “it is easier to run into a wall than to suddenly halt one inch in front of it” (12).

<sup>41</sup> Some studies have failed to find any markedness bias in phoneme errors of this type, but see section 3 of Goldrick 2002 for an extensive methodological criticism of these studies.

<sup>42</sup> The model parameters were set as follows:  $p=0.3$  for all connections,  $q=0.6$ ,  $\delta=0.1$ .

(28) Introduction of new phoneme: no markedness



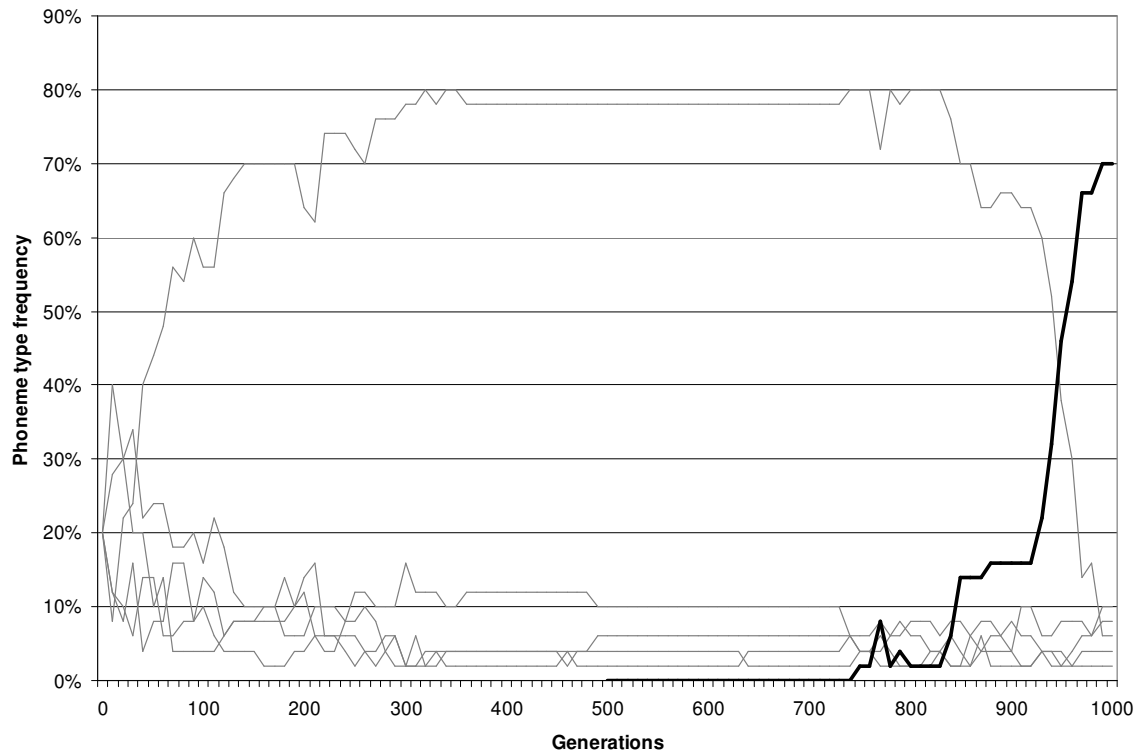
In order to model the effects of markedness, I assigned weights to each phoneme node, just as I did with the lexical nodes in the simulation in §1.6. I ran the simulation again, but this time gave the node representing the new phoneme *f* a weight of 1.5 (all other phoneme nodes had weights of 1.0). Remember that a node's weight is used as a multiplier when the node's activation is calculated, so the higher weight gives words containing the new phoneme an advantage. This represents the node's connections to a set of gestures that are easier to articulate. The equation used to compute activation is repeated in (29).

(29) Calculating activation with weights

$$A(j, t_i) = w_j[A(j, t_{i-1}) + \sum_{k=1}^n p_k A(c_k, t_{i-1})](1 - q) + noise$$

When markedness is implemented in this way, the new phoneme increases in frequency despite its initial rarity, as shown in (30).<sup>43</sup>

(30) Introduction of new phoneme: with markedness



Of course, this does not mean that the frequencies of all sounds are strictly determined by markedness. For two phonemes with roughly equal weights, it will be essentially random which one is more frequent in a given language—this could be the case for /p/ and /t/, as was shown in (21).

<sup>43</sup> To confirm that the difference between the two simulations in (28) and (30) was not a fluke, I ran the simulation 10 times under each set of conditions. Without markedness, the new phoneme was never the most frequent phoneme at any point in any of the 10 simulation runs; with markedness, the new phoneme ended the simulation as the most frequent phoneme all 10 times.

## 2.6. Conclusion

In tracing the history of a subset of the French vocabulary, I have shown that Boermsa's (1998) conjecture is correct: a phonotactic preference for sounds that are easier to articulate has resulted in a statistical bias towards initial /b/ over initial /d/ in the lexicon of modern French. Because for most of the history of the language, /b/ has been less frequent than /d/, this preference cannot simply be derived from the lexical statistics; the hypothesis that speakers simply prefer words that are similar to the words they already know, as is predicted by a markedness-free model of speech production, cannot account for the consistent bias towards /b/ that is observed in the history of Latin and French.

I have also shown that this preference for unmarked sounds can be modeled using a speech production network in which nodes representing unmarked sounds are given higher weights than those representing marked sounds. When phonemic nodes are weighted in this way, the model predicts that mismatches between frequency and markedness, which can arise due to “blind” sound change, will be gradually repaired over time, just as has apparently happened in the history of the Romance languages.

### **3. Phoneme cooccurrence and lexical bias**

In the model of speech production and lexical competition I have developed so far, a word's success depends on its phonotactic properties because its lexical node is connected to nodes representing phonological units such as phonemes or syllables. In this chapter, I will present evidence that long-distance dependencies between segments also play a role in competitions among words—specifically, that words having highly similar consonants in close proximity are at a disadvantage.

The chapter is organized as follows. I begin in §3.1 by establishing that a gradient similarity avoidance constraint is present in many languages, including English. In §3.2 and §3.3 I present evidence from English words and American first names that shed light on the effect this constraint has in the evolution of the lexicon. In §3.4 I discuss the role played by processing in similarity avoidance, and finally in §3.5 I show how the speech production model established in chapter 1 can be modified to account for the data presented here.

#### **3.1. Gradient OCP effects in English**

Frisch et al. (2004) showed that in Arabic trilateral verbal roots, similar consonants tend not to occur in adjacent positions—this restriction is categorical when the consonants are identical (e.g., there are no roots of the form /d d m/), but gradiently holds when the consonants are merely similar (e.g., roots like /d s m/, whose first two consonants are both coronal, are attested, but statistically underrepresented). The same similarity avoidance tendency has been found in the

lexicons of many languages: Ngbaka (Broe 1995), Javanese (Mester 1986), Russian (Padgett 1995), Muna (Coetzee and Pater 2005), Japanese (Kawahara et al. 2005), and Bengali (Khan to appear), to take but a few examples. Berkley (1994, 2000) found that the same is true of English: words like *king* or *mop*, in which two sounds with the same place of articulation are separated by a vowel, are underrepresented when compared to words with heterorganic consonants like *sing* and *mat*.

In the remainder of this chapter I will consider how this gradient OCP phonotactic constraint has shaped the lexicon of English. Instead of examining the cooccurrence frequencies of all English consonants, I will focus on just two, the liquids /l/ and /r/, for two reasons. First, unlike most other consonants, the cooccurrence frequency of the liquids is relatively unaffected by stress (Frisch 1996), so using these phonemes allows me to ignore this factor. The other reason is practical: much of the data I will analyze in this chapter is in orthographic form. In rhotic varieties of English, the graphemes *l* and *r* have the advantage of a nearly one-to-one mapping with the phonemes /l/ and /r/ respectively—orthographic data can therefore accurately represent the occurrence of liquids when phonetic transcriptions are not available.

### **3.2. Liquid cooccurrence in English neologisms**

Data from the lexical database CELEX (Baayen et al. 1993) confirms Berkley's (2000) finding that sequences of identical liquids separated by a vowel are underrepresented. In Modern English, /r/ and /l/ are of roughly equal frequency, with

/r/ having a slight advantage: of all the words containing exactly one liquid in CELEX (where cooccurrence with another liquid is not a factor), 43.3% (10,650 out of 24,598) contain /l/ and 56.7% (13,948 out of 24,598) contain /r/. Given these frequencies for the individual segments, we would expect that in words containing exactly two liquids, the liquids will be identical about half of the time (50.9% to be exact<sup>44</sup>).

In fact, identical liquids cooccur much less often than this, particularly when they are in close proximity in the word. Of the 1,739 words in CELEX which contain exactly two liquids separated by a vowel,<sup>45</sup> the liquids are identical in only 355 words, for an identity rate of 23.5%. It is clear that this is below the average expected rate of 50.9%, but it is not clear whether the difference is significant. In order to determine significance, I will use a Monte Carlo procedure (Kessler 2001) to approximate the distribution of the expected rate. The procedure is described in detail in the following section.

---

<sup>44</sup> The expected value is calculated as the probability that both liquids are /l/ added to the probability that both are /r/:  $.443^2 + .567^2 = .509$ .

<sup>45</sup> Because the phonetic transcriptions in CELEX represent a non-rhotic variety of British English, these calculations are based on the orthographic representations of the words (this also makes it easier to compare this data to the data sets presented later in the chapter, for which transcriptions were not available). This list therefore represents all words spelled with two liquids (the double letters *ll* and *rr* were treated as single letters) in which the string intervening between the two liquids consists only of one or more of the English vowel letters (*a, e, i, o, u, y*), or any vowel letter followed by *w* or *y*. Using orthography in this way results in a small amount of noise in the data, as in words like *liar*, where the two liquids are separated by more than a single vowel phoneme, but because the strength of the OCP lessens with distance, this noise is likely to make the OCP effects discussed in this chapter appear weaker than they actually are.

### 3.2.1. *The Monte Carlo test for significance*

The Monte Carlo test is performed as follows. First, a list of the words in question is compiled, and then converted into two lists, one consisting of the first liquid in each word, and another consisting of the second liquid in each word. The process is illustrated in (31), using the first 10 two-liquid words in the CELEX list as an example.

(31) Preparing a word list for Monte Carlo test

Word		Liquid 1	Liquid 2
accelerando		l	r
accelerate		l	r
acceleration		l	r
acrylic		r	l
admiral	→	r	l
admiralty		r	l
admirer		r	r
adorer		r	r
advalorem		l	r
adventurer		r	r

Once these parallel lists of cooccurring liquids have been compiled, the liquids can be recombined at random—this is most simply done by fixing the order of the liquids in the *Liquid 1* list, and placing the liquids in the *Liquid 2* list in a randomly determined order. After each such shuffling, the number of identical liquid pairs can be calculated, and the entire process repeated as many times as necessary. The process is illustrated in (32), with identical liquid pairs indicated by shading.



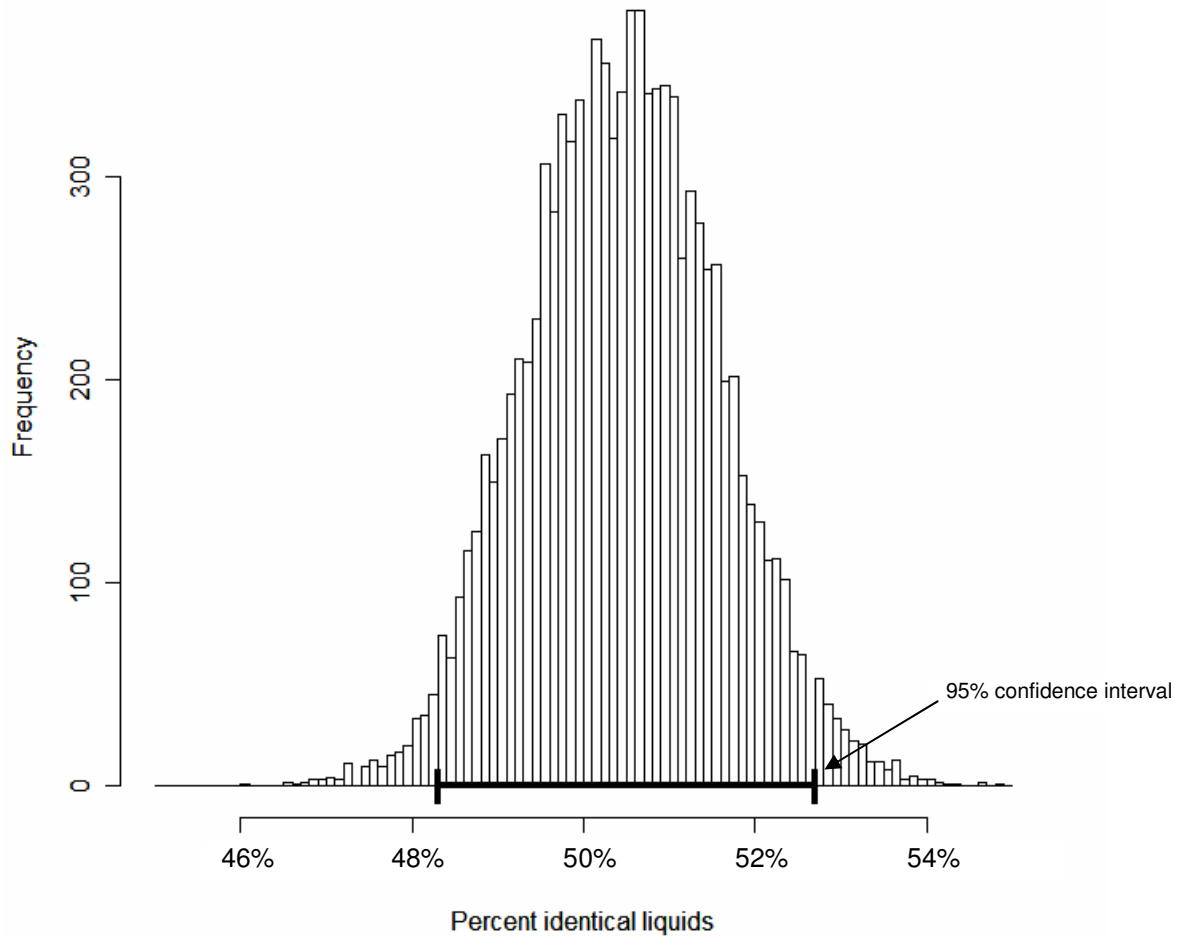
(32) Performing the Monte Carlo test

Initial state		→	First shuffle		→	Second shuffle		...
Liquid 1	Liquid 2		Liquid 1	Liquid 2		Liquid 1	Liquid 2	
l	r		l	r		l	l	
l	r		l	l		l	r	
l	r		l	r		l	r	
r	l		r	r		r	l	
r	l		r	r		r	r	
r	l		r	r		r	r	
r	r		r	l		r	l	
r	r		r	l		r	r	
l	r		l	r		l	r	
r	r		r	r		r	r	

If the shuffling is repeated sufficiently many times, the result will be a reliable estimate not only of the average expected number of identical liquids that would occur by chance, but of the entire distribution of this expected value. With this information, we can determine exactly how likely the actual value is. The histogram in (33) presents the results of the Monte Carlo procedure on the entire list of 1,739 words in CELEX which contain exactly two liquids separated by a vowel. The *x*-axis represents the percent of liquid pairs in which the liquids were identical, and the *y*-axis represents the number of shuffles (out of 10,000 total<sup>46</sup>) in which a given identity rate occurred.

<sup>46</sup> All of the Monte Carlo tests reported in this dissertation were performed using 10,000 iterations.

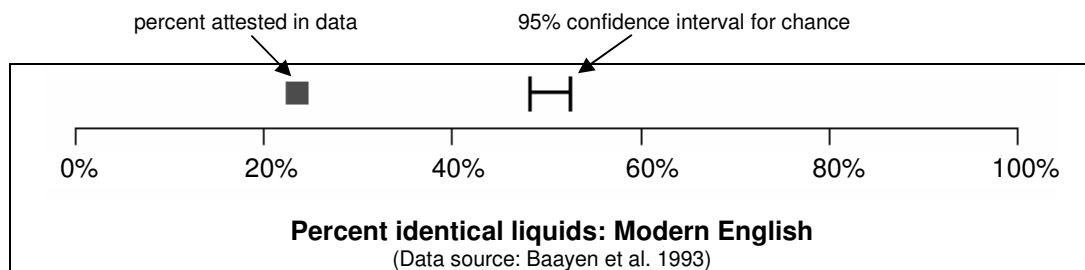
(33) Results of Monte Carlo test on CELEX two-liquid words



The histogram shows that the values generated in the Monte Carlo test are approximately normally distributed around the median value of 50.5%. The horizontal I-shaped bar indicates the 95% confidence interval around this median value—the range in which 95% of the values lie. The actual percentage of identical liquids in the CELEX data, 23.5%, is not only well below this interval, but is lower than even the lowest value generated in any of the 10,000 iterations of the Monte Carlo procedure. From this we can conclude that the actual identity rate is significantly ( $p < 0.0001$ ) below the identity rate that would be expected by chance.

The results of the Monte Carlo test can be more succinctly summarized by omitting the histogram and simply reporting the confidence interval and actual value. This is done in the chart in (34), which represents the same test reported in (33). In this chart, the gray square indicates the actual fraction of two-liquid words with identical liquids, while the horizontal bar represents the 95% confidence interval derived from the Monte Carlo test.

(34) Comparing attested CELEX liquid pairs to Monte Carlo results

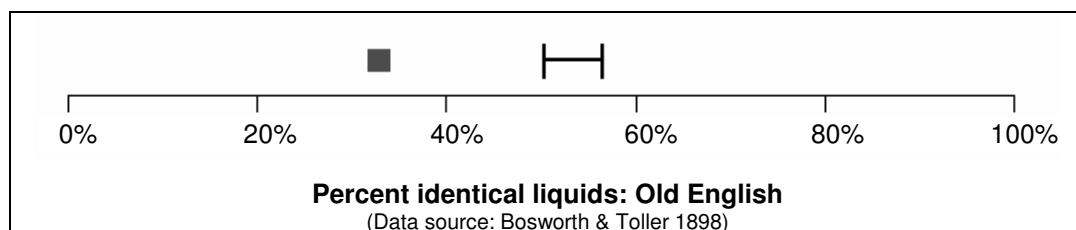


All of the Monte Carlo results in this and the following chapter will be reported using similar graphs.

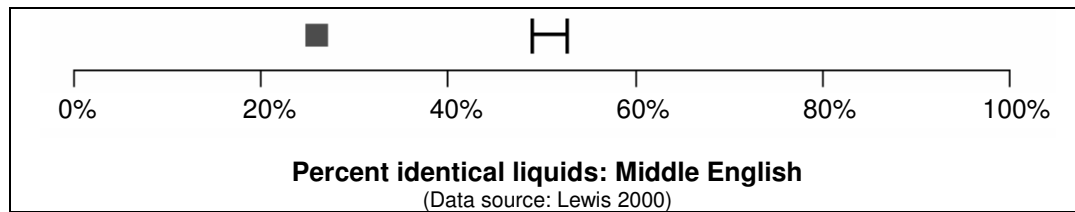
### 3.2.2. *The OCP in several stages of English*

This OCP effect is not merely an accidental property of current English; the charts in (35) and (36) show that the same bias is found in two older stages of the language: Old English (450-1100 AD) and Middle English (1100-1500 AD).

(35) Sequences of identical liquids are underrepresented in Old English



(36) Sequences of identical liquids are underrepresented in Middle English



Words with identical liquids have been underrepresented in the lexicon of English throughout the history of the language, despite the fact that in Modern English retains only 10-15% of the vocabulary of Old English (Stockwell and Minkova 2001).

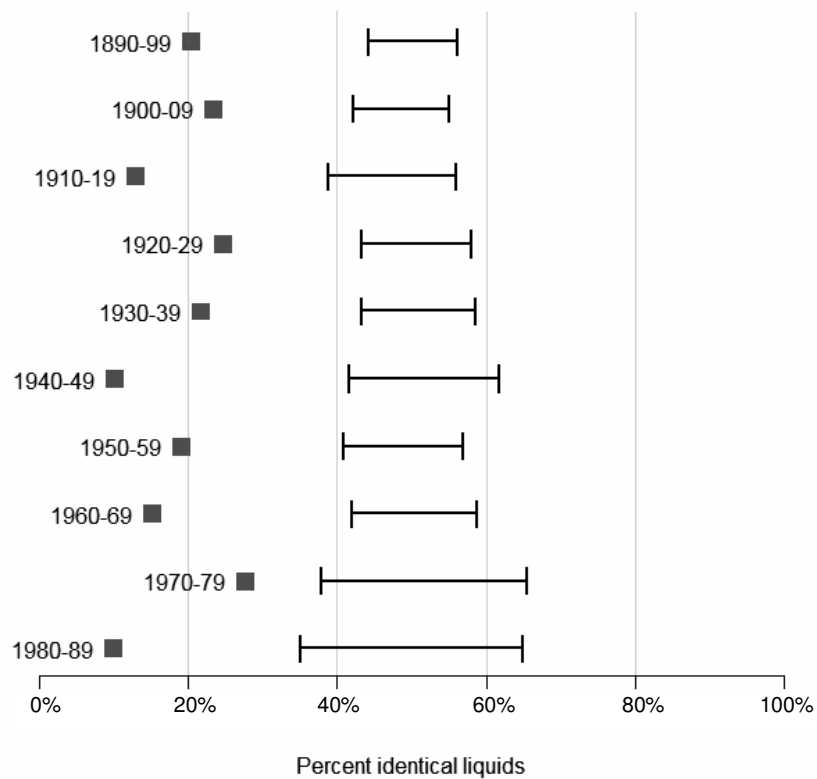
The underrepresentation in Modern English might be attributed to the large number of Latinate words in the current language—Latin had a rule of liquid dissimilation that is reflected in a number of Modern English words (e.g., *annu-al*, but *lun-ar*, not *\*lun-al*). The fact that Old English exhibits the same OCP effect, however, shows that the effect is not solely due to the influence of Latin, as the number of Latin words borrowed into Old English was quite small: less than 3% of the Old English vocabulary consisted of loanwords from any language (Culpeper 2005).

This consistent underrepresentation could be the result of sound change. If there were a tendency to misperceive or mispronounce /l/ as /r/, for example, when followed by another /l/, this could over time result in a lexicon skewed away from /l...l/ sequences. In order to examine this possibility, I collected a list of recent neologisms from the *Oxford English Dictionary* that first entered the language over the period of a century, from 1890 to 1989 (i.e., words whose earliest cited quotation falls between those years). Words that are entirely new to the language are much less

likely to have had time to accumulate sound change—thus, if sound change alone is responsible for the OCP effect in English, we would expect liquids in these novel words to cooccur relatively freely.

For each of the 38,232 neologisms from this time period listed in the OED, I considered only those words containing two liquids separated by a vowel (there were a total of 1,363 words fitting this description). The chart in (37) gives the percent of this subset of words in which the liquids are identical (as well as the corresponding chance confidence interval), divided by decade.

(37) OED neologisms by decade: liquid identity rates



As is clear from the chart, identical liquids are underrepresented in neologisms just as strongly as in older words. This strongly suggests that it is phonotactic preferences, and not sound change, that is driving OCP effects in the English lexicon.

Of course, many of these neologisms could be borrowed from languages which themselves underrepresent identical liquid pairs. A closer examination of the data as well as information on the lexical statistics of the languages English borrowed from during this period would be necessary to determine the degree to which the underrepresentation results from biases in English speakers. In the next section I examine data from baby names that demonstrates that English speakers are in fact biased by phonotactic preferences, and sheds further light on how a gradient phonotactic like the OCP can be maintained in an ever-changing lexicon.

### **3.3. Liquid cooccurrence in American baby names**

American given names tend to be characterized by the same phonotactic restrictions that hold for English words—no common names begin with [ɲ] or end with [h], for example, and *Bnick* is as unlikely to catch on as a first name as it is to become an English word.<sup>47</sup> In this section I will show that names also conform to gradient phonotactics; in particular, the OCP constraint that is the focus of this chapter.

---

<sup>47</sup> Of course, not all Americans are native English speakers, and so the analogy between American names and words of English is not perfect, but I will assume that the most popular names are at the very least heavily influenced by the phonotactic intuitions of the 82% of Americans that are monolingual English speakers (2000 US Census data; available at <http://factfinder.census.gov>).

The data I will describe on American first names comes from the U.S. Social Security Administration, consisting of the 1,000 most common names each for boys and girls for every decade in the twentieth century.<sup>48</sup> One advantage of this data is the degree of precision with which it allows us to examine naming trends, a precision that is often unavailable in lexicographic studies of other kinds of words. With the name data, we can pinpoint to within a year or so not only when a certain name became popular, but also when it fell out of use as a baby name. As I show below, this level of detail will allow us to ask questions about how lexical biases change that could not be posed with traditional dictionary or corpus data.

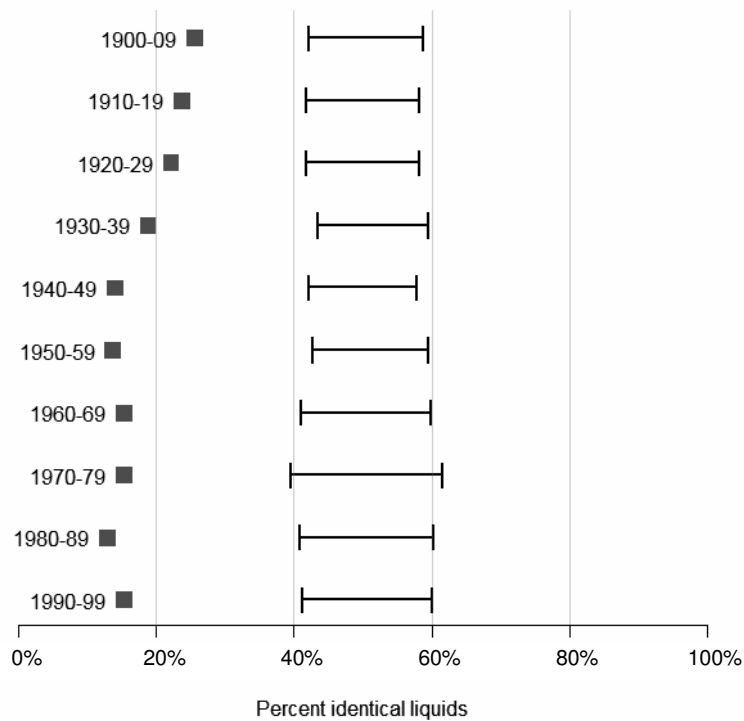
I used the Social Security Administration data to examine the effects of the OCP in English—just as words like *rare* or *lull* are underrepresented compared to words like *lair* or *real*, names like *Gerard* or *Leila* are less frequent than would be predicted by chance.<sup>49</sup> The chart of observed/expected values in (38), calculated over all names containing exactly two liquids separated by a vowel, shows that the same is true for every decade in the twentieth century.

---

<sup>48</sup> Available at <http://www.ssa.gov/OACT/babynames/>.

<sup>49</sup> The name data presented throughout this chapter assumes a rhotic dialect of English which is commonly spoken in the United States. Note that the theory makes different predictions for a non-rhotic dialect such as British English—names like *Gerard*, in which the second orthographic *r* is unpronounced, should be more common than names like *Leila* or *Rory*. I have not yet collected sufficient data on British names to test this prediction.

(38) Liquid pairs in popular names by decade



The OCP bias in names has remained remarkably stable<sup>50</sup> despite large changes in the actual names that were popular—a comparison of the top-1,000 lists for the first and last decades of the century shows that they only have 36% of the names in common.

Using the name data, we can also examine the role played by the OCP in whether a name made it into, or fell out of, the top-1,000 list. To do this, I divided the two-liquid names into two groups: “winners,” names that appear in one decade but did not appear in the previous decade, and “losers,” names that appear in one decade but do not appear in the following decade (a name may be both a winner and loser). Among the winners, the two liquids are identical only 11.8% (17/144) of the time,

---

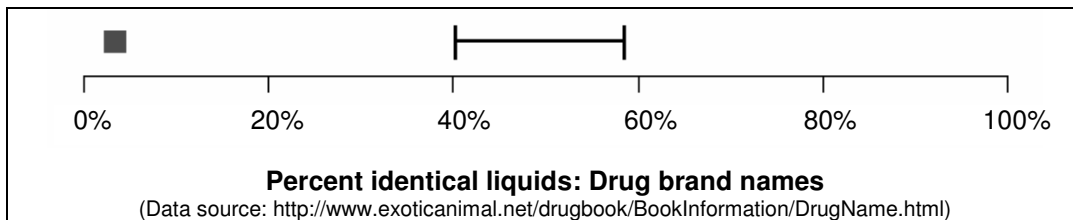
<sup>50</sup> Note the downward trend in the occurrence of identical liquids over the first half of the twentieth century. This is consistent with the theory advanced in §1.6 that phonotactic factors came to play a larger and larger role in naming choice as the century progressed.



while the losers have a higher identity rate of 19.4% (35/180). This difference is significant (one-tailed  $p < 0.05$  by Fisher's Exact Test), and indicates that a name with two identical liquids has a lower chance of becoming popular, and once popular, a higher chance of falling out of popularity.

The OCP is not just a feature of common names, which, it could be argued, are chosen from a pool of existing names which itself may have been shaped by sound change in the past. The same effect can be seen in newly coined names, which presumably have not had time to undergo sound change. The charts below show that identical liquids are underrepresented in drug brand names (39), a list of names invented for a fantasy role-playing game (41), and a list of “unusual, made-up” baby names (43). Examples from each data set are given in (40), (42), and (44), respectively.

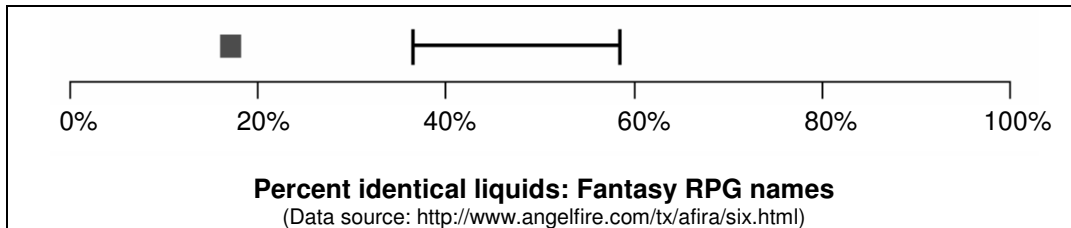
(39) Identical liquids are underrepresented in drug brand names



(40) Examples of two-liquid drug brand names ( $N = 88$ )

<b>l..r</b>	<b>r..l</b>	<b>r..r</b>	<b>l..l</b>
Choloromycetin Inteflora Seleron	Droleptan Demerol Oralet	<i>none</i>	Dalalone Imazalil Hemicelulose

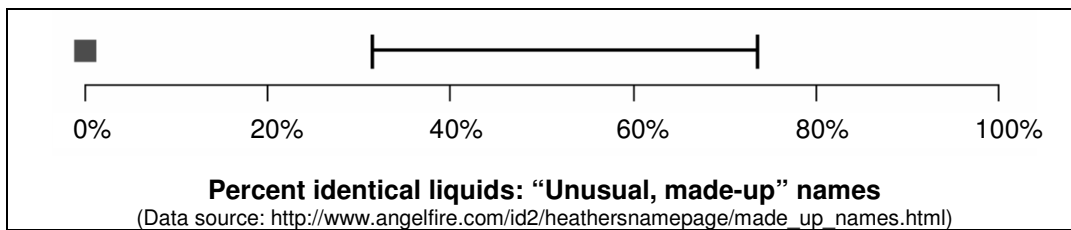
(41) Identical liquids are underrepresented in names for fantasy role-playing game characters



(42) Examples of two-liquid role-playing game character names ( $N = 82$ )

<b>l..r</b>	<b>r..l</b>	<b>r..r</b>	<b>l..l</b>
Balor	Cynoril	Adraeran	Laeli
Falyrias	Xarolith	Aurora	Loili
Larn	Imril	Azhrarn	Lylas

(43) Identical liquids are underrepresented in unusual baby names



(44) Examples of two-liquid unusual baby names ( $N = 20$ )

<b>l..r</b>	<b>r..l</b>	<b>r..r</b>	<b>l..l</b>
Alera	CamBrel	<i>none</i>	<i>none</i>
Clarendy	Raleda		
Islara	Veralidaine		

These results suggest that the OCP bias in English does not only affect a word's ability to spread throughout the speech community; rather, it appears to directly affect the creation of new words (or names).

### 3.4. Processing and the OCP

Dell et al. (1997) propose a spreading activation model of speech which can sequence units of speech. The model uses three mechanisms to insure that phonemes

(or other phonological units) are activated in the correct sequence: one which activates the unit presently being prepared, a second which deactivates the units that have already been produced (so they are not confused with the current unit), and a third which primes (i.e., starts to activate) the units that will be produced after the current unit. Together these three mechanisms ensure that at any point during an utterance, the phoneme being produced is the most active node.<sup>51</sup> Given this model, there are several possible theories regarding how sequences of identical or similar sounds could interfere with processing. Below I summarize three such theories, and discuss their implications for my model.

#### 3.4.1. *Refractory period*

One way that sequences of similar or identical segments could impede processing involves the mechanism responsible for deactivating past nodes—what Dell et al. (1997) call the *turn-off function*. Imagine, for example, that someone is attempting to produce the word *lull*. This involves first activating the phoneme node for /l/, then the node for the vowel /ʌ/, and then the node for /l/ again. When /l/ is activated the second time, however, it has just been deactivated by the turn-off function—if we assume either that there is a refractory period during which a deactivated node cannot be activated again (or that the deactivation and activation functions may overlap, causing them to interfere with one another), using the same

---

<sup>51</sup> The question of how the production system represents temporal sequences, such that it knows which phonemes to activate in which order, has spawned several theories in the literature. Some models invoke associations between successive phonemes (Elman 1990), other posit frames into which phonemes are slotted (Dell 1986), and yet others propose a time-varying control signal to encode serial order (Vousden et al. 2000). This issue is orthogonal to my concerns in this chapter.

phoneme twice in a row will take more time than using two different phonemes (MacKay 1970, Frisch 2004). The second /l/ in a word like *lull* will be activated more slowly than the first. This will have two consequences—first, it will put *lull* at a disadvantage in the competition to be selected when compared to synonyms that do not contain sequences of identical consonants. Second, it will make it more probable that another phoneme will reach activation first, resulting in a speech error. Experimental and corpus studies of speech errors have confirmed this prediction—a sequence of identical consonants increases the chance of making an error on that sequence (Shattuck-Hufnagel 1979, Dell 1986, Stemberger 1990, Wilshire 1999).

Assuming that nodes for features or gestures also have refractory periods, sequences of consonants that share many features will also lead to processing difficulties. This account thus predicts that the type frequency of a given sequence of consonants in the lexicon should be inversely correlated with the similarity of the consonants. Identical consonant pairs should be most strongly underrepresented, followed by pairs of highly similar consonants. Although this is true for the English liquids, as well as consonants in Hawaiian and Croatian (MacKay 1970), a different pattern is found in other languages. In Ngbaka (Broe 1995), Muna (Coetzee and Pater 2005), and Japanese (Kawahara et al. 2005), as well as many other languages, identical consonants cooccur freely—highly similar consonant pairs are underrepresented, but identity serves as an “escape hatch” that permits violations of the OCP (MacEachern 1999 also lists categorical constraints of this type from languages with laryngeal cooccurrence constraints).

English, in fact, exhibits the same identity effect for certain segments in certain contexts—identical segments cooccur more frequently in stressed syllables, and less sonorous identical segments cooccur more freely than more sonorous segments (Frisch 1996, Berkley 2000). Frisch and Zawaydeh (2001) also found experimental evidence for the special status of identical consonants—native Arabic speakers rated nonwords with sequences of identical consonants as more well-formed than nonwords with sequences of similar but nonidentical consonants, judgments that are in the opposite direction of the lexical statistics of Arabic, where identical consonant pairs are rarer than similar consonant pairs.

The refractory period hypothesis cannot explain the special status of identical segments, which it predicts should cause the greatest processing difficulty. In the next section I examine another hypothesis, which can explain why identity is sometimes better than near-identity.

### 3.4.2. *Confusability*

Another source of difficulty for production can be located in the effects of noise on the mechanisms proposed by Dell et al. (1997), which can result in speech errors. Perseveratory errors (e.g. *walk the beak* for *walk the beach*) result when an already-produced unit is not deactivated quickly enough and is mistaken for the current unit; anticipatory errors (e.g., *cuff of coffee* for *cup of coffee*) occur when future units are primed too quickly and are accidentally chosen instead of the current unit.

As with the refractory period theory, this confusability theory predicts a correlation between similarity and error rate: consonants similar to the target segment will be more highly activated (because they are connected to many of the same feature nodes as the target) and thus more likely to be selected instead of the target. Sequences of highly similar consonants are therefore more likely to result in errors, as can be seen by the preponderance of such sequences in tongue twisters (e.g., *She sells seashells by the seashore*). Identical consonants, however, are a special case—substituting one segment for an identical segment has no perceptible effect, and would not be categorized as an error. If underrepresentation results from a mechanism which avoids words that are likely to cause errors, words like *sash* should be underrepresented, but not words like *sass*.

### 3.4.3. *Repetition blindness*

Another possibility is that repetition causes problems for perception. Evidence from both visual and auditory processing indicates that two identical stimuli presented in rapid succession are often perceptually fused, causing subjects to report that they perceive only a single instance of the stimulus. This phenomenon, known as *repetition blindness* (Kanwisher 1987, Bavelier 1994), could make it difficult to perceive identical or similar consonants as distinct. Boersma (1998) proposes that this perceptual bias is the basis of a phonological constraint on adjacent identical elements. Frisch (2004) further argues that repetition blindness could also be the cause of the

long-distance OCP effects described in this chapter—sequences of identical consonants are avoided because they are difficult to perceive.

This hypothesis makes the same prediction as the refractory period theory: completely identical segments are the most likely to be perceptually fused, and so should be the most strongly underrepresented. The token frequency of consonant sequences should thus be inversely correlated with similarity. Just as with the refractory period theory, the repetition blindness theory can account for liquid cooccurrence in English, but cannot explain cases in which pairs of identical segments occur with greater frequency than pairs of highly similar segments.

#### 3.4.4. *Why is repetition difficult?*

This survey of theories that attempt to explain the burden imposed on processing by repetition leaves us with a conundrum. Two theories predict that identical segments should be the worst combination, while one predicts that they should be the best. The problem is that both predictions are true for different languages, and even for different contexts within the same language.

One possible solution to this puzzle is suggested by Zuraw's (2002) theory of Aggressive Reduplication. She uses extensive evidence from the active phonology of Tagalog to argue for a family of Optimality Theoretic markedness constraints, IDENT- $\kappa\kappa$ , which enforce identity between different substrings within the same word.<sup>52</sup>

Zuraw uses this framework to discuss an identity escape clause in a laryngeal

---

<sup>52</sup> A separate set of constraints is responsible for determining which substrings are in correspondence (*coupled*, in her terminology) with each other; for details, see Zuraw 2002.

cooccurrence restriction in Peruvian Aymara reported by MacEachern (1999): a morpheme may not contain more than one ejective consonant, unless the ejectives are identical (a parallel restriction holds for aspirated consonants). Zuraw argues that this is plausibly the result of Aggressive Reduplication—in words with multiple identical ejectives, the vowels following the ejectives are more likely to be identical than chance would predict. Thus, the special status of identical consonants in Aymara appears to be part of a more general phenomenon involving a pressure towards identity of entire strings. Zuraw speculates that Aggressive Reduplication constraints could play a role in lexical learning, causing learners to mislearn words in such a way as to inflate the numbers of words with identical subparts (as is evidenced in errors made by English speakers like *orangutan* → *orangutang*). However, the same constraints could also act to bias lexical competitions, leading speakers to prefer words that contain identical substrings.

Because the effects of Aggressive Reduplication constraints are dependent on their language-specific ranking, it is not surprising that languages vary in the statistical representation of identical segments. It is not only the treatment of identity, however, that differs across languages—Coetzee and Pater (2005) and Khan (to appear) propose that different languages give different weights to various phonological features when computing the similarity of two segments, so that what counts as similar is also in part language-specific. How these differences arise, and whether they are learned, or somehow derived from other properties of a language's



phonological system, remains unclear; I leave a comprehensive study of these issues to future research.

### **3.5. Extending the model**

The fact that different languages (or as in English, different contexts within the same language) treat identical consonant pairs differently makes it difficult to construct a language-independent model of speech production that can account for lexical OCP effects. For this reason, rather than implementing a single concrete model of lexical selection, I will discuss more generally how the processing difficulties predicted by each of the three theories described in the previous section place constraints on such a model.

Together, the three theories predict two different types of processing difficulty: processing slowdown, and errors in production or perception. The first, processing slowdown, is predicted only by the refractory period hypothesis. If it is difficult to activate the same phoneme node twice in rapid succession, then words containing sequences of identical (or similar) consonants will take longer to activate, putting them at a disadvantage in the race to be selected over other synonyms. This can be straightforwardly handled in the speech production model I proposed in chapter 1, in which a word's fitness is a direct result of how quickly it can be activated.

The second type of difficulty, production or perception errors, is predicted by all three theories, although for different reasons. Under the refractory period theory, the slower activation of the second consonant in a series allows nodes for other

consonants to become active and thereby misselected. The confusability theory predicts that sequences of similar consonants are more susceptible to errors because of their many shared features. The repetition blindness theory predicts that such sequences have a higher chance of being misperceived due to imperfections in the part of the perceptual system responsible for deciding how many events constitute a given stimulus.

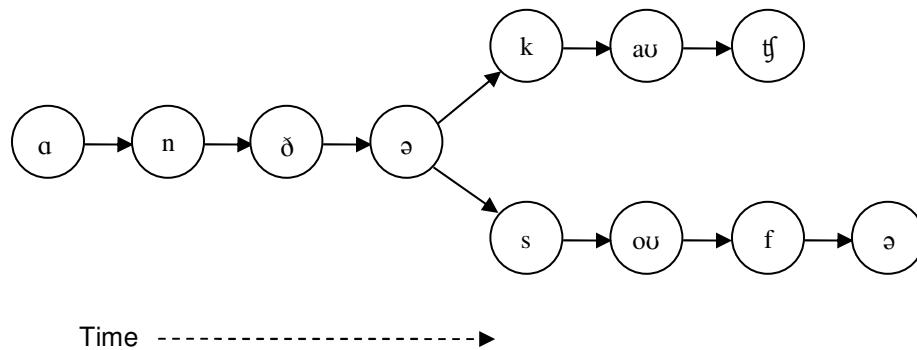
The issue of how error patterns end up being reflected in lexical statistics is more complex. It could be that some persistent speech or perception errors cause sound changes within individual lexical items, but much of the data in this chapter has shown that not all OCP biases can be the result of sound change. Another possibility is that errors affect the resting activations of lexical nodes. Normally, when a lexical item is used or heard, its resting activation (and thus the probability that it will be selected in the future) increases. Perhaps when an error (in production or perception) occurs, the resting activation is not increased, or is decreased. This error-driven feedback would eventually result in difficult-to-process words becoming underrepresented, as they are replaced by less problematic synonyms.

The problem with this story is that it does not predict any generalization on the part of speakers or hearers. Experiments such as those performed with Arabic speakers by Frisch and Zawaydeh (2001) show that speakers dislike nonce words that contain difficult-to-process sequences, even though they have had no chance to make errors using those words. It is therefore more plausible that speakers' avoidance of difficult words is a consequence of generalized knowledge—speakers know, not just

which words tend to result in errors, but which sequences tend to result in errors, and avoid words containing those sequences. This knowledge could either be an innate component of Universal Grammar, or be constructed by language users from their experience in making and hearing errors.

How specifically could the model of speech production I have developed in this dissertation be modified to incorporate this knowledge of error patterns? The answer to this question depends on how lexical selection and sequencing interact. If lexical selection is completed before sequencing begins—that is, if one lexeme is definitively chosen out of a set of synonyms, and only then is sequenced, then difficulties in sequencing should not affect selection. If the processes overlap, however, and sequencing is begun on multiple synonyms in parallel before one is chosen over the others, then long-distance OCP violations will have an opportunity to influence the selection process. This parallel sequencing process is illustrated schematically in (45); in this example, the two lexemes *couch* and *sofa* are competing to express a single concept. Note that this diagram does *not* represent a spreading-activation network—it merely represents the linear sequencing of phonemes, with arrows indicating temporal order.

(45) Sequencing synonyms in parallel: *on the (couch/sofa)*



Evidence that synonyms are sequenced in parallel comes from blend speech errors, in which parts of two words are combined into one (e.g., *frowl* from *frown* and *scowl*). These errors are usually formed from synonyms, or words that are closely related semantically (Wells 1951, Fromkin 1971, Poulisse 1999), and preserve the linear order of segments in both words. It is thus likely that such errors originate during sequencing, but before either synonym has been definitively selected.

How does the speaker's knowledge of error patterns impinge on lexical selection in this model? One possibility is suggested by theories of self-monitoring in speech production. These models assume that speakers are able to monitor their "inner speech"—the words that have been phonologically encoded and sequenced, but not yet spoken—in order to detect speech errors as quickly as possible. Levelt (1983) concludes from a study of a corpus of self-repairs that this monitor can notice errors in a word even before articulation of the word has begun.

One can imagine that the duties of this self-monitor include not only spotting errors that have already occurred, but also noticing and avoiding potential trouble

spots in the course of processing that are likely to lead to errors. When this happens, the monitor can suppress the activation of the offending lexical item. If there is a synonym available that doesn't violate the OCP (or violates it less severely), that synonym will have a higher probability of being selected. Over time, this bias will lead to the underrepresentation of identical or similar consonant sequences.

Some support for this picture comes from the fact that similarity avoidance constraints are strongest at the beginning of the word (Frisch 1996, 2000). This follows if we assume that the overlap between lexical selection and sequencing is partial rather than total. In other words, even if sequencing begins on multiple synonyms before one is chosen, selection may occur before sequencing is completed. Thus, a word-initial sequence of similar consonants will almost always play a role in selection, but the same sequence at the end of a longer word will have little influence, because by the time the sequencer reaches that point, a winner in the synonym competition will have already emerged. In other words, an /VI/ sequence at the end of the word is just as hard to process as one at the beginning of the word, but the word-final sequence simply occurs too late to affect selection, and consequently has less impact on the shape of the lexicon.

#### **4. Morphologically driven phonotactic preferences**

In this chapter I examine a type of phonotactic preference that is governed not by universal facts of articulatory or processing ease, but by language-specific phonotactic patterns. I will show that phonotactic restrictions that hold categorically within morphemes tend to “leak” into larger domains, resulting in weaker, gradient versions of the same restrictions across morpheme boundaries. I will argue that this is the result of a kind of phonotactic preference—speakers prefer to form complex words that obey stem-internal phonotactics.

I will present evidence for this from three languages, English, Navajo, and Turkish. In all of these languages, compounds that violate a stem phonotactic are attested, but are rarer than compounds that obey the phonotactic. In Navajo, for example, compounds are permitted to violate a sibilant harmony constraint that is unviolated in stems, but disharmonic compounds are statistically underrepresented. I show that this can be modeled as a side effect of generalizations formed during early phonotactic learning.

##### **4.1. The proposal**

I propose that the low type frequency of these compounds in each language is the result of their being assigned intermediate well-formedness by the phonotactic grammar. This in turn, I argue, is the result of multiple, overlapping generalizations formed by learners in the process of acquiring a language’s phonotactics. I focus on

the interaction between two types of generalization: one that takes into account morphological structure, *structure-sensitive*, and another that ignores morphological structure, *structure-blind*. Crucially, both types of generalization, although they may make conflicting predictions, are combined by the grammar when assigning a probability to a potential output. This is why a complex word, which may be perfect according to structure-sensitive generalizations, may be nonetheless penalized by a structure-blind generalization.

I formalize this as a grammar consisting of weighted constraints, coupled with a learning algorithm which uses the principle of Maximum Entropy (Goldwater and Johnson 2003, Hayes and Wilson to appear) and a smoothing term that penalizes complex grammars. Maximum Entropy grammars can account for both categorical and gradient generalizations, and are thus ideally suited to explain the data discussed here, in which some configurations are banned outright within morphemes, but only gradiently dispreferred across morpheme boundaries. I show that such a learner, if equipped with both types of constraints described above, automatically produces a bias against complex words that violate morpheme-internal phonotactics, even when no such bias exists in the learning data; the model thus predicts the correlation between tautomorphemic and heteromorphemic phonotactics observed in the three case studies presented here.

The chapter is organized as follows. In §4.2.1, §4.3, and §4.4 I describe the English, Navajo, and Turkish data respectively. In §4.5 I discuss the theoretical consequences of the data for a learning model. In §4.6 I informally describe the

leaning model, which is then formally presented in §4.7 and §4.8. Finally, in §4.9 and §4.10 I show how the phonotactic learner can be integrated with the speech production model, and speculate on how these lexical biases are maintained over time.

## **4.2. English consonant clusters**

The phonotactic grammar of English places restrictions on the consonant clusters that may occur within a morpheme—no English morpheme, for example, contains the sequence /pf/. These restrictions typically do not apply across morpheme boundaries: the sequence /pf/ is perfectly acceptable if it occurs in a compound like *grapefruit*. I will argue, however, that the same constraints that block certain tautomorphemic clusters do in fact apply to heteromorphemic clusters, albeit in a gradient rather than categorical form.

In order to determine which clusters are legal within morphemes, I used the CELEX-derived list of monomorphemes compiled by Hay et al. (2003).<sup>53</sup> I define a tautomorphemically legal cluster as one which occurs word-medially in at least one word in this list. The chart in (46) shows which clusters are considered legal according to these criteria (attested clusters are indicated with a check mark).

---

<sup>53</sup> I am grateful to Janet Pierrehumbert for supplying me with this list.



(46) Legal word-medial CC clusters in English monomorphemes<sup>54</sup>

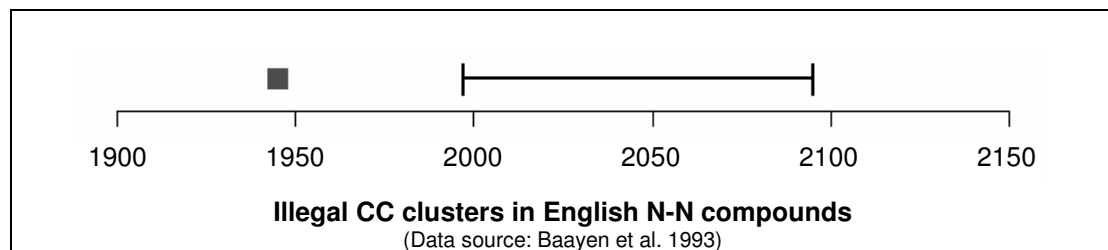
		Second Consonant																									
First Consonant		p	t	k	b	d	g	f	v	θ	ð	s	z	ʃ	ʒ	tʃ	dʒ	m	n	ŋ	l	r	j	w	h		
	p		✓	✓								✓		✓		✓		✓	✓		✓	✓	✓				
	t			✓				✓				✓		✓				✓	✓		✓	✓	✓	✓	✓		
	k		✓		✓	✓		✓	✓			✓		✓		✓		✓	✓		✓	✓	✓	✓			
	b					✓		✓	✓			✓					✓		✓		✓	✓	✓				
	d	✓		✓					✓									✓	✓		✓	✓	✓	✓			
	g				✓													✓	✓		✓	✓	✓	✓			
	f		✓								✓								✓		✓	✓	✓				
	v																		✓			✓	✓	✓			
	θ																	✓	✓		✓	✓	✓	✓			
	ð																	✓				✓					
	s	✓	✓	✓				✓		✓							✓		✓	✓		✓	✓	✓	✓		
	z				✓													✓	✓			✓		✓			
	ʃ	✓						✓															✓	✓	✓		
	ʒ																										
	tʃ				✓																						
	dʒ		✓																								
	m	✓	✓	✓	✓	✓		✓	✓				✓	✓	✓				✓			✓	✓	✓			
	n		✓	✓		✓		✓	✓	✓			✓	✓	✓		✓	✓	✓			✓		✓			
	ŋ		✓	✓			✓						✓										✓				
	l	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓		✓	✓	✓	✓			✓	✓	✓		
	r																							✓	✓		
	j																										
	h																							✓			
	w																										

In order to compare this set of tautomorphemically attested consonant clusters, I extracted all of the words marked as noun-noun compounds from the lemmatized version of the CELEX database (Baayen et al. 1993), a total of 4,758 words, and compiled a list of the consonant clusters created in each compound (i.e., the final segment of the first compound member followed by the initial sound of the second member). Each such cluster was then designated as legal or illegal, depending on

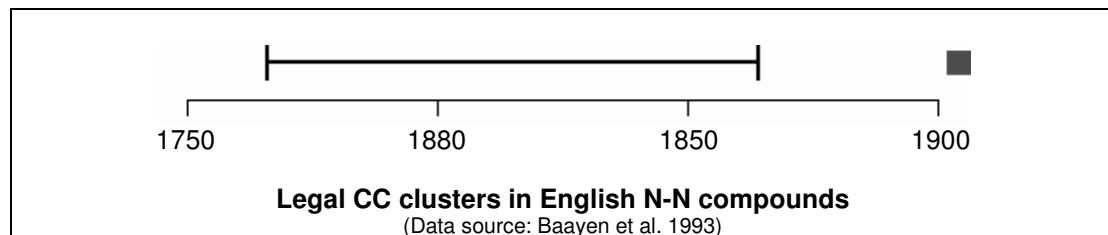
<sup>54</sup> The low number of attested clusters with /r/ as the first consonant is a consequence of the fact that CELEX, the database from which this chart was constructed, uses transcriptions based on a non-rhotic British dialect of English.

whether it occurred in any of the monomorphemes described in (46); 1,945 of the 4,758 compounds (40.9%) contain tautomorphemically illegal clusters. Finally, I performed a Monte Carlo test (see §3.1) on the set of compound clusters to determine the expected number of illegal clusters predicted by chance. The results, presented in (47), show that consonant clusters that are illegal within morphemes are underrepresented in compounds. Tautomorphemically legal clusters, on the other hand, are overrepresented, as shown in (48).

(47) Illegal non-geminate clusters are underrepresented in compounds



(48) Legal clusters are overrepresented in compounds



This correlation is evidence that the categorical phonotactic restrictions that hold within morphemes also hold gradiently across morpheme boundaries. This could result from a tendency for speakers to avoid forming compounds that would create illegal consonant clusters, or from the tendency of such compounds, once formed, to be replaced by competing synonyms.

However, there are problems with determining which clusters are legal by examining a corpus as I have—some clusters may occur in words that did not happen to appear in the corpus, for example. It also seems probable that the set of illegal clusters is a category with fuzzy boundaries—what, for example, is the status of clusters that occur in only one or two low-frequency words? For these reasons, in the remainder of this chapter, rather than consider the entire set of illegal consonant clusters, I will restrict my discussion of English to a subset of illegal clusters for which native speaker intuitions are clear: geminate consonants.

#### 4.2.1. *Geminates in English*

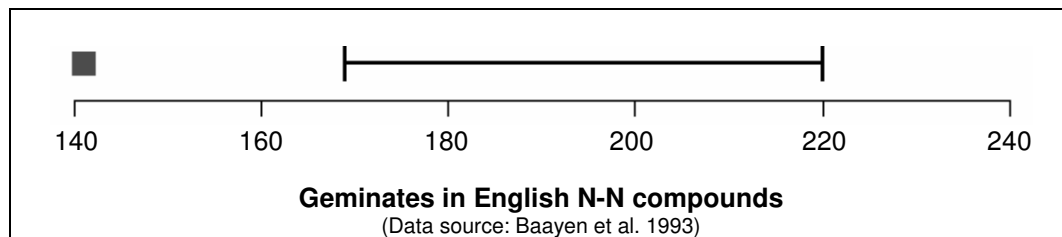
Geminate consonants in English are permitted only across morpheme boundaries (Hammond 1999, Ladefoged 2001, Kaye 2005). Words like *unknown*, *solely*, and *bookcase* are typically pronounced with geminates that have been created by combining morphemes that end and begin with the same consonant. These morphologically-created geminates are often called “false geminates” to differentiate them from morpheme-internal long consonants—the two types of geminate often exhibit different phonological behavior. Minimal pairs differing only in consonant length, as in the compounds *carpool* and *carp pool*, may be found in multimorphemic words; in monomorphemic words, no such minimal pairs exist—the hypothetical word [hæppi], which would form a minimal pair with existing *happy* [hæpi], is not a possible monomorpheme of English. In the following sections, I show that just as with illegal consonant clusters in general, geminate consonants created by

morphological concatenation are statistically underrepresented in the lexicon of English. The results are discussed separately for compounds and affixed forms in §4.2.2 and §4.2.3 respectively.

#### 4.2.2. *Compounds with geminates*

The data discussed here utilize the same 4,758 noun-noun CELEX compounds described in §4.2. Of these, 141 (3.0%) words contain false geminates—e.g., *bus stop*, *hat trick*, *penknife*, *bookkeeper*. The results of a Monte Carlo test on the CELEX compounds are shown in (34).

(49) Geminates are underrepresented in English compounds



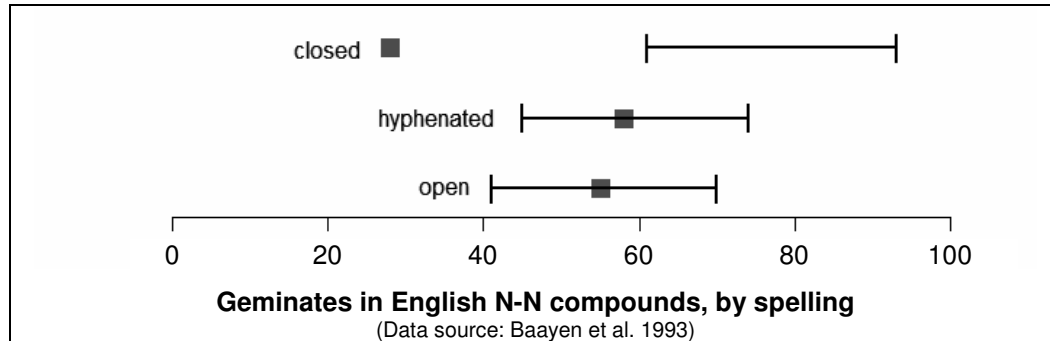
As the chart makes clear, the number of geminates found in the actual compounds, 141 (3.0%), is significantly lower than expected ( $p < .001$ ).

In a corpus-based study of how English compounds are spelled, Sepp (2006) found that whether or not a word has a geminate affects the way it is spelled.

Compounds in English can be spelled one of three ways: *open*, with a space between the compound members (e.g., *sand dune*), *hyphenated* (e.g., *roller-skate*), or *closed* (e.g., *joystick*). The chart in (50) shows that compounds with geminates are more likely to be spelled hyphenated or open than closed (this chart depicts the same set of

words from CELEX described in (34), divided according to how they are spelled in the CELEX entry).<sup>55</sup>

(50) Geminates are underrepresented in compounds spelled closed



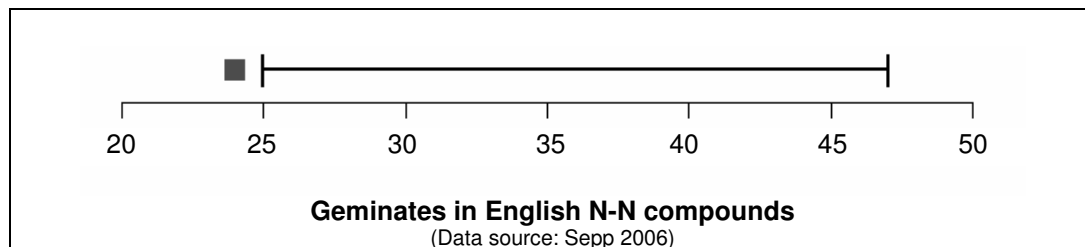
This difference in spelling suggests a possible source for the underrepresentation of geminates seen in (34). If the parsing algorithm used to identify compounds in CELEX failed to identify some open compounds as compounds (which is more likely than failing to identify a closed compound), then the underrepresentation of geminates could be an artifact of the parsing process, combined with people's tendency to spell compounds according to their junctural phonotactics. To show that this is not the case, I ran the same Monte Carlo test on the list of compounds compiled by Sepp (2006) from a 14-million-word corpus of American English (for details of the construction of this corpus, see Sepp 2006). Sepp used a part-of-speech tagger and computational parser to extract all potential noun-

<sup>55</sup> Although the same compound can be spelled different ways by different writers, each compound is listed with a single spelling in CELEX. It is unclear how this spelling was determined. My intuition is that nearly all of the words spelled with hyphens in CELEX would be most often spelled open by native speakers (e.g., *space-vehicle*, *rabbit-hutch*, *slot-machine*), a suspicion that is strengthened by the nearly indistinguishable behavior of hyphenated and open compounds in (50). This accords with Sepp's findings that less than 5% of the noun-noun compounds in her corpus are spelled with a hyphen more often than either open or closed (many of those are either dvandva compounds (*Clinton-Gore*, *hip-hop*), or involve abbreviations(*op-ed*)).

noun compounds, and then further filtered the list by hand, removing all non-compounds. Because every compound was checked by hand, the likelihood of undercounting open compounds is much lower than it would be if all parsing were done by algorithm.

Of the 708 compounds with a frequency of 35 or more in Sepp's corpus (including both those with open and closed spellings), 24 (3.4%) contain false geminates. The results of a Monte Carlo test on these compounds, shown in (51), demonstrate that, just as with the CELEX compounds, the actual number of geminates is significantly lower ( $p < .05$ ) than the mean expected number of 35.9 (5.5%).

(51) Geminates are underrepresented in compounds in Sepp corpus



Thus, even when the risk of a counting bias is minimized by careful hand-checking, geminates are still underrepresented in compounds overall. This suggests that any orthographic bias that people may have is in addition to a general bias against forming compounds that create geminates.

#### 4.2.3. *Suffixed words with geminates*

Geminates are underrepresented not only in compounds, but in affixed forms as well. As noted above, false geminates may be created at the boundary between a

stem and a level 2 affix, as in *solely* [soulli] (potential geminates created at level 1 morpheme boundaries, as in *innate* [inert], are repaired through degemination). In this section I present evidence that words suffixed with *-ness*, *-ly*, and *-less* contain fewer geminates than expected.

The suffix *-ness* attaches to adjectives to form nouns, as in *random* → *randomness*. In order to determine how many geminates were created in words with this suffix, I first extracted all adjectives (i.e., potential bases) from the CELEX database. I removed all suffixed words<sup>56</sup> so as to avoid complications induced by suffixes interacting with each other (proper names and words spelled with a space were also omitted), leaving a total of 1,736 adjectives. I then searched CELEX for all words consisting of one of these adjectives suffixed with *-ness*, resulting in a total of 281 suffixed forms (e.g., *roughness*, *vagueness*, *alertness*). Of these 281 words, 18 (6.4%) contain geminates (e.g., *cleanness*, *openness*). This does not differ from chance, although this is likely due to the small sample size. I therefore considered *-ness* suffixed words that contain any tautomorphically illegal cluster (a total of 46 words).

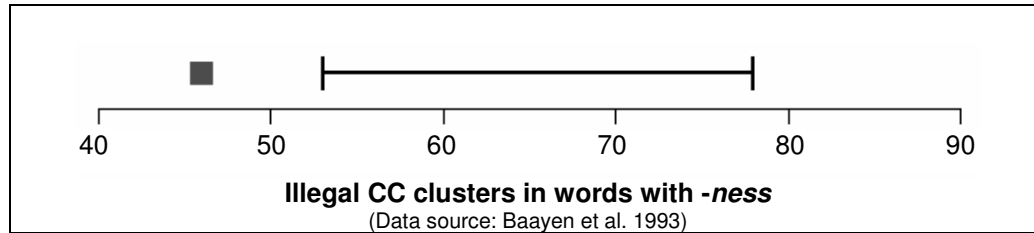
To determine the number of illegal clusters predicted by chance, I performed a Monte Carlo test by choosing 281 words (the number of suffixed words) at random from the set of 1,736 adjectives (the number of potentially suffixed words), and determining how many illegal clusters would be created if these words were suffixed

---

<sup>56</sup> The morphological parsing given for each word in the CELEX database (EML.CD) was used to determine which words were suffixed. In addition, words ending in the orthographic strings *-ing*, *-ed*, and *-en* were removed, as CELEX does not parse out these inflectional suffixes.

with *-ness*. The results of performing this Monte Carlo test 10,000 times are given in (52).

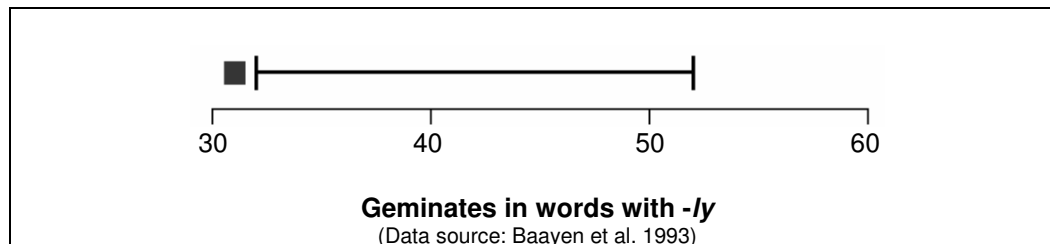
(52) Illegal consonant clusters are underrepresented in *-ness* suffixed words



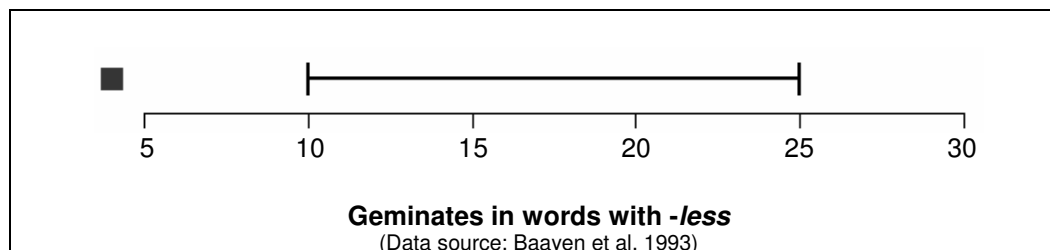
As the chart shows, suffixed words with illegal clusters are underrepresented, just as in compounds.

I performed similar tests, using the same method, for adjectives suffixed with *-ly* (e.g., *quickly*) and nouns suffixed with *-less* (e.g., *hopeless*); in these cases larger numbers allowed me to consider the creation of geminates specifically, rather than all illegal clusters. The results of these tests are shown in (53) and (54).

(53) Geminates are underrepresented in *-ly* suffixed words



(54) Geminates are underrepresented in *-less* suffixed words



For these two suffixes, as with compounds, geminates are underrepresented.



#### 4.2.4. *Summary of English data*

In this section I have presented evidence that geminates created by level 2 morphology are underrepresented in the English lexicon. This is true for both compounds and for the suffixes *-ness*, *-ly*, and *-less*. Interestingly, no corresponding effect was found for the prefix *un-*, despite its being roughly as productive as the suffixes described above—the number of geminates in prefixed forms was at chance. The lack of an effect at a prefix-stem boundary may be the result of the well-known tendency for prefixes to be prosodically more loosely affiliated with their bases than suffixes (Peperkamp 1997), although more research would be required to confirm this hypothesis.

### 4.3. Navajo sibilant harmony

All sibilants in a Navajo root must agree in their specification for the [anterior] feature; thus, a single root can only contain sibilants that are either all anterior or all posterior (Sapir and Hojier 1967, Kari 1976, McDonough 1991, 2003, Fountain 1998). The two sets of consonants are summarized in the chart below.

(55) Navajo sibilant classes

[+anterior]	[-anterior]
s	ʃ
z	ʒ
ts <sup>h</sup>	tʃ <sup>h</sup>
ts	tʃ
ts'	tʃ'

Thus, for example, roots like /tʃ'ɔʃ/ ‘worm’ or /ts'ózí/ ‘slender’ are attested, but \*/soʃ/ is not a possible Navajo root.

This is not only a cooccurrence restriction on roots—sibilants in affixes must also agree in anteriority with sibilants in the root, resulting in alternations in sibilant-bearing affixes (Sapir and Hojier 1967). The examples in (56) demonstrate the alternations in prefixed forms (sibilants are in bold).

(56) Examples of sibilant harmony (Fountain 1998)<sup>57</sup>

- |                          |   |                       |                  |
|--------------------------|---|-----------------------|------------------|
| (a) /ji- <b>s</b> -lééʒ/ | → | [ji- <b>ʃ</b> -tlééʒ] | ‘it was painted’ |
| (b) /ji- <b>s</b> -tiz/  | → | [ji- <b>s</b> -tiz]   | ‘it was spun’    |

Typically, assimilation proceeds from the root to the prefixes.

In compounds, however, which contain multiple roots, sibilant harmony does not necessarily apply, meaning that such words can contain disagreeing sibilants:<sup>58</sup>

(57) Exceptions to sibilant harmony in compounds (Young and Morgan 1987)

- |                              |                |            |
|------------------------------|----------------|------------|
| (a) <b>tʃ</b> éí-            | <b>ts'</b> iin | ‘rib cage’ |
| heart                        | bone           |            |
| (b) <b>ts<sup>h</sup></b> é- | <b>tʃ</b> ééʔ  | ‘amber’    |
| stone                        | resin          |            |

In the next section, I will show that just as English compounds may violate the constraint against geminates but tend not to, compounds in Navajo, although they may violate sibilant harmony, tend to combine roots whose sibilants already agree.

<sup>57</sup> A note on transcriptions: Navajo examples are given in IPA, with acute accents marking high tones (low tones are unmarked). In order to accommodate accent marks, nasal vowels are indicated with a hook below the relevant symbol (e.g., [a̤] for IPA [ã]).

<sup>58</sup> A handful of compounds do undergo sibilant harmony, such as **tsaa-nééʒ** ‘mule’, from /tʃaa/ ‘ear’ + /nééʒ/ ‘long’ (Sapir and Hojier 1967). I suspect that these words undergo harmony because they have been stored as single units by speakers due to their semantic opacity, but I have included them in the analysis in their underlying (i.e., disagreeing) form, on the assumption that the sibilants disagreed when the compound was originally formed.

#### 4.3.1. Navajo compounds

The data described here are taken from Young and Morgan 1987, the largest existent dictionary of Navajo. From this dictionary a list of all compounds containing exactly two sibilants, each sibilant in a different root, was constructed, a total of 140 words—this represents all the words that *could* violate sibilant harmony. The effect of sibilant harmony in compounds is sensitive to distance (A. Martin 2005); the data discussed here are thus limited to the subset of these words in which the sibilants are in adjacent syllables (there were no cases in which sibilants were in the same syllable, but different roots),<sup>59</sup> a total of 97 words. Representative examples are given in (58) (roots are underlined).

(58) Examples of compounds with two sibilants in adjacent syllables (one per root)

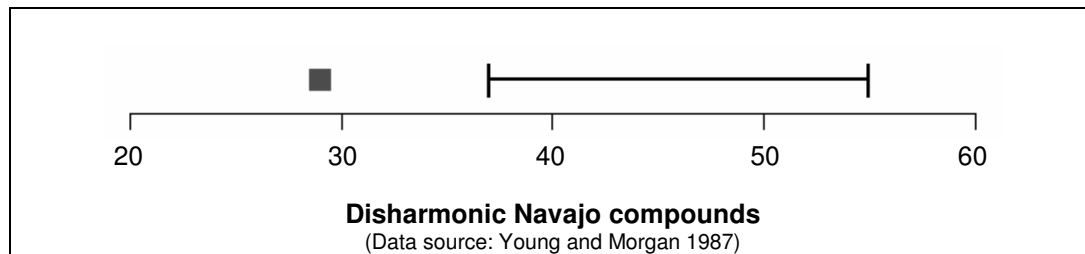
- |     |                          |                |              |
|-----|--------------------------|----------------|--------------|
| (a) | <u>ts<sup>h</sup>ee-</u> | <u>ts'iin</u>  | ‘tailbone’   |
|     | tail                     | bone           |              |
| (b) | <u>k'iif-</u>            | <u>ʒin-</u> ii | ‘blue beech’ |
|     | alder                    | black one      |              |
| (c) | <u>ts<sup>h</sup>é-</u>  | <u>zéí</u>     | ‘gravel’     |
|     | rock                     | crumbs         |              |

Of these 97 words, 29 (29.9%) contain disagreeing sibilants, violating the stem-internal phonotactic. A Monte Carlo test confirms that this is significantly below chance ( $p < 0.001$ ), as illustrated in (59).

---

<sup>59</sup> Navajo syllables are maximally CVC.

(59) Disharmonic compounds are underrepresented in Navajo



Although the phonotactic constraint in Navajo is entirely different than that in English, both languages are the same in that the stem-internal phonotactic is obeyed, albeit more weakly, across morpheme boundaries.

The gradient harmony constraint in Navajo compounds is unlikely to be the result of language-independent performance factors—due to the gradient OCP effects discussed in chapter 3, we would expect most languages to avoid sequences of similar sibilants. This is the case, for example, in English; /s..f/ sequences are *over*represented compared to /s..s/ sequences, the opposite of the pattern seen in Navajo compounds (Berkley 2000). From this I conclude that the pattern in Navajo is learned, and should be modeled as part of Navajo speakers’ grammatical competence.

#### 4.4. Turkish vowel harmony

In Turkish, just as in English and Navajo, compound words are biased towards obeying a phonotactic that holds within stems. In the case of Turkish, the phonotactic is vowel harmony. Vowels in a Turkish word must agree in backness; this requirement can cause backness features to spread from the root onto following

suffixes (Lewis 1967). The chart in (60) lists the vowels of Turkish; each vowel is given in Turkish orthography, accompanied by the IPA equivalent in brackets.

(60) Turkish vowel system

Front		Back	
i [i]	ü [y]	ı [ɯ]	u [u]
e [e]	ö [ø]		o [o]
		a [a]	

Unlike the English and Navajo cases, the vowel harmony phonotactic in Turkish stems is not exceptionless; native words with disagreeing vowels like *dahi* ‘also’ are attested (Clements and Sezer 1982). Harrison et al. (2002) report that fully 25% of Turkish stems (including borrowings) are disharmonic. Despite these exceptions, however, the phonotactic is still somewhat productive—although disharmonic loanwords are tolerated, like Fr. *microbe* > T. *mikrop*, some are repaired, for example It. *medaglia* ‘medal’ > T. *madalya*, or Ar. *mumkin* ‘possible’ > T. *mümkün* (Lewis 1967). Below I will show that this phonotactic, despite its gradient nature, influences the formation of compounds.

#### 4.4.1. Turkish compounds

Noun-noun compounds in Turkish can be divided into two types—*izafet* compounds, and “single-word” constructions (Lewis 1967, Birtürk and Fong 2001). *Izafet* compounds involve a concatenation of two nouns followed by the third-person singular possessive suffix, as in the following examples:<sup>60</sup>

<sup>60</sup> I refer here only to what are called indefinite *izafet*; I do not discuss definite *izafet* constructions, which involve a genitive suffix on the first member and are much closer to syntactic phrases.

(61) *Izafet* compounds<sup>61</sup>

(a) baş + ağrı + sı → başağrısı ‘headache’  
head pain POSS

(b) balık + ağ + ı → balıkbağı ‘fishing net’  
fish net POSS

(c) deniz + kız + ı → denizkızı ‘mermaid’  
sea girl POSS

*Izafet* compounding is highly productive and tends to be semantically transparent

(Birtürk and Fong 2001).

Single-word compounds also concatenate two nouns, but lack the possessive suffix, as shown in (62).

(62) Single-word compounds

(a) baş + bakan → başbakan ‘prime minister’  
head minister

(b) orta + okul → ortaokul ‘middle school’  
middle school

(c) ön + ayak → önyak ‘pioneer’  
front foot

Single-word compounding is not productive, according to Birtürk and Fong (2001), and single-word compounds often have non-compositional meaning. Note also that vowel harmony does not apply within compounds; in (62c), front and back vowels are permitted to coexist in the same word.

I compiled a list of single-word compounds by the following procedure: I began with a list of nouns taken from a large machine-readable list of Turkish words originally designed for use by spell-checking software (Solak and Oflazer 1993). I

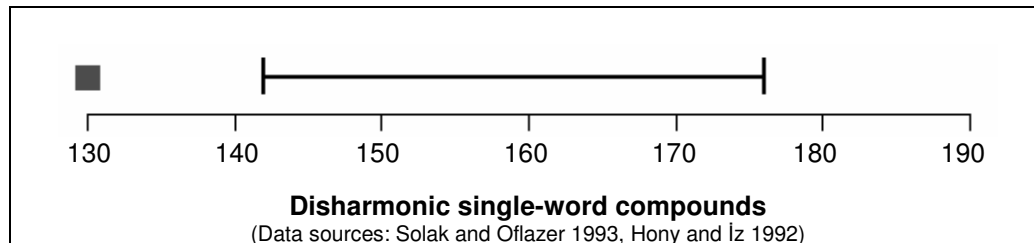
---

<sup>61</sup> Turkish words are given in native orthography; ⟨i⟩ represents a back high unrounded vowel (IPA [ɯ]), and ⟨ö⟩ and ⟨ü⟩ front rounded vowels (IPA [ø] and [y] respectively).

then extracted all potential noun-noun compounds from this list; that is, all words that could be parsed as a concatenation of two other nouns. I then looked up each potential compound in the Oxford Turkish-English dictionary (Hony and İz 1992) to determine whether it was in fact a compound (the dictionary uses a diacritic to represent compound boundaries). This resulted in a total of 326 compounds.

Because some of the stems used in the compounds are disharmonic, I examined only the final vowel of the first member and the initial vowel of the second member, on the assumption that any phonotactic effect would be strongest in the closest pair of vowels. Out of the 326 compounds, only 130 (39.9%) violate the harmony constraint across the boundary. The chart in (63) shows that this is significantly below chance ( $p < .001$ ).

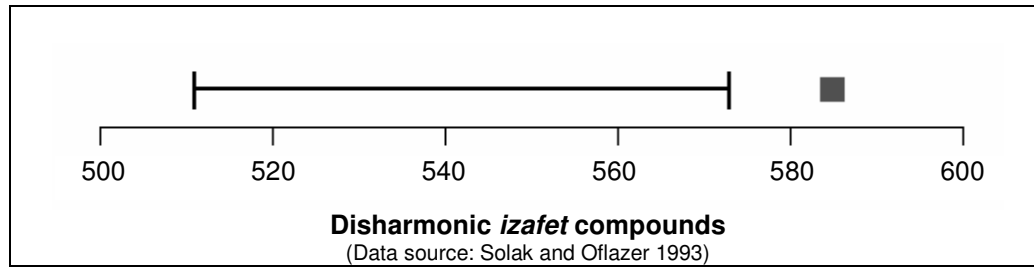
(63) Disharmonic stems are underrepresented in Turkish single-word compounds



Thus, despite the fact that the stem-internal vowel harmony phonotactic is itself gradient in Turkish, speakers prefer not to violate it when forming compounds (although the effect is weaker across morpheme boundaries).

A different pattern is found in the more productive *izafet* compounding process. The chart in (64) shows that in a list of 1,121 *izafet* compounds (taken from Solak and Oflazer's (1993) spell-checking word list), disharmonic compounds are *overrepresented*.

(64) Disharmonic stems are overrepresented in Turkish *izafet* compounds



It is not surprising that *izafet* compounds fail to exhibit the same pattern of disharmony avoidance as single-word compounds—given that *izafet* compounding is highly productive and semantically transparent, such compounds are unlikely to be stored in the lexicon. Single-word compounds may, in fact, be *izafet* compounds that have been lexicalized over time, losing their possessive marker and semantic transparency. This would explain the overrepresentation of disharmonic vowel sequences in current *izafet* compounds, if we think of the space of possible compounds as being divided between the two compound types.<sup>62</sup> Because disharmonic stem combinations tend not to be lexicalized as single-word compounds, speakers wishing to express these combinations must form *izafet* compounds, leading to the overrepresentation of these compounds in corpus data.

#### 4.5. Discussion

The phonotactics in English, Navajo, and Turkish discussed above all obey the same generalization: some phonotactic constraint holds within morphemes, and a

<sup>62</sup> This assumes that there is a blocking effect; i.e., speakers do not tend to form a *izafet* compound from two stems if there is already an available single-word compound formed from the same stems. This seems plausible, and is supported in my data by the fact that there is no overlap between the two lists—no combination of stems appears in both the single-word and *izafet* lists.

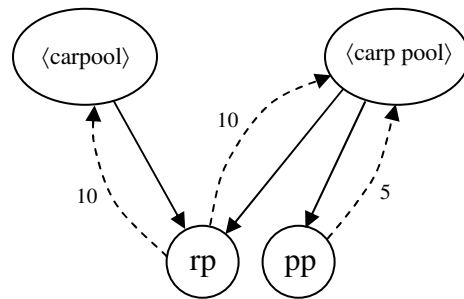


weaker version of the same constraint holds across morpheme boundaries. Morpheme boundaries, in other words, are like semi-permeable membranes: they license violations of phonotactic constraints, but only up to a point. In terms of the model developed so far in this dissertation, this means that complex words that obey stem-internal phonotactics have more of an advantage in competitions with rival synonyms than words that violate phonotactics.

One way to model the English case that will not work would be to posit individual nodes that represent consonant clusters—the lexical node for *carpool*, for example, would be connected to the node for /rp/, while *carp pool* would be linked additionally to the /pp/ node. The tautomorphemically legal /rp/ node, because it is connected to lexical nodes for both monomorphemes and multimorphemic words, will be connected to more words than the node for illegal /pp/. Words containing /rp/ will thus receive more feedback from that node than words containing /pp/.

This simple extension of the model, however, will not work, because *carp pool* contains *both* /rp/ and /pp/—it contains a superset of the consonant clusters present in *carpool*, and so in a spreading activation model should always receive more activation, and thus be fitter, than *carpool*. This is shown in (65), in which the amount of feedback sent from the cluster nodes to the lexical nodes is represented by numbers next to the dashed arcs; in this example, /pp/ sends back a smaller amount of activation because of the fewer words it is connected to.

(65) Network with consonant cluster nodes



If a node's activation is simply the sum of all inputs, *carp pool* (with 15 units of incoming activation) should have an advantage over *carpool* (10 units).

To avoid this problem, we could change how activation spreads in the model. If, for example, a node's activation is determined by averaging the activations from neighboring nodes, instead of summing them, then a low-probability consonant cluster (i.e., a node connected to few words) could penalize a word containing it by lowering the average incoming activation to that word's lexical node. In the example in (65), *carp pool* has an average incoming activation of 7.5 (the average of 10 and 5), while *carpool* has an average of 10.

This model has at least two major problems. First, it would not be able to account for the ill-formedness of monomorphemes with geminates. The hypothetical monomorphemic word *carppool*, with a geminate /p/, would receive just as much activation as the compound *carppool*. In reality, it is extremely unlikely that native English speakers would accept *carppool* as a word unless it were parsed into two units, *carp* and *pool*. The system simply cannot encode the effects of morphology on phonotactics.

We could attempt to get around this by augmenting the nodes to refer to morpheme boundaries. In this model, *carp pool* would be linked to an /rp/ node and a /p+p/ node, and monomorpheme *carppool* would only be linked to an /rp/ node—there would be no /pp/ node due to the absence of words containing this sequence morpheme-internally. If a node's activation is determined by averaging its inputs, however, this would predict that *carppool*, with the morpheme-internal geminate, would have higher fitness than *carp pool*. The only way to get this model to work would be to posit a /pp/ node with negative weight, which strongly inhibited words containing this sequence. It is unclear, however, why the system responsible for speech planning would require nodes for nonattested sequences, and I am aware of no independent evidence (from speech errors, for example) for the existence of such nodes in the speech production system.

The second problem involves the nature of the representations encoded by the nodes at the phonological level. In the Navajo and Turkish cases, the network would need nodes for very abstract properties: nodes representing whether two sibilants in the same word agree or not in Navajo and similar nodes for vowel backness agreement in Turkish. Even ignoring the objection that the speech production system would derive no obvious benefit from encoding abstract, long-distance relationships like this, there is simply no evidence from speech errors that the system makes use of such nodes.

These generalizations, then, must be encoded somewhere outside of the system responsible for directly planning and encoding speech. The obvious location

for these generalizations is a grammar, a mechanism which stores generalizations extracted from the lexicon during the course of language learning. The grammar in turn influences competitions among lexical items. In the remainder of this chapter, I will describe a phonotactic learner that constructs such grammars, and argue that the phonotactic effects in English, Navajo, and Turkish are the result of a tension in this learner between pressures for simplicity and accuracy.

#### **4.6. The phonotactic learner**

As befits one of the core topics of phonological theory, there is an extensive literature on the learning of phonotactic generalizations, accompanied by a wide range of proposed learning algorithms (e.g., Boersma 1997, Prince and Tesar 1999, Hayes 2000, Pater 2005, Heinz 2007, Pater et al. 2007). Because of the gradient nature of the phonotactic preference in complex words, I will be concerned here with the subset of possible learning algorithms that can learn statistical generalizations. The goal will be to construct a learner which, when given a list of words as input, outputs a grammar which assigns probabilities to all possible words based on the properties of the input data. In the case of English,<sup>63</sup> we want the final grammar to assign very low probabilities to words containing stem-internal geminates, high probabilities to words containing only legal clusters, and intermediate probabilities to words containing geminates across morpheme boundaries.

---

<sup>63</sup> Throughout the rest of the chapter, I will use the English case to illustrate how the theory works; the Navajo and Turkish cases can be assumed to be analyzable in the same way.

Some learners of this type generalize beyond the training data by directly assigning probabilities not to whole words, but to subparts of words. The grammar can then compute the probabilities to completely novel words, as long as they are composed of subparts that are attested in the training data, by computing the joint probability of their subparts. Words containing high-frequency subparts will be rated as more probable than words containing low-frequency subparts. Where these learners differ is in how they analyze words into subparts—whether frequencies are computed over phonemes, strings of phonemes, syllables, bundles of features, or some combination of these.

The simplest possible learner of this type is an *n-gram learner*. This class of learners, when presented with data in the form of strings of symbols, simply tabulates the frequency of each substring of length  $n$  that occurs in the data (Jurasfky and Martin 2000). A novel word is given a probability equivalent to the joint probability of all of its  $n$ -grams. A unigram learner over phonemes, for example, would simply count the frequency of all the phonemes in the training data—the word /kabe/ would have a probability  $P(/kabe/) = P(/k/) \times P(/a/) \times P(/b/) \times P(/e/)$ . A bigram learner would assign the same word a probability  $P(/kabe/) = P(/#k/) \times P(/ka/) \times P(/ab/) \times P(/be/) \times P(/e#/)$ .

Could an  $n$ -gram learner learn the English phonotactics? Let us consider how such a learner would handle a simple, toy language designed to emulate the English facts. This language, which I will call the *p-t language*, has only two segments, [p] and [t]. All words in this language consist of a string of two segments which may or

may not have an intervening morpheme boundary (indicated by a “+”). All of the logically possible words in this language are listed in (66).

(66) Logically possible words in the p-t language

p <sub>1</sub> t	t <sub>1</sub> p	p <sub>1</sub> p	t <sub>1</sub> t
p <sub>2</sub> +t	t <sub>2</sub> +p	p <sub>2</sub> +p	t <sub>2</sub> +t

“Words” without a morpheme boundary are considered to be morpheme-internal clusters (words written without the “+” symbol therefore do not contain a boundary). Because this language is designed to resemble English, I will stipulate that the words pp and tt are unattested in this language. The set of actual words in the p-t language is given in (67).

(67) Attested words in the p-t language

p <sub>1</sub> t	t <sub>1</sub> p		
p <sub>2</sub> +t	t <sub>2</sub> +p	p <sub>2</sub> +p	t <sub>2</sub> +t

Let us now construct a lexicon for the p-t language. The first lexicon I consider will resemble English in that “compounds” with geminates are legal, but underrepresented—I will call this the *biased lexicon*. Because of the small number of possible words, I will represent type frequencies as token frequencies—if the lexicon contains 2,000 tokens of p<sub>1</sub>t, it should be interpreted to mean that there are 2,000 words with morpheme-internal non-geminate clusters, not that there are 2,000 tokens of the same word (another way to think of it is that the data described here consists only of the word-medial consonant clusters taken from all the words in the lexicon). The structure of the biased lexicon is shown in (68).

(68) Biased lexicon for p-t language

Word	Type	Number
pt	monomorpheme	1,000
tp	monomorpheme	1,000
pp	monomorpheme	0
tt	monomorpheme	0
p+t	compound	1,000
t+p	compound	1,000
p+p	compound	800
t+t	compound	800

The frequencies have been chosen to very roughly mirror the three-way wellformedness distinction apparent in the English lexicon: compounds without geminates are more frequent than compounds with geminates, which are much more frequent than monomorphemes with geminates.

A trigram learner confronted with this lexicon (assuming the morpheme boundary is simply treated as a symbol) would assign frequencies commensurate with the frequencies in (68), and would assign probabilities to all eight logically possible words as in (69).

(69) Probabilities assigned to biased lexicon

Word	Type	P(word)
pt	monomorpheme	0.18
tp	monomorpheme	0.18
pp	monomorpheme	0.00
tt	monomorpheme	0.00
p+t	compound	0.18
t+p	compound	0.18
p+p	compound	0.14
t+t	compound	0.14

This simple statistical learner, not surprisingly, correctly learns the pattern in the lexicon. If we assume that these probabilities bias lexical competitions, then

compounds with geminates will be less fit than those with legal clusters, and the result will be an underrepresentation of geminates in the lexicon.

The problem with this learner as a model of what humans do is that it is too good at tracking frequencies. As I showed above, a simple trigram learner can learn a lexical bias against geminates, but it could learn the lack of a bias just as well. The chart in (70) shows that if the lexicon is unbiased with respect to heteromorphemic geminates, the resultant grammar will also be unbiased.

(70) Probabilities assigned to unbiased lexicon

<b>Word</b>	<b>Type</b>	<b>Number</b>	<b>P(word)</b>
pt	monomorpheme	1,000	0.167
tp	monomorpheme	1,000	0.167
pp	monomorpheme	0	0.00
tt	monomorpheme	0	0.00
p+t	compound	1,000	0.167
t+p	compound	1,000	0.167
p+p	compound	1,000	0.167
t+t	compound	1,000	0.167

If the learner is able to learn any kind of lexicon, why do actual lexicons tend to be biased? This simple learner fail to explain the apparent connection between tautomorphemic and heteromorphemic phonotactics observed in the English, Navajo, and Turkish cases. Some additional mechanism would have to be invoked to account for the fact that weaker versions of stem-internal phonotactics occur heteromorphemically.

This is because the learner computes tautomorphemic and heteromorphemic phonotactics independently. As far as the trigram learner is concerned, a pp sequence and a p+p sequence are entirely unrelated; a given cluster can only be counted as one

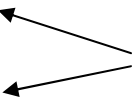


or the other, not both. I will call generalizations like these, which take morphological structure into account, *structure-sensitive*. In order to model the correlation between the two, I will argue that learners also form *structure-blind* generalizations, which ignore morphological structure. If, for example, the learner were to compare the frequencies of p (+) t and p (+) p (i.e., clusters regardless of whether they have an intervening boundary), it would find that geminates are less frequent than non-geminate clusters, even in an unbiased lexicon, simply as a result of the fact that geminates cannot occur within morphemes (see (71)).

(71) Structure-sensitive vs. structure-blind frequency counts: unbiased lexicon

	Word	Number
structure-sensitive	p t	1,000
	t p	1,000
	p p	0
	t t	0
	p+t	1,000
	t+p	1,000
	p+p	1,000
	t+t	1,000
structure-blind	p (+) t	2,000
	t (+) p	2,000
	p (+) p	1,000
	t (+) t	1,000

apparent bias if morphological structure is ignored



The learner I will propose makes use of both types of generalization, combining them in a single grammar. Because the probabilities of structure-blind and structure-sensitive sequences are not independent, however, the grammar cannot simply multiply the probabilities of a word's sequences to obtain the probability of the word. In order to deal with the probabilities of overlapping categories, a more

sophisticated mathematical approach is required. In the next section I show that the Maximum Entropy framework will allow us to combine structure-blind and structure-sensitive generalizations, and that with a grammar of this type, the phonotactic biases described in English, Navajo, and Turkish emerge as a natural consequence, even from a beginning state with no such bias.

#### **4.7. Maximum Entropy**

The Maximum Entropy (“maxent”) formalism has long been a staple of the machine learning literature, and has recently been applied to problems of phonological learning (Berger et al. 1996, Della Pietra et al. 1997, Goldwater and Johnson 2003, Jäger 2004, Wilson 2006, Hayes and Wilson to appear). A maxent learning algorithm learns a probability distribution over the members of some set given a sample drawn from that distribution. Crucially for our purposes, the algorithm has a principled way of calculating probabilities involving overlapping categories.

A maxent grammar consists of a set of numerically weighted constraints. These constraints ban structures in the output (e.g., \*pp “no geminates within a morpheme”), and are equivalent to the markedness constraints used in Optimality Theory (Prince and Smolensky 1993/2004). The grammar defines a probability distribution over constraint violations, so that the set of constraints determines the units whose frequency is counted—if the algorithm is given only constraints against n-grams, for example, it will behave as an n-gram learner.

Each constraint has a weight, represented by a nonzero real number, which represents the “strength” of the relevant constraint. The set of constraints and their weights (which together constitute the grammar) determine a probability for every possible candidate output, which is a function of the set of constraints violated by the output and their weights. Specifically, a word’s probability is a function of its *score*<sup>64</sup>  $\Phi$ , which is calculated by simply summing the (weight  $\times$  number of violations) for every constraint in the grammar, as shown in (72).

(72) Definition of score

$$\phi(x) = \sum_{i=1}^M w_i C_i(x)$$

where

$M$ : number of constraints

$w_1, w_2, \dots, w_M$ : constraint weights

$x$ : representation of candidate

$C_i(x)$ : number of violations assigned to  $x$  by constraint  $C_i$

For a grammar with three constraints,  $C_1$  (weight 1.0),  $C_2$  (weight 2.0), and  $C_3$  (weight 3.0), an output form  $x$  violating  $C_1$  twice and  $C_3$  once would be assigned a score of  $\Phi(x) = (1.0 \times 2) + (2.0 \times 0) + (3.0 \times 1) = 2.0 + 0 + 3.0 = 5.0$ .

A word’s probability is an exponential function of its score; it is given by the equation in (73), where  $Z$  is a normalizing term defined as the sum of  $e^{-\Phi(x)}$  for all possible words.

(73) Determining candidate probability

$$P(x) = \frac{e^{-\phi(x)}}{Z}$$

---

<sup>64</sup> This terminology, as well as the specific implementation of the maximum entropy learning algorithm discussed here, is taken from Hayes and Wilson (to appear).

Throughout the rest of this chapter, because I will be comparing words for which the denominator in (73) is the same, I will simply refer to the value of the numerator, which Hayes and Wilson (to appear) call the *maxent harmony* ( $h(x) = e^{-\Phi(x)}$ ), rather than the actual probability of a given word. This value represents the share of the total probability apportioned to a given word.

Given a set of constraints and a set of training data, the learning algorithm adjusts the constraint weights so as to maximize the probability of the data—the algorithm thus represents an example of *maximum likelihood* learning. The probability of the data is calculated by simply multiplying the probabilities of all of the words in the data to arrive at their joint probability. The learner maximizes the log of this probability, which is equivalently stated as the sum of the log probabilities of each word, as in (74).

(74) Probability of training data

$$\sum_{i=1}^N \log P(x_i)$$

If the algorithm is simply asked to maximize this function, however, there is a danger of overfitting the data. Because the learner is given a finite sample of data drawn from an infinite language, a pure maximum likelihood learner will tend to overestimate the probability of items that are in the sample, and underestimate the probability of items that didn't happen to occur in the sample (many low-probability items, for example, will not occur and thus be assigned probabilities of zero). In other words, the probability distribution learned from finite sample will be too strongly skewed in the direction of the observed data.

The standard way to avoid overfitting is to introduce a smoothing term into the learning function (S.C. Martin et al. 1999). The smoothing term penalizes skewed distributions and causes the learner to favor more uniform distributions, which ameliorates the tendency to overfit. Many smoothing methods have been developed for maxent learning algorithms; I will use a Gaussian prior over the constraint weights (see Chen and Rosenfeld 2000 for arguments in favor of this smoothing technique compared to others). The prior term is subtracted from the likelihood term, resulting in the learning function in (75).

(75) Maxent learning function

$$\sum_{i=1}^N \log P(x_i) - \sum_{j=1}^M \frac{(w_j - \mu_j)^2}{2\sigma_j^2}$$

The Gaussian prior assesses a penalty for constraint weights that deviate from their ideal weights, represented by  $\mu_j$ . In the implementation of the algorithm I will use,  $\mu$  is set to zero for all constraints, so that the prior penalizes any nonzero weight, with the size of the penalty increasing with the square of the weight. This pressure towards low constraint weights translates in a bias against highly skewed distributions—because the prior term increases with the square of each constraint weight, it prefers grammars with many low-weighted constraints over grammars with a few high-weighted constraints. This means that if multiple constraints are each capable of explaining a given property of the data, the learner will assign all of the constraints low weights rather than choose one and assign it a high weight. This property of the prior will prove crucial in modeling the English, Navajo, and Turkish data.

The learning function thus embodies a trade-off between a pressure to model the data as accurately as possible (the likelihood term) and a pressure to have as simple (i.e., uniform) a grammar as possible. The value of the free parameter  $\sigma^2$  determines the relative importance of each of these factors. As we will see in the next section, modeling the connection between tautomorphemic and heteromorphemic phonotactics will rely crucially on this trade-off.

#### **4.8. Testing the maxent learner on the p-t language**

The phonotactic “leakage” seen in English, Navajo, and Turkish can be modeled as the effects of two crucial components: the existence of structure-blind constraints in the grammar, and a bias against high constraint weights (represented in the maxent learner as a Gaussian prior). Below I will first show that given these two components, the maxent learner learns a bias against violating stem-internal phonotactics even from an unbiased lexicon. Then I will demonstrate that both components are necessary by showing that the learner fails to learn a bias in the absence of either.

The training data I will use for all the demonstrations of the learner is given in (76). The numbers of each word type were chosen so that there would be an equal number of monomorphemes and compounds, and an equal number of compounds with geminates and compounds without geminates.

(76) Training data (unbiased)

Cluster	Structure	Number of examples
p t	monomorpheme	2,000
t p	monomorpheme	2,000
p+t	compound	1,000
t+p	compound	1,000
p+p	compound	1,000
t+t	compound	1,000

The constraints the learner will start with are given in (77). Note that a plus sign in parentheses indicates an optional morpheme boundary, while the absence of a plus sign (as in \*pp) indicates that no morpheme boundary intervenes between the consonants.

(77) Constraints

**Structure-blind**

- \*p(+)p no geminates
- \*t(+)p no non-geminate consonant clusters

**Structure-sensitive**

- \*pp no geminates within a morpheme
- \*tp no non-geminate consonant clusters within a morpheme
- \*p+p no geminates across a morpheme boundary
- \*t+p no non-geminate clusters across a morpheme boundary

When the maxent learning algorithm (with the smoothing term) is given these constraints, and exposed to the data in (76), it arrives at the grammar in (78).

(78) Final grammar, all constraints

Constraint	Weight
*p(+)p	0.04
*t(+)p	0.00
*pp	4.01
*tp	0.13
*p+p	0.00
*t+p	0.00

The algorithm assigns a high weight to \*pp, which is unsurprising due to the lack of pp sequences in the training data. More surprising is the fact that \*p(+)p also receives a small but nonzero weight. This is the effect of the prior term in the learning function, which is optimized by making the distribution of weights as uniform as possible.

Assigning a weight to \*p(+)p lowers the probability of pp sequences, which allows the weight on \*pp to be lower. The price of this more uniform distribution is accuracy in modeling the data—the weight on \*p(+)p also lowers the probability of p+p sequences.<sup>65</sup>

The table in (79) shows that compounds with geminates (p+p) are evaluated as less probable (i.e., have a lower maxent harmony) than compounds without geminates (t+p).

---

<sup>65</sup> Note that the learner also assigns a nonzero weight to \*tp, despite such words being plentiful in the training data. This is a byproduct of the simplified nature of the example—the software that implements the learning algorithm is designed to expect words up to a maximum length equal to the longest word encountered in the training data. Because words of length 3 are encountered (e.g., t+p; the morpheme boundary is counted as a symbol), the learner expects to see all possible words of this length, including words like tpt or ppp. Since these words do not occur, the learner slightly increases the weights on both \*pp and \*tp (the weight on \*tp cannot be increased very much, because words like tp are attested). Because this is an artifact of the way the algorithm has been implemented, and not a fundamental property of the algorithm itself, the weight given to \*tp can be safely ignored.



(79) Example outputs as evaluated by grammar, all constraints

<b>x</b> (potential output)	<i>constraints</i>				<b>Φ(x)</b> (score)	<b>h(x)</b> (maxent harmony)
	<b>*pp</b> (w 4.01)	<b>*tp</b> (w 0.13)	<b>*p(+)p</b> (w 0.03)	<b>*t(+)p</b> (w 0)		
(a) pp	4.01		0.03		4.04	<b>0.02</b>
(b) tp		0.13		0	0.13	<b>0.87</b>
(c) p+p			0.03		0.04	<b>0.96</b>
(d) t+p				0	0	<b>1.00</b>

This bias against compounds with geminates disappears if the structure-blind constraints are removed from the constraint set, and the learner constructs a grammar using only structure-sensitive constraints when given the same training data. The results of this structure-sensitive-only learning are shown in (80).

(80) Final grammar, structure-sensitive constraints only

<b>Constraint</b>	<b>Weight</b>
*pp	4.02
*tp	0.13
*p+p	0.00
*t+p	0.00

Note that the weight for \*pp is 4.02, as compared to the weight of 4.01 that was assigned to the same constraint by the learner using the structure-blind constraints. This shows that a constraint's weight is dependent not just on the properties of the data, but on the other constraints that are present in the grammar. In this case, \*pp gets a higher weight when there is no other constraint that could also explain the absence of pp in the data. When the structure-blind constraints are included in the grammar, this generalization is split between two constraints, \*pp and \*p(+)p, which allows the weight on \*pp to be slightly lower.

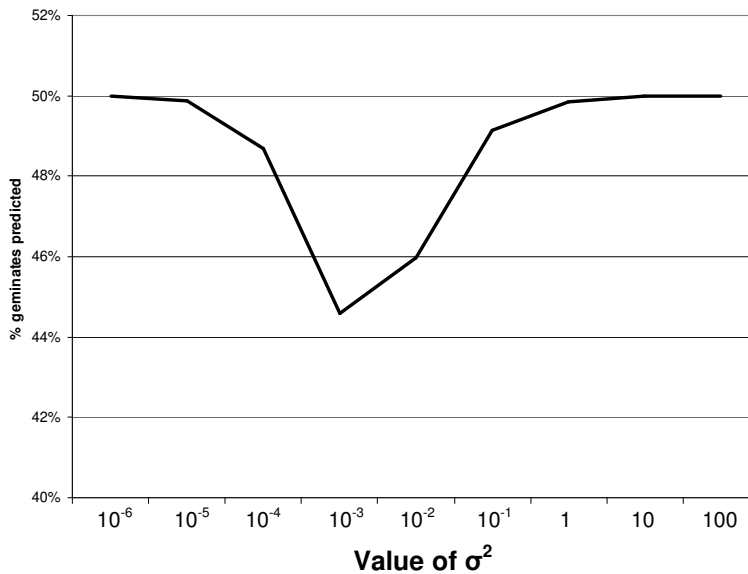
The table in (81) shows that when the structure-blind constraints are removed, the grammar learned by the algorithm evaluates compounds with and without geminates as equally probable.

(81) Example outputs as evaluated by grammar, structure-sensitive constraints only

<b>x</b> (potential output)	<i>constraints</i>				<b><math>\Phi(\mathbf{x})</math></b> (score)	<b>h(x)</b> (maxent harmony)
	<b>*pp</b> (w 4.02)	<b>*tp</b> (w 0.12)	<b>*p+p</b> (w 0)	<b>*t+p</b> (w 0)		
(a) pp	4.02				4.02	<b>0.02</b>
(b) tp		0.12			0.12	<b>0.89</b>
(c) p+p			0		0	<b>1.00</b>
(d) t+p				0	0	<b>1.00</b>

The bias also disappears if the structure-blind constraints are included, but the  $\sigma^2$  parameter is increased, making the prior less important. The graph in (82) shows how the predicted ratio of geminates to non-geminates in compounds changes as a function of  $\sigma^2$ . This chart represents grammars containing both structure-sensitive and structure-blind constraints.

(82) Effect of  $\sigma^2$  on geminate ratio in compounds



For very low values of  $\sigma^2$ , the prior is so strong that the weights of all the structure-blind constraints are forced to zero, and no bias is apparent; when  $\sigma^2$  is very high, the prior is essentially turned off, and again the learner has no bias against geminates in compounds. For values in the middle, however, the prior is strong enough to put weights onto the structure-blind constraints against geminates, but not so strong that the distribution of weights is completely flat, and the result is a bias against geminates in compounds. Crucially, for no value of  $\sigma^2$  is there a reverse bias in which geminates are preferred to non-geminates.

For human learners, of course, the value of  $\sigma^2$  has presumably been set by natural selection. Values that are too low would result in learners that are incapable of learning, and simply prefer a uniform distribution no matter what the data looks like. Values that are too high would result in overfitting, and a failure to generalize beyond the specific data the learner is exposed to. It is likely that the human phonotactic learner represents an optimal compromise between these two extremes—a learner that can generalize, but is still able to acquire language-specific patterns. The cost of this compromise is a learner for whom phonotactics in different domains are entangled rather than independent.

I have shown that if structure-blind constraints are present, the learner will automatically be biased against sequences that do not occur within morphemes. But why would speakers make use of structure-blind constraints? One possibility is that learners come with an innate bias for simpler constraints. Just as the Gaussian prior

causes the learner to prefer simpler grammars, this bias could cause learners to posit constraints that are maximally simple, even if those constraints lead to a less accurate characterization of the data.

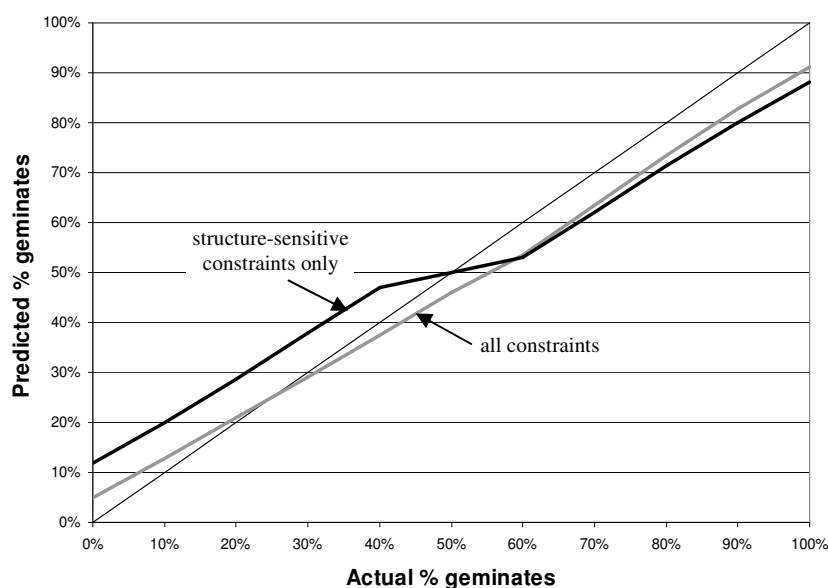
Another possible explanation is that structure-blind constraints represent a holdover from an early stage of phonotactic learning. There is substantial evidence for the fact that children learn a fair amount about the phonotactic patterns in their language before they are able to parse the speech stream into morphemes or even words (Peters 1983, Jusczyk 1997). Generalizations formed at this stage are by necessity structure-blind, and may themselves be used to discover morphological structure. Of course, once they master morphology, children are able to construct structure-sensitive constraints, but it is plausible that the structure-blind constraints they used at the earlier stage remain in the grammar, and make their presence felt, in the form of phonotactic preferences, into adulthood.

#### **4.9. The nature of the learning bias**

I have shown that given an unbiased lexicon, the maxent learner, equipped with a bias against constraint weights and structure-blind constraints, automatically learns a bias against complex words that violate stem-internal phonotactics. In this section I look at how the learner performs on biased lexicons—if geminates are *over*represented in compounds, for example, will the learner still construct a grammar that is biased against them?

The graph in (83) compares the learner with structure-blind constraints to the learner with only structure-sensitive constraints on a range of training data. In each case, the number of compounds in the data was held constant at 4,000, but the ratio of geminates to non-geminates in compounds was varied from all geminates to all non-geminates. The  $x$ -axis in (83) represents the percentage of compounds that contained geminates; for example, the 30% geminates data set consisted of 1,200 compounds with geminates and 2,800 compounds without geminates (all of the training data sets included 4,000 monomorphemes, none of which contained geminates). The  $y$ -axis represents the ratio of geminates predicted by the final grammar. For comparison, a hypothetical perfectly accurate learner that always predicts the exact ratio of geminates that is present in the training data is represented by the thin line that passes through the origin.

(83) The learning bias with different lexicons<sup>66</sup>



The graph reveals that when only structure-sensitive constraints are used (the black line), the only learning bias is the one introduced by the prior. It produces a predicted ratio that is closer to uniform than is supported by the data—in other words, it underpredicts the number of words of whichever type is overrepresented in the data. The learner that also incorporates structure-blind constraints (the gray line) also has this bias, but also has a slight additional bias towards underrepresenting geminates.

The bias towards extending stem-internal phonotactics to apply across morpheme boundaries is therefore a “soft” bias—the learner can acquire a grammar with the opposite bias, if given data that is sufficiently biased in that way, but *ceteris paribus* it will prefer grammars that underpredict the number of words that violate stem phonotactics.

<sup>66</sup> The free parameter  $\sigma^2$  is set to 0.01 for the learning depicted in this chart, so that the differences between the two learners are large enough to be clearly visible. Larger values of  $\sigma^2$  result in differences that are quantitatively smaller but qualitatively identical (until  $\sigma^2$  gets very large, at which point the learners behave identically).

#### **4.10. Consequences for the model**

What does this mean for my model of phonotactic preferences? The output of the phonotactic grammar could be used to influence a word's fitness by simply making each lexical node's weight (which, recall, acts as a multiplier of any incoming activation) partially a function of the probability output by the grammar for the word represented by that node, in addition to other factors such as token frequency or sociolinguistic associations. Words with higher probabilities, like compounds without geminates, would overall be more successful than words with lower probabilities, like compounds with geminates.

In terms of historical change, this means that if learners are using structure-blind constraints, a lexicon with no bias in compounds will not be stable. Such learners would prefer compounds without geminates (to use English as an example), and over time compounds with geminates would come to be underrepresented. As I showed in chapter 1, the eventual stable ratio of geminates to non-geminates that the system settles on is a function of how large a role phonotactics play in lexical selection as compared to other factors—the more emphasis speakers put on phonotactics, the more the lexicon will be dominated by the fittest forms.

## 5. Conclusion

### 5.1. Summary of findings

In this dissertation I have striven to paint a picture of the lexicon that is more nuanced than the traditional view that sees the lexicon as little more than a “trash heap”—a repository of unpredictable facts that the language learner has no choice but to simply memorize. The lexicon is the result of unconscious choices made by generations of speakers and listeners, and to the extent that these choices are biased, the lexicon itself will be biased. I have argued that these biases, which I call phonotactic preferences, can and do skew the lexicon through the adoption or retention of words, even in the absence of sound change.

In chapter 2, I presented evidence to support Boersma’s (1998) conjecture that phonotactic preferences can be based on articulatory ease. In early Latin, a sound change turned a highly marked Proto-Indo-European sound (possibly /p’/) into the less-marked /b/, resulting in a cross-linguistically unusual distribution in which /b/ was less frequent than /d/. Over time, words beginning with /b/ were more likely to be formed and retained, resulting in /b/-initial words eventually coming to outnumber /d/-initial words in the modern Romance languages. This demonstrates both that sound change is not the only mechanism that shapes the phonological makeup of the lexicon, and that phonotactic preferences can be driven by factors such as articulatory ease, and not solely by a language’s current lexical statistics.

In chapter 3, I showed that a long-distance consonant cooccurrence restriction in English also acts to bias the creation and survival of words. English words



containing two identical liquids, such as *rare* or *lull*, are underrepresented when compared to words with different liquids, such as *lair* or *real*. Using data on novel words in the *Oxford English Dictionary* and trends in American baby names, I showed that this represents a true phonotactic preference (as opposed to diachronic sound change or a phonological process of dissimilation), and argued that this preference is motivated by an avoidance of sequences that are difficult to process.

Finally, in chapter 4 I demonstrated that tautomorphemic phonotactic restrictions are accompanied by weaker, gradient versions of the same restrictions that hold across morpheme boundaries, using data from the unrelated languages English, Navajo, and Turkish. In Navajo, for example, sibilants that are separated by a morpheme boundary in compounds gradiently obey a sibilant harmony constraint that holds categorically within stems. I presented a phonotactic learning algorithm from which these effects follow—learners overgeneralize due to a pressure to learn the simplest possible grammar.

## **5.2. Summary of the model**

To account for the existence and nature of phonotactic preferences, I have proposed a model of speech production in which lexical items compete with synonymous items to be produced—the result over time is a lexicon largely consisting of words whose properties make them good at winning these competitions. The model draws on existing proposals that the speech production system consists of a network of nodes, representing concepts, lexemes, and pieces of phonological

structure, in which activation spreads among connected nodes, allowing concepts to activate appropriate words, which in turn activate their constituent phonemes. In this model, competitions among lexical items can be interpreted as a race to become activated most quickly, with the first lexeme to reach a critical threshold selected and eventually produced as part of the intended utterance.

Extensive evidence from speech errors and experimental data indicates that the production network incorporates feedback from lower to higher levels (e.g., from phonemes to lexical lemmas), meaning that these competitions may be biased by a word's phonological properties—phonemes, for example, that become activated more quickly than others will boost the activation of words containing those phonemes, giving them an advantage over words containing “slower” phonemes.

Although I have referred to competitions among words throughout this dissertation, there is no need to limit the theory to individual words, or even individual lexical items. In the process of attempting to construct an utterance that communicates the speaker's intended message, the production system may well generate a wide variety of words, phrases, and clauses, all of which compete to be part of the final utterance. In such a system, a word may compete with a phrase to express the same concept. If the phrase wins often enough, for enough speakers, the word will fall out of use, and the concept will thenceforth be expressed periphrastically. When the opposite occurs (i.e., a word coming to replace a phrase), the result will be the birth of a new word, as what was once described periphrastically is now expressed with a single lexical item. Extending the model in this way could

explain how lexical biases develop even if it turns out that competitions among individual synonymous words are relatively rare.

Another potential source of bias that I have not mentioned is the process by which new words (or larger linguistic structures) are created. It is clear how existing words compete—they already have lexical entries which are integrated into the network. But where do truly novel words (or phrases) come from? I have treated this generation process as akin to mutation in biological evolution, a neutral source of variation on which natural selection works. However, in the case of language it is possible that the formation of new lexical entries is itself biased. A complete theory of phonotactic preferences will have to take these biases into consideration, and determine how they differ from the biases introduced by competition.

### **5.3. Directions for future research**

As I pointed out in chapter 1, the study of how lexical statistics change over time has played almost no role in modern linguistic theory. Because of this, a single dissertation cannot hope to answer, or even address, all of the many questions raised by the topic. This chapter discusses some of the remaining issues and problems, and how they could be further explored.

#### **5.3.1. *Data collection***

The first step in pursuing a complete theory of phonotactic preferences involves increasing the amount of available data. The great majority of the data used

in this dissertation comes from a small subset of the world's languages. In order to do the kind of detailed statistical analyses that are required to test my theory, large, annotated dictionaries and corpora are necessary. Unfortunately, such resources currently exist for only a small handful of languages, most of them European.

Before a serious theory of the diachrony and typology of lexical statistics can be begun, therefore, more data must be collected and collated, with an emphasis on providing coverage of a wide range of unrelated languages. This will of course require collaboration on a large scale, but it is the only way progress can be made on the study of lexical statistics from a typological and historical perspective.

### 5.3.2. *Correlations between historical change and processing ease*

The thesis I have argued for here makes novel predictions concerning correlations between historical lexical change and the effects of phonotactic properties on processing tasks. For example, I claim that the lexicon is shaped by the accumulation of many competitions among synonyms, and that these competitions take the form of a race during speech production, in which lexical entries that are accessed and encoded quickly have an advantage. From this it follows that sounds that are historically favored should be those that contribute to faster lexical access, which can be measured by means of speech error data or reaction times in experimental production tasks.

This means that, for example, words containing marked sounds should be more prone to lexical substitution speech errors (i.e., those in which an entire word is

replaced with another word, as in *detector* → *protector*), and should tend to be replaced by words containing less marked sounds. Apart from Harley and MacAndrew's (2001) finding that longer words are more likely to undergo a substitution error, I know of no data in the existing literature that bears on this prediction. Studies using speech error corpora or experimental error elicitation techniques could test predictions like this, and the results could be compared to the statistical patterns that are stable across languages or across time within a language.

### 5.3.3. *Experimental tests of phonotactic preferences*

The evidence I have presented for the existence of phonotactic preferences has been limited to data from lexical statistics. This kind of evidence is not only indirect, but relies on dictionaries and corpora, which are at best approximations of the actual lexicons internalized by an entire population of speakers. The theory would receive stronger support if phonotactic preferences could be demonstrated experimentally.

Such an experiment could take the form of an artificial language learning task. Subjects would be shown a series of words containing a gradient phonotactic that does not exist in their native language. They would then be given pictures of objects and asked to choose new words in the artificial language for these objects from a range of options. Their choices could be analyzed to see how they are biased by the subjects' native phonotactics and the phonotactics of the artificial language. The task could be modified to address other issues—for example, subjects could be asked to borrow some of the artificial words into their own language to look at the effects of

phonotactics on borrowing. Correlations between the results of experiments like these and actual historical trends would lend support to the theory.

#### 5.3.4. *The interaction of phonotactics and morphology*

In chapter 2, when discussing French and Latin, I discussed only unprefixated words, because of the existence of highly productive prefixes containing marked sounds. Is morphological productivity affected at all by phonotactic considerations? What determines how productive a given affix is? This question can be seen as parallel to the larger question asked in this dissertation—what makes a word successful?—and so using some of the techniques I have employed here may prove enlightening.

#### 5.3.5. *Avoidance in children*

Phonotactic preferences in adult speaker represent a form of avoidance of less well-formed words. Children are often known to exhibit a much stronger version of this, completely avoiding words which contain sounds they find difficult (Ferguson and Farwell 1975, Leonard et al. 1981, Schwartz and Leonard 1982). In my model, this could be explained by very strong markedness biases (implemented as feedback from the phoneme or articulatory level), coupled with a small lexicon. In adults, markedness may cause a word to lose out to a synonym or paraphrase, but in children a synonym may not be available, and the result is that no winner emerges and the concept simply cannot be expressed.

#### 5.3.6. *The interaction of phonotactics and sociolinguistic variables*

In chapter 1, I presented evidence that the effects of phonotactics on lexical selection are strongest when the influence of social factors is weakest. I represented these social factors in my model with a single parameter—the reality is obviously much more complex. Working out how these variables interact in determining a word's success represents a fertile area for future study, one which would benefit from collaboration between sociolinguists and generative linguists. Incorporating phonotactic preferences into existing quantitative sociolinguistic models could give them greater predictive power, and would shed light on how speakers achieve compromise among conflicting pressures when deciding what to say and how to say it.

## References

- Andersen, Henning. 2006. Synchrony, Diachrony, and Evolution. In Ole Nedergaard Thomsen (ed.), *Competing Models of Linguistic Change*. Amsterdam: John Benjamins. 59–90.
- Aulestia, Gorka. 1989. *Basque-English Dictionary*. Reno, Nevada: University of Nevada Press.
- Baayen, R.H., Piepenbrock, R., Gulikers, L. 1995. The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.
- Bailey, T. M. and U. Hahn. 2001. Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language*. 44: 568–591.
- Barabási, Albert-László and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286: 509–512.
- Baronchelli, Andrea, Maddalena Felici, Vittorio Loreto, Emanuele Caglioti and Luc Steels. 2006. *Journal of Statistical Mechanics: Theory and Experiment*.
- Baugh, Albert and Thomas Cable. 1993. *A History of the English Language*. London: Routledge.
- Bavelier, Daphne. Repetition blindness between visually different items: the case of pictures and words. *Cognition* 51: 199–236.
- Berg, Thomas. 1998. *Linguistic Structure and Change: An Explanation from Language Processing*. Oxford: Clarendon Press.
- Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22: 39–71.
- Berkley, Deborah Milam. 1994. The OCP and gradient data. *Proceedings of FLSM V. Studies in the Linguistic Sciences* 24(1/2): 59-72.
- Berkley, Deborah. 2000. Gradient OCP Effects. Ph.D. dissertation, Northwestern University.
- Birtürk, Aysenur Akyuz and Sandiway Fong. 2001. A Modular Approach to Turkish Noun Compounding: The Integration of a Finite-State Model. In *Proceedings*



*of the 6th Natural Language Processing, Pacific Rim Symposium (NLPRS2001), Tokyo, Japan.*

- Blackmore, Susan. 2000. *The Meme Machine*. Oxford: Oxford University Press.
- Blevins, Juliette. 2004. *Evolutionary Phonology*. Cambridge: Cambridge University Press.
- Blevins, Juliette. 2006. A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics* 32: 2. 117–166.
- Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.
- Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences, Amsterdam* 21: 43–58.
- Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.
- Boersma, Paul. 2007. The evolution of phonotactic distributions in the lexicon. Talk given at the Presentation Workshop on Variation, Gradience and Frequency in Phonology, Stanford University.
- Bolinger, Dwight. 1950. ‘Shivaree’ and the Phonestheme. *American Speech* 25: 135–135.
- Bosworth, Joseph and TN Toller. 1898. *An Anglo-Saxon Dictionary*. Oxford: Oxford University Press.
- Breen, Jim. 2000. A WWW Japanese Dictionary. *Japanese Studies* 20:313–317. Available at [http://www.csse.monash.edu.au/~jwb/j\\_edict.html](http://www.csse.monash.edu.au/~jwb/j_edict.html).
- Brighton, Henry, Kenny Smith and Simon Kirby. 2005. Language as an evolutionary system. *Physics of Life Reviews* 2(3): 177–226.
- Briscoe, Ted. 2002. *Language Evolution Through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.
- Broe, Michael. 1995. Specification theory and Ngbaka co-occurrence constraints. Ms., Northwestern University.
- Chen, Stanley and Ronald Rosenfeld. 2000. A Survey of Smoothing Techniques for ME models. *IEEE Transactions on Speech and Audio Processing* 8: 37–50.
- Chomsky, Noam, and Howard Lasnik. 1977. Filters and Control. *Linguistic Inquiry* 8: 425–504.

- Clark, Eve. 1993. *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.
- Clements, George. 1990. The role of the sonority cycle in core syllabification. In J. Kingston, and M. E. Beckman (eds.), *Papers in Laboratory Phonology I: between the grammar and physics of speech*. Cambridge: Cambridge University Press. 283–334.
- Clements, George and Engin Sezer. 1982. Vowel and Consonant Disharmony in Turkish. In Harry van der Hulst and Norval Smith (eds.), *The structure of phonological representations (Part II)*. Dordrecht: Foris Publications. 213–255.
- Coetzee, Andries and Joe Pater. 2005. Lexically Gradient Phonotactics in Muna and Optimality Theory. Ms., University of Massachusetts.
- Coleman, J., & Pierrehumbert, J. B. 1997. Stochastic phonological grammars and acceptability. In *Computational phonology: Third meeting of the ACL special interest group in computational phonology*. Somerset, NJ: Association for Computational Linguistics. 49–56.
- Croft, William. 2006. The relevance of an evolutionary model to historical linguistics. In Ole Nedergaard Thomsen (ed.), *Competing Models of Linguistic Change*. Amsterdam: John Benjamins. 91–132.
- Croft, William. 2000. *Explaining Language Change: An evolutionary approach*. Harlow, Essex: Longman.
- Crystal, David. 2006. *Words, Words, Words*. Oxford: Oxford University Press.
- Culpeper, Jonathan. 2005. *History of English*. New York: Routledge.
- Dankovičová, J., P. West, J. S. Coleman, and A. Slater. 1998. Phonotactic grammaticality is gradient. Poster paper presented at the 6<sup>th</sup> International Conference on Laboratory Phonology (LabPhon 6), University of York.
- Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex*. New York: D. Appleton and Company.
- Davis, Alva, and Raven McDavid. 1949. ‘Shivaree’: An Example of Cultural Diffusion. *American Speech* 24: 249–255.
- Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.
- Dell, Gary. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93: 283–321.

- Dell, Gary, Lisa Burger and William Svec. 1997. Language Production and Serial Order: A Functional Analysis and a Model. *Psychological Review* 107: 123–147.
- Dell, Gary and Jean Gordon. 2003. Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities*. Berlin: Mouton de Gruyter.
- Dell, G. S., K.D. Reed, D.R. Adams, and A.S. Meyer. 2000. Speech errors, phonotactic constraints, and implicit learning: A study of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26: 1355–1367.
- Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 380–393.
- Del Viso, S., J.M. Igoa, and J.E. Garcia-Albea. 1991. On the autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research* 20: 161–185.
- Dennett, Daniel. 1995. *Darwin's Dangerous Idea*. New York: Simon and Schuster.
- Elman, Jeffrey. 1990. Finding Structure in Time. *Cognitive Science* 14: 179–211.
- Ernout, A. and A. Meillet. 1959. *Dictionnaire Étymologique de la Langue Latine*. Paris: Librairie C. Klincksieck.
- Ferguson, Charles A. 1963. Assumptions about nasals; a sample study in phonological universals. In J. Greenberg (ed.), *Universals of Language*. Cambridge: MIT Press. 53–60.
- Ferguson, C.A. and C. Farwell. 1975. Words and sounds in early language acquisition: English initial consonants in the first 50 words. *Language* 51: 419–439.
- Ferreira, Victor S. and Zenzi M. Griffin. 2003. Phonological influences on lexical (MIS)selection. *Psychological Science* 14(1): 86–90.
- Firth, John Rupert. 1930. Speech. In Peter Stevens (ed.), *The Tongues of Men and Speech*. Oxford: Oxford University Press.

- Fountain, Amy. 1998. An Optimality Theoretic approach to Navajo prefixal syllables. Ph.D. dissertation, University of Arizona. Available at Rutgers Optimality Archive (<http://roa.rutgers.edu>) as ROA-238.
- Frisch, Stefan. 1996. Similarity and Frequency in Phonology. Ph.D. dissertation, Northwestern University.
- Frisch, Stefan. 2004. Language Processing and Segmental OCP Effects. Phonetically-based Phonology, ed. B. Hayes, R. Kirchner and D. Steriade. Cambridge: Cambridge University Press. 346–371.
- Frisch, Stefan, Nathan Large and David Pisoni. 2000. Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords. *Journal of Memory and Language* 42: 481–496.
- Frisch, Stefan, Janet Pierrhumbert and Michael Broe. 2004. Similarity Avoidance and the OCP. *Natural Language and Linguistic Theory* 22: 179–228.
- Frisch, Stefan, and Bushra Zawaydeh. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77: 91–106.
- Fromkin, Victoria. 1971. The Non-anomalous Nature of Anomalous Utterances. *Language* 47(1): 27–52.
- Gamkrelidze, Thomas V., and Vjačeslav V. Ivanov. 1972. Lingvističeskaja tipologija i rekonstrukcija sistemy indoevropskix smyčnyx. Konferencija po sravnitel'no-istoričeskoj grammatike indoevropskix jazykov: Predvaritel'nye materialy. Moskva: Nauka. 15–18.
- Gathercole, S.E., C.R. Frankish, S.J. Pickering, and S. Peaker. Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25: 84–95.
- Giles, H. and P. Smith. 1979. Accommodation theory: Optimal levels of convergence. In: Giles, H., St. Clair, R. (eds.), *Language and Social Psychology*. Oxford: Blackwell.
- Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. 111–120.
- Greenberg, Joseph. 1966. *Language Universals, with special reference to feature hierarchies*. The Hague: Mouton.

- Grimm, Jakob. 1819/1837. *Deutsche Grammatik*. Göttingen: 4 Theile.
- Hammond, Michael. 1999. *The Phonology of English: A Prosodic Optimality-theoretic Approach*. Oxford: Oxford University Press.
- Harley, Trevor and Siobhan MacAndrew. 1992. Modeling paraphasias in normal and aphasic speech. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum. 378–383.
- Harley, Trevor and Siobhan MacAndrew. 1995. Interactive models of lexicalization: Some constraints from speech error, picture naming, and neuropsychological data. In J.P. Levym, D. Bairaktaris, J.A. Bullinaria, P. Cairns (eds.), *Connectionist Models of Memory and Language*. London: UCL Press. 311–331.
- Harley, Trevor and Siobhan MacAndrew. 2001. Constraints Upon Word Substitution Speech Errors. *Journal of Psycholinguistic Research* 30: 395–418.
- Harrison, K. David, Mark Dras and Berk Kapicioglu. 2002. Agent-Based Modeling of the Evolution of Vowel Harmony. In M. Hirotani (ed.), *Proceedings of the Northeast Linguistic Society* 32. 217–236.
- Haspelmath, Martin. 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18.2: 180–205.
- Hay, Jennifer, Janet Pierrehumbert and Mary Beckman. 2003. Speech Perception, Well-formedness and the Statistics of the Lexicon. In R. Ogden, J. Local & R. Temple (eds.), *Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press.
- Hayes, Bruce. 2000. Gradient Well-formedness in Optimality Theory. In Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford University Press. 88–120.
- Hayes, Bruce and Colin Wilson. To appear. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*.
- Heine, Bernd. 1999. *Ik Dictionary*. Köln: Köppe.
- Heinz, Jeffrey. 2007. Inductive Learning of Phonotactic Patterns. Ph.D. dissertation, UCLA.
- Hock, Hans Heinrich and Brian Joseph. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Berlin: Walter de Gruyter.

- Hony, H.C. and F. İz. 1992. *The Oxford Turkish Dictionary*. Oxford: Oxford University Press.
- Hopper, Paul J. 1973. Glottalized and murmured occlusives in Indo-European. *Glossa* 7. 141–166.
- Huffman, D. A. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers*. 1098–1101.
- Hull, David L. 1988. Interactors versus Vehicles. In Plotkin, Henry C. (ed.), *The Role of Behavior in Evolution*. Cambridge, MA: MIT Press. 19–50.
- Imbs, P. 1994. *Trésor de la langue française*. Paris: Centre National de la Recherche Scientifique. Online access at <http://atilf.atilf.fr/tlf.htm>.
- Inkelas, Sharon, Aylin Küntay, Orhan Orgun, and Ronald Sprouse. 2000. Turkish Electronic Living Lexicon (TELL). *Turkic Languages*, 4: 253–75. Online access at <http://socrates.berkeley.edu:7037>.
- Jäger, Gerhard. 2004. Maximum entropy models and stochastic Optimality Theory. Ms., University of Potsdam.
- Jescheniak, Jörg and Willem Levelt. 1994. Word Frequency Effects in Speech Production: Retrieval of Syntactic Information and of Phonological Form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(4): 824–843.
- Jescheniak, Jörg and Herbert Schriefers. Discrete serial versus cascaded processing in lexical access in speech production : Further evidence from the coactivation of near-synonyms. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24(5): 1256–1274.
- Jurafsky, Dan and James Martin. 2000. *Speech and Language Processing*. Englewood Cliffs: Prentice Hall.
- Jusczyk, Peter. 1997. *The Discovery of Spoken Language*. Cambridge: MIT Press.
- Kanwisher, Nancy. 1987. Repetition blindness: type recognition without token individuation. *Cognition* 27: 117–143.
- Kari, James. 1976. *Navajo Verb Prefix Phonology*. New York: Garland.
- Kawahara, Shigeto, Hajime Ono and Kiyoshi Sudo. 2005. Consonant Cooccurrence Restrictions in Yamato Japanese. T. Vance (ed.), *Japanese/Korean Linguistics* 14: 27–38.

- Kaye 2005. Gemination in English. *English Today* 21: 43–55.
- Kelz Sperling, Susan. 2005. *Poplollies & Bellibones: A Celebration of Lost Words Along with Tenderfeet and Ladyfingers: A Compendium of Body Language*. New York: William S. Konecky Associates.
- Khan, Sameer. To appear. Similarity Avoidance in East Bengali Fixed-Segment Reduplication. *Proceedings of the Western Conference on Linguistics*. Fresno, California.
- Kirby, S. and J. Hurford. 2001. The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi (eds.), *Simulating the Evolution of Language*. Springer Verlag.
- Kirchner, Robert. 2001. *An Effort Based Approach to Consonant Lenition*. New York: Routledge.
- Kochetov, Alexei. 2002. *Production, Perception and Emergent Phonotactic Patterns: A Case of Contrastive Palatalization*. New York: Routledge.
- Labov, William. 1973. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 2001. *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.
- Ladefoged, Peter. 2001. *A Course in Phonetics*. 4th edition. Fort Worth, TX: Harcourt.
- Lass, Roger. 1990. How to Do Things with Junk: Exaptation in Language Evolution. *Journal of Linguistics* 26: 79–109.
- Lass, Roger. 1997. *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.
- Lahiri, Aditi. 2002. Pertinacity in Representation and Change. Paper presented at the Workshop on Pertinacity, Schloss Freudental, July 10-14, 2002.
- Leonard, L.B., R.G. Schwartz, B. Morris and K. Chapman. 1981. Factors influencing early lexical acquisition: lexical orientation and phonological composition. *Child Development* 52: 882-887.
- Levelt, Willem. 1983. Monitoring and self-repair in speech. *Cognition* 14: 41-104.
- Levelt, Willem, Ardi Roelofs and Antje Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22: 1-38.

- Levelt, Willem and L. Wheeldon. 1994. Do speakers have access to a mental syllabary? *Cognition* 50: 239–269.
- Levitt, A. G., and A. F. Healy. 1985. The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language* 24: 717–733.
- Lewis, Charlton and Charles Short. 1879. *A Latin Dictionary*. Oxford: Clarendon Press.
- Lewis, G.L. 1967. *A Turkish Grammar*. Oxford: Oxford University Press.
- Lewis, Robert. 2000. *Middle English Dictionary*. Ann Arbor: University of Michigan Press.
- Lidell, Henry George, Robert Scott, and Henry Stuart Jones. 1940. *A Greek-English Lexicon*. Oxford: Clarendon Press.
- Livingstone, Daniel. 2002. The evolution of dialect diversity. In Angelo Cangelosi and Domenico Parisi (eds.), *Simulating the Evolution of Language*. New York: Springer-Verlag. 99–118.
- Livingstone, Daniel and Colin Fyfe. 1999. Modelling the Evolution of Linguistic Diversity. In Floreano, D., Nicoud, J-D. and Mondada, F.(eds), *Proceedings of the 5th European Conference in Artificial Life, ECAL'99*, Lausanne, Switzerland, September 1999. Springer.
- Luce, R.D. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.
- MacEachern, Margaret. 1999. *Laryngeal Cooccurrence Restrictions*. New York: Routledge.
- MacKay, Donald. 1970. Phoneme Repetition and the Structure of Languages. *Language and Speech* 13: 199–213.
- Maddieson, Ian. 1984. *Patterns of Sound*. Cambridge: Cambridge University Press.
- Magay, T. and L. Országh. 1990. *A Concise Hungarian-English Dictionary*. Oxford: Oxford University Press.
- Magnus, Margaret. 2001. What's in a Word? Studies in Phonosemantics. Ph.D dissertation, University of Trondheim.



- Markman, Ellen. 1984. The acquisition and hierarchical organization of categories by children. In C. Sophian (ed.), *Origins of cognitive skills*. Hillsdale, NJ: Erlbaum. 371–406.
- Martin, Andrew. 2005. The Effects of Distance on Lexical Bias: Sibilant Harmony in Navajo. MA Thesis, UCLA.
- Martin, Nadine, Deborah Gagnon, Myrna Schwartz, Gary Dell, and Eleanor Saffran. 1996. Phonological Facilitation of Semantic Errors in Normal and Aphasic Speakers. *Language and Cognitive Processes* 11(3): 257–282.
- Martin, S.C., H. Ney, and J. Zaplo. 1999. Smoothing methods in maximum entropy language modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I. 545–548.
- Matasović, Ranko. 1994. Proto-Indo-European \*b and the Glottalic Theory. *Journal of Indo-European Studies* 22, 133–149.
- McDonough, Joyce. 1991. On the representation of consonant harmony in Navajo. In *Proceedings of the Tenth West Coast Conference on Formal Linguistics*. 319–335.
- McDonough, Joyce. 2000. How to use Young and Morgan’s “The Navajo Language.” In K. M. Crosswhite & J. S. Magnuson (eds.), *University of Rochester Working Papers in the Language Sciences* 1(2): 195–214.
- McDonough, Joyce. 2003. *The Navajo Sound System*. Dordrecht: Kluwer Academic Publishers.
- McFarland, George. 1944. *Thai-English Dictionary*. Stanford: Stanford University Press.
- McFedries, P. 2004. *Word Spy. The Word Lover’s Guide to Modern Culture*. New York: Broadway Books.
- McMahon, April. 1994. *Understanding Language Change*. Cambridge: Cambridge University Press.
- Mesoudi, Alex, Andrew Whiten and Kevin Laland. 2004. Towards a unified science of cultural evolution. *Behavioral and Brain Sciences* 29: 329–383.
- Mester, Armin. 1986. Studies in Tier Structure. Ph.D. Dissertation, University of Massachusetts.
- Meyer-Lübke, W. 1935. *Romanisches Etymologisches Wörterbuch*. Heidelberg: Carl Winters Universitätsbuchhandlung.

- Milroy, Lesley. 1987. *Observing and analysing natural language*. Oxford: Blackwell.
- Motley, M. T. and B. J Baars. 1975. Encoding sensitivities to phonological markedness and transitional probability: Evidence from spoonerisms. *Human Communication Research* 2: 351–361.
- Niyogi, Partha. 2006. *The Computational Nature of Language Learning and Evolution*. Cambridge: MIT Press.
- Niyogi, Partha and Robert Berwick. 1997. Evolutionary consequences of language learning. *Linguistics and Philosophy* 20: 697–719.
- Nowak, Martin and Natalia Komarova. 2001. Towards an evolutionary theory of language. *TRENDS in Cognitive Science* 5(7): 288–295.
- Ohala, John. 1981. The listener as a source of sound change. In C.S. Masek, R.A. Hendrick, and M.F. Miller (eds.), *Papers from the Parasession on Language and Behavior*. Chicago: Chicago Linguistic Society. 178–203.
- Ohala, John. 1997. Aerodynamics of phonology. *Proceedings of the 4th Seoul International Conference on Linguistics*. 11–15 Aug 1997. 92–97.
- Ohala, John and Carol Riordan. 1979. Passive vocal tract enlargement during voiced stops. In J. J. Wolf and D. H. Klatt (eds.), *Speech Communication Papers*, New York: Acoustical Society of America. S. 89–92.
- Oldfield, R.C. and A. Wingfield. 1965. Response Latencies in Naming Objects. *Quarterly Journal of Experimental Psychology* 17(4): 273–281.
- Padgett, Jaye. 1995. *Stricture in Feature Geometry*, CSLI Publications, Stanford, California.
- Pater, Joe. 2005. Learning a stratified grammar. In Alejna Brugos, Manuella R. Clark-Cotton, and Seungwan Ha (eds.), *Proceedings of the 29th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press. 482–492.
- Pater, Joe, Rajesh Bhatt and Christopher Potts. 2007. *Linguistic Optimization*. Ms, University of Massachusetts, Amherst.
- Pedersen, Holger. 1951. Die gemeinindoeuropäischen und die vorindoeuropäischen Verschlusslaute. *Historisk-filologiske Meddelelser* 32/5. København: Munksgaard.
- Peperkamp, Sharon. 1997. *Prosodic Words*. HIL dissertations 34. The Hague: Holland Academic Graphics.

- Peters, Ann. 1983. *The Units of Language Acquisition*. Cambridge: Cambridge University Press.
- Peterson, Robert and Pamela Savoy. 1998. Lexical Selection and Phonological Encoding During Language Production: Evidence for Cascaded Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24(3): 539–557.
- Pharr, Pauline. Onomastic Divergence: A Study of Given-Name Trends among African Americans. *American Speech* 68: 400–409.
- Pinker, Steven. 1984. *Language learnability and language development*. Cambridge, MA: Harvard University Press.
- Plotkin, H.C. 1994. *Darwin Machines and the Nature of Knowledge: Concerning Adaptations, Instinct and the Evolution of Intelligence*. Harmondsworth: Penguin.
- Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Technical report, Rutgers University and University of Colorado at Boulder, 1993. ROA 537, 2002. Revised version published by Blackwell, 2004.
- Prince, Alan, and Bruce Tesar. 1999. Learning phonotactic distributions. Technical Report RuCCS-TR-54, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick. ROA-353.
- Pulleyblank, Douglas, and William Turkel. 2000. Learning Phonology: Genetic Algorithms and Yoruba Tongue Root Harmony. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford: Oxford University Press. 554–591.
- Rapp, Brenda and Matthew Goldrick. 2000. Discreteness and Interactivity in Spoken Word Production. *Psychological Review* 107(3): 460–499.
- Redford, Melissa, Chun Chi Chen and Risto Miikkulainen. 1998. Modeling the Emergence of Syllable Systems. In Morton Ann Gernsbacher and Sharon J. Derry (eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (COGSCI-98)*. 882– 886.
- Richardson, James. 1967. *A New Malagasy-English Dictionary*. Farnborough, Hants: Gregg Press.
- Richardson, Peter, and Robert Boyd. 2004. *Not by Genes Alone: How Culture Transformed Human Evolution*. Chicago: University of Chicago Press.

- Ritt, Nikolaus. 2004. *Selfish Sounds and Linguistic Evolution: A Darwinian Approach to Language Change*. Cambridge: Cambridge University Press.
- Sapir, Edward and Harry Hoijer. 1967. *The phonology and morphology of the Navaho language*. University of California Publications in Linguistics 40. Berkeley: University of California Press.
- Schwartz, R.G. and L.B. Leonard. 1982. Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language* 9: 319-336.
- Selkirk, Elizabeth. 1982. The syllable. In H. van der Hulst and N. Smith (eds.), *The structure of phonological representations*. Dordrecht, Netherlands: Foris.
- Sepp, Mary. 2006. Phonological Constraints and Free Variation in Compounding: A Corpus Study of English and Estonian Noun Compounds. Ph.D. dissertation, City University of New York.
- Shattuck-Hufnagel, Stephanie. 1979. Speech errors as evidence for a serial-ordering mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale, N. J.: Lawrence Erlbaum.
- Sherman, D. 1975. Stop and fricative systems: a discussion of paradigmatic gaps and the question of language sampling. *Working Papers on Language Universals* 17: 1–33.
- Simon, Herbert A. 1955. On a Class of Skew Distribution Functions. *Biometrika* 42: 425–440.
- Smith, Grant. 1996. Name sounds in recent U.S. elections. Paper presented at International Congress of Onomastic Scientists, August, Aberdeen, Scotland, UK.
- Smith, Grant. 1998. The political impact of name sounds. *Communication Monographs* 65: 154–172.
- Smith, Grant. 2007. The influence of name sounds in the congressional elections of 2006. Paper presented at LSA Annual Meeting, January, Anaheim, California.
- Solak, Aysin and Kemal Oflazer. Design and Implementation of a Spelling Checker for Turkish. *Literary and Linguistic Computing* 1993 8(3): 113–130.
- Steinmetz, Sol, and Barbara Ann Kipfer. *The Life of Language: The Fascinating Ways Words are Born, Live, and Die*. New York: Random House.

- Steels, L., F. Kaplan, A. McIntyre, and J. Van Looveren. 2000. Crucial factors in the origins of word-meaning. Paper given at ENST, Paris. 3-6 April 2000.
- Stemberger, Joseph. 1984. Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology* 1: 281–313.
- Stemberger, Joseph. 1985. An interactive activation model of language production. In A.W. Ellis (ed.), *Progress in the Psychology of Language*. LEA, London, Vol. 1.
- Stemberger, Joseph. 1990. Wordshape errors in language production. *Cognition* 35: 123–57.
- Stevenson, Robert and F.H. Eveleth. 1953. *Judson's Burmese-English Dictionary*. Rangoon: Baptist Board of Publications.
- Stockwell, Robert and Donka Minkova. 2001. *English Words: History and Structure*. Cambridge: Cambridge University Press.
- Storkel, Holly and Margaret Rogers. 2000. The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics and Phonetics* 14: 407–425.
- Tambovtsev, Yuri and Colin Martindale. 2007. Phoneme Frequencies Follow a Yule Distribution. *SKASE Journal of Theoretical Linguistics* 4: 2.
- Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. 2000. English speakers' sensitivity to phonotactic patterns. In M. B. Broe & J. B. Pierrehumbert 213 (eds.) *Papers in laboratory phonology V: Acquisition and the lexicon*. Cambridge: Cambridge University Press. 269–282.
- Trubetzkoy, Nikolaj. 1931. Die phonologischen Systeme. *Travaux du Cercle Linguistique de Prague* 4: 96–116.
- Vitevitch, Michael. 1997. The neighborhood characteristics of malapropisms. *Language and Speech* 40: 211–228.
- Vitevitch, Michael. 2002. The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28: 735–747.
- Vitevitch, Michael, Jonna Armbrüster, and Shinying Chu. 2004. Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30: 514–529.

- Vitevitch, Michael, and Mitchell Sommers. 2003. The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory and Cognition* 31: 491–504.
- Vitevitch, Michael, and Melissa Stamer. 2006. The curious case of competition in Spanish speech production. *Language and Cognitive Processes* 21: 760–770.
- Vousden, Janet, Gordon Brown and Trevor Harley. 2000. Serial Control of Phonology in Speech Production: A Hierarchical Model. *Cognitive Psychology* 41: 101–175.
- Wang, Jun, Les Gasser, and Jim Houk. 2006. Convergence Analysis for Collective Vocabulary Development. *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems*. Hakodate, Japan. 1378–1380.
- Wedel, Andrew. 2004. *Self-Organization and Categorical Behavior in Phonology*. PhD. dissertation, University of California, Santa Cruz.
- Wedel, Andrew. 2006. Exemplar models, evolution and language change. *The Linguistic Review* 23(3) 247–274.
- Weide, R.L. 1998. Carnegie Mellon Pronouncing Dictionary. Release 0.6. Available at <http://www.speech.cs.cmu.edu>.
- Weinreich, Uriel, W. Labov and M. I. Herzog. 1968. Empirical Foundations for a Theory of Language Change. In W. P. Lehmann (ed.), *Directions for Historical Linguistics: A Symposium*. Austin: University of Texas Press. 95–195.
- Wells, R. 1951. Predicting Slips of the Tongue. *Yale Scientific Magazine*, December. 9–12.
- Westbury, John and Patricia Keating. 1986. On the naturalness of stop consonant voicing. *Journal of Linguistics* 22. 145–166.
- Wilshire, Carolyn. 1999. The “tongue twister” paradigm as a technique for studying phonological encoding. *Language and Speech* 42: 57–82.
- Wilson, Colin. 2006. Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science* 30(5): 945–982.
- Young, Robert, and William Morgan. 1987. *The Navajo Language: A Grammar and Colloquial Dictionary*. Albuquerque: University of New Mexico Press.

Yule, G. Udny. 1924. A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B*. 213. 21–87.

Zipf, George. 1935. *The Psycho-biology of Language*. Boston: Houghton Mifflin.

Zuraw, Kie. 2002. Aggressive reduplication. *Phonology* 19: 395–439.