UCLA
UCLA Electronic Theses and Dissertations

Title
Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning

Permalink
https://escholarship.org/uc/item/0rz210sn

Author
Zymet, Jesse

Publication Date
2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Lexical propensities in phonology:

corpus and experimental evidence, grammar, and learning

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Linguistics

by

Jesse Zymet

2018

© Copyright by

Jesse Zymet

2018

ABSTRACT OF THE DISSERTATION

Lexical propensities in phonology:
corpus and experimental evidence, grammar, and learning

by

Jesse Zymet

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2018

Professor Bruce P Hayes, Co-Chair

Professor Kie Ross Zuraw, Co-Chair

Traditional theories of phonological variation propose that morphemes be encoded with descriptors such as [+/- Rule X], to capture which of them participate in a variable process. More recent theories predict that morphemes can have LEXICAL PROPENSITIES: idiosyncratic, gradient rates at which they participate in a process—e.g., [0.7 Rule X]. This dissertation argues that such propensities exist, and that a binary distinction is not rich enough to characterize participation in variable processes. Corpus investigations into Slovenian palatalization and French liaison reveal that individual morphemes pattern across an entire propensity spectrum, and that encoding individual morphemes with gradient status improves model performance. Furthermore, an experimental investigation into French speakers' intuitions suggests that they internalize word-specific propensities to undergo liaison.

The dissertation turns to modeling language learners' ability to acquire the idiosyncratic behavior of individual attested morphemes while frequency matching to statistical generalizations across the lexicon. A recent model based in Maximum Entropy Harmonic Grammar (MaxEnt) makes use of general constraints that putatively capture statistical generalizations across the lexicon, as well as lexical constraints governing the behavior of individual words. A series of learning simulations reveals that the approach fails to learn statistical generalizations across the lexicon: lexical constraints are so powerful that the learner comes to acquire the behavior of each attested form using only these constraints, at which point the general constraint is rendered ineffective. A GENERALITY BIAS is therefore attributed to learners, whereby they privilege general constraints over lexical ones. It is argued that MaxEnt fails to represent this property in its current formulation, and that it be replaced with the hierarchical MIXED-EFFECTS LOGISTIC REGRESSION MODEL (MIXED-EFFECTS MAXENT), which is shown to succeed in learning both a frequency-matching grammar *and* lexical propensities, by encoding general constraints as fixed effects and lexical constraints as a random effect. The learner treats the grammar and lexicon differently, in that vocabulary effects are subordinated to broad, grammatical effects in the learning process.

The dissertation of Jesse Zymet is approved.

Timothy Hunter

Megha Sundara

Bruce P Hayes, Committee Co-Chair

Kie Ross Zuraw, Committee Co-Chair

University of California, Los Angeles

2018

To my mother, Lynn, my father, Mark, and my brother, Corey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

Six years to the Ph.D., transformative almost entirely due to the colleagues and friends that were by my side (and, perhaps a trifle, to finally seeing the Great American West, well past the limits of my New Jersey hometown).

I begin with my advisors, Bruce Hayes and Kie Zuraw. Just how I could begin to think as Bruce Hayes does, with such vivid clarity and an eye for simple yet sweeping innovation in areas ripe for it, is surely a challenge that I can only ever hope to meet halfway. I cannot count the number of times I slapped my leg after an idea he proposed in solution to a Hulk-sized problem, whispering—yelling—to myself, "*Of course!*" Kie Zuraw knows basically everything. She was such an inspiring teacher of Graduate Phonology 1 that she pulled me from morphosemantics into phonology. She knew the answers to nearly every novel, often technical question that I hurled at her during advising meetings, supplying said answers immediately. She also shares and sympathizes with my plight of having a last name at the very end of the alphabet. I cannot thank Bruce and Kie enough for being incredibly dedicated, supportive advisors, for setting such a high standard for me, and for believing that I can achieve it. They have made just about every moment in their meetings with me incredibly useful.

Many thanks to Robert Daland, for years of insightful commentary on computational modeling, phonology, and life. To Megha Sundara, for teaching me how to design, run, and analyze an experiment, and for taking a critical eye to my work. To Tim Hunter, for stepping in as fourth member of my dissertation committee soon after landing at UCLA, and for long discussions about both finite-state and stochastic phonology. To Ed Stabler, for early advising, for teaching me about finite-state automata and formal learning theories, and for letting me store my car in his driveway over winter break. To Karen Jesney, for generously meeting with me throughout the dissertation process. And, finally, to Eric Baković, for reaching over from UC San Diego to check up on how I was doing, and for being an inspiration, both as an academic and as a person—oh, and also, for great times at LSA'15.

My deepest gratitude to Margit Bowler, Philippe Côté-Boucher, Scott Gaines, and Michael Lefkowitz, for sharing their friendship and immense knowledge and experiences with me, and for their support. To Margit, for sending me dozens of sun emoticons and getting me to eat new foods; to Phil, for unanticipated trips to Turbo Bingo or to an "Intelligent Dance Music" concert; to Scott, for episodes of The Untold History of the United States and a documentary about CIA blunders; and to Michael, for many hours of juggling and letting me play The Witness. To Roslyn Burns, Meaghan Fowlie, Eleanor Glewwe, John Gluckman, Margaret Kroll, Isabelle Lin, and Adeline Tan, for fun hangouts, lingchat, cats and dogs. To the phonetics lunchers, for much needed non-linguistic conversations up to and during the dissertation days. To the original cohort—Margit Bowler, Philippe Côté-Boucher, John Gluckman, Dana Kleifield, Margaret Kroll, Eunhwa Lee, Nicoletta Loccioni, Jennifer Kang, and Deniz Ozyildiz—for an extremely stimulating first year. And to a variety of other UCLA linguists, including Maayan Abenina-Adar, Dustin Bowers, Canaan Breiss, Marc Garellek, Laura Kalin, Nicoletta Loccioni, Travis Major, Connor Mayer, Laura McPherson, Chase O'Flynn, Adam Royer, Yu Tanaka, Kaeli Ward, Lauren Winans, Deborah Wong, Meng Yang, Z.L. Zhou, and many others, for wisdom, friendship, and support. To Coralie Seizilles de Mazancourt of the bridge into the forest-contained courtyard, Alec Stein, Laura Finch, Brandon Stevenson, and Patricia Wiley for great times with UCLA non-linguists beginning at Weyburn.

As for those Californians outside of UCLA. My deepest gratitude as well to Erik Zyman, for great times in general, for inspiration and motivation over the years and for believing in me, and for (long) conversations over the phone to get past the TRIALS OF GRADUATE SCHOOL, going far beyond the call of duty of a friend. To Jennifer Black and Max McGowen, for great friendship and deep conversations, for dozens of episodes of Project Runway, or for trips through: Sacramento, Lake Tahoe, Reno, Mono Lake, Bishop, Mojave, Lancaster; Tucson, Phoenix, Sedona, Flagstaff, the Grand Canyon; that area northeast of Littlefield, Utah on the 15, Hurricane ([hɚəkən]), Zion, Antelope Canyon, Monument Valley; Tijuana, Ensenada; Yosemite; Malibu, Santa Barbara, the Cuesta Grade, King City/the 25; San Pedro, Long Beach, Huntington Beach, Newport Beach; Sonoma, Napa; Palm Springs; Lassen, Mount Shasta, the Redwood Forest; Catalina Island; Helsinki, Hong Kong; Nothing; San Diego, La Jolla, Laguna Niguel; the Angeles National Forest, Calico, Baker, Primm, Jean, Paradise, Las Vegas, Red Rock Canyon, the Valley of Fire State Park, the Hoover Dam, the Mojave, Kelso, Wonder Valley; San Luis Obispo, San Simeon, Gorda, Big Sur, Carmel; and everything in between. When are we going to Death Valley? To Skyla Pojednic and friends, for Vaporwave and the desert. To my parallelism co-conspirator Jeff Adler, who everyone wants in their linguistics department—good catch, UCSC. And also to my other UCSC friends, including Nick Kalivoda, Andrew Hedding, Jason Ostrove, Deniz Rudin, and Clara Sherley-Appel, for making for a great home away from home. To great folks at USC, including Brian Hsu, Charlie O'Hara, Caitlin Smith, and Rachel Walker, for making conferences and joint USC/UCLA Phonology Seminar events so enjoyable. And to anyone else I've regrettably forgotten. Thank you all.

To my invaluable NJ-NY metro area and Rutgers friends, for sticking it out with me, even though I'm located on the other side of the country for the majority of the year. To my prior teachers and fellow spellers from the National Spelling Bee days, for setting a high standard for me early in life. And, finally, to my wonderful family: my parents, Lynn and Mark, and my brother, Corey, for love and support throughout this phase of life, and for splitting flight costs with me so that I could visit frequently enough such that it never really felt like I left New Jersey in the first place.

VITA

2011	Robert Jeffers Prize for Outstanding Contributions to the Linguistics Department Rutgers, The State University of New Jersey, New Brunswick, New Jersey
2011	Aresty Research Prize Rutgers, The State University of New Jersey
2011	B.A., with majors in Linguistics (departmental honors) and Quantitative Economics, and minors in Mathematics and Cognitive Science Rutgers, The State University of New Jersey
2012, 2015	Andrew W. Mellon Fellowship for Humanistic Studies University of California, Los Angeles
2013, 2014	Graduate Summer Research Mentorship Fellowship University of California, Los Angeles
2013 – 2014	Graduate Research Mentorship Fellowship University of California, Los Angeles
2014	National Science Foundation Graduate Research Fellowship Honorable Mention
2014	M.A., Linguistics University of California, Los Angeles
2013 – 2017	Teaching Assistant/Associate/Fellow, Instructor Department of Linguistics, University of California, Los Angeles
2017 – 2018	Dissertation Year Fellowship University of California, Los Angeles

PUBLICATIONS

2014 Zymet, Jesse. Distance-based decay in long-distance phonological processes. *Proceedings of the 32nd West Coast Conference in Formal Linguistics*. Cascadilla Press.

2017 Adler, Jeffrey & Jesse Zymet. Irreducible parallelism in phonology. *Proceedings of the North Eastern Linguistics Society Annual Meeting 47*, ed. by Andrew Lamont and Katerina Tetzloff. Amherst: University of Massachusetts, GLSA.

2017 Zymet, Jesse. Contradictory markedness preferences across morphological domains. *Proceedings of the 35th West Coast Conference in Formal Linguistics*. Cascadilla Press.

Chapter 1:

Introduction

Phonological variation—situations where a single morpheme can be realized with multiple phonetic forms in a single environment—has been the subject of great interest among phonologists in recent decades (Labov 1989; Kiparsky 1993b; Anttila 1997; Pater 2000; Zuraw 2000; Bailey & Hahn 2001; Frisch & Zawaydeh 2001; Albright & Hayes 2003; Ernestus & Baayen 2003; Becker 2009; Coetzee & Pater 2011; Moore-Cantwell & Pater 2016; among many others). Some perennial questions that have emerged are the following: How is variation represented in the linguistic system? How is it learned? How does it arise in language and change throughout time?

Several theories of variation propose that encoded in morphemes is a binary scale ([+/- Rule X]) that determines whether they trigger or undergo a phonological process (esp. Walther & Wiese 1999; Chomsky & Halle 1968, Kenstowicz & Kisseberth 1977, Anttila 1997, Pater 2000, Becker 2009, Jurgec 2016, among others). Other, more recent theories raise the possibility that encoded in morphemes are gradient parameters ([0.7 Rule X]), predicting that they should display **LEXICAL PROPENSITIES**—idiosyncratic, gradient rates at which they trigger or undergo a process (*cf*. Moore-Cantwell & Pater 2016, Zuraw 2016, Smolensky & Goldrick 2016, Zuraw & Hayes 2017). If it turns out that individual morphemes display such propensities, and that speakers have knowledge of these propensities, then that would have significant ramifications for theories of morphophonological representation and learning. It would favor theories that are capable of encoding a morpheme's status on an entire spectrum, and would suggest that learners are capable of *tracking* morpheme-specific rates of allomorphy—a problem in the field yet to be fully addressed (*cf.* Moore-Cantwell & Pater 2016, Smolensky & Goldrick 2016). Moreover, it would validate theories that allow encoding on a morpheme-by-morpheme basis, and suggests that variation cannot be explained by referring to idiosyncrasies of stored larger constituents alone (*cf.* Zuraw 2000, 2010; Bybee 2001, 2002). In this investigation, I argue that individual morphemes can display differing propensities to participate in a variable process, and that learners internalize these propensities. These propensities can be associated with both triggering and undergoing morphemes. The evidence for these claims comes from a series of corpus investigations into variable Slovenian palatalization and French liaison, and a nonce probe investigation into the intuitions of native French speakers.

Consider the case of variation in French liaison. Smolensky & Goldrick (2016) raise the question of whether lexical propensities are associated with this variable process, in which a consonant spelled at the end of a given word (*Word1*) is pronounced only if the following word (*Word2*) is vocoid-initial, as in the data below. While *très*, 'very' appears to undergo liaison categorically before vocoid-initial words, *plus*, 'more' undergoes the process *optionally* before vocoid-initial words — it cannot be categorized categorically liaising *or* non-liaising.

(1a) *très beau*, [trɛ bo]
'very beautiful'

très intelligent, [trɛ **z** ɛ̃tɛliʒɑ̃]
'very intelligent'

(1b) *plus beau*, *[ply bo]*
'more beautiful'

plus intelligent, [ply **z** ɛ̃tɛliʒɑ̃] ~[ply ɛ̃tɛliʒɑ̃]
'more intelligent'

An investigation into French liaison obtains the following: in a corpus of spoken French, individual words display differing propensities to undergo liaison, and a model of the data is improved if Word1 identity is encoded, even after factors previously found to affect liaison are controlled for. The corpus data come from the *Phonologie du Français Contemporain* (Durand, Laks & Lyche 2002, 2009; Durand & Lyche 2008), a database of spoken French containing around 54,000 Word1-Word2 pairs. A histogram of Word1s occurring 100 or more times in the corpus reveals that though most of them are clear liaisers and non-liaisers, a healthy minority of Word1's undergo liaison at medial rates.

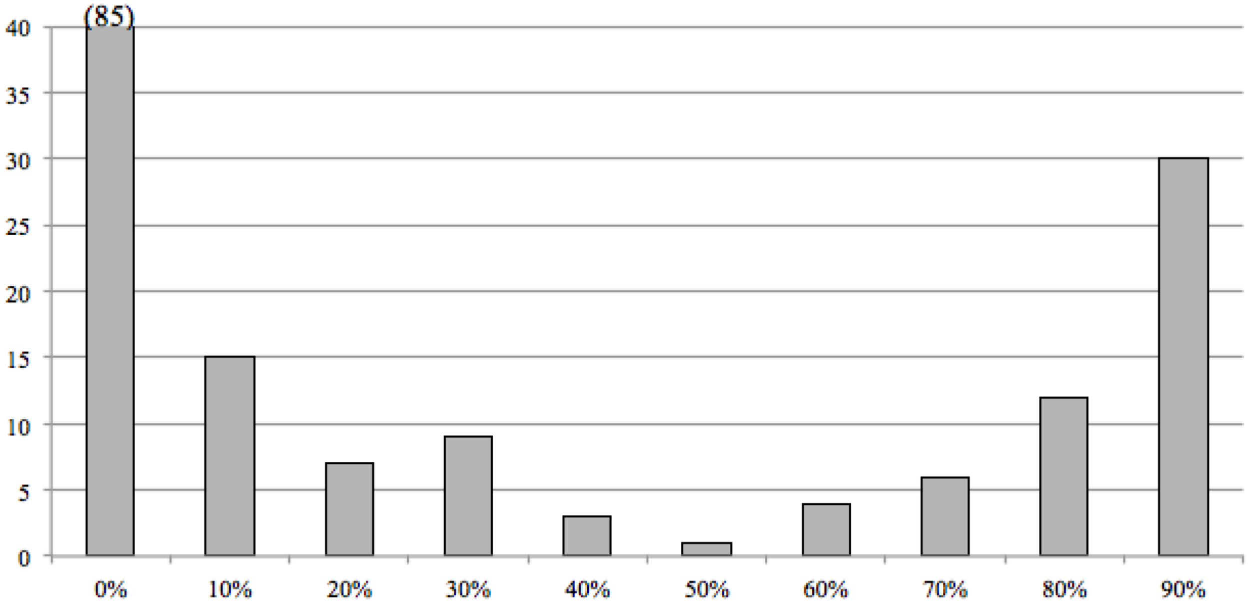


Figure 1: *Histogram showing number of common Word1's with a particular liaison rate*

Moreover, a nonce probe investigation into French speakers' intuitions suggests that French learners internalize the liaison propensities of different Word1s. Prior corpus studies and my PFC corpus study of French liaison give different rates for the adverbs *très* ('very', 97%, Mallet 2008), *plus* ('more', 64%, Mallet 2008), *bien* ('very', 43%, Mallet 2008), *moins* ('less', 26%, my PFC study) and *pas* ('not', 1%, Mallet 2008). In a nonce probe study, native speakers of French were presented with *très*, *plus*, *bien*, *moins*, and *pas* followed by nonce vowel-initial adjectives (e.g., "*très arvant*", "*bien agrivieux*") and were asked whether they preferred the liaised or non-liaised form. Overall, participants replicated the distinctions found in corpora.

These findings support theories that endow individual morphemes with gradient parameters, and theories that permit encoding on individual morphemes rather than exclusively on larger constituents, whole domains, or groups of morphemes.

In light of the fine-grained, quantitative nature of phonological variation, the dissertation turns to an investigation into theories that predict the existence of lexical propensities. These models aim to capture: the language learner's behavior in nonce probe studies to frequency match to statistical generalizations found across the lexicon; the idiosyncratic behavior of individual attested morphemes to abide by or deviate from these generalizations. A recent model for learning a frequency-matching grammar together with lexical propensities based in Maximum Entropy Harmonic Grammar (MaxEnt) makes use of general constraints that putatively capture statistical generalizations across the lexicon, as well as lexical constraints governing the behavior of individual words (e.g., Moore-Cantwell & Pater 2016, Zuraw & Hayes 2017, Tanaka 2017). MaxEnt treats general constraints and lexical constraints as *equally viable* explanatory variables for learning the dataset and its patterns. A series of learning simulations reveals that the approach fails to learn general, grammatical trends for this very reason, as it runs into a **GRAMMAR-LEXICON BALANCING PROBLEM**: lexical constraints are so powerful in explaining the dataset that that the learner comes to acquire the behavior of each form using only these constraints, at which point the general constraint is rendered ineffective. A **GENERALITY BIAS** is therefore attributed to learners, whereby they privilege general constraints over lexical constraints. It is argued that MaxEnt—essentially an simple logistic regression model—fails to represent this property, and that it be replaced with a mixed-effects logistic regression model, **MIXED-EFFECTS MAXIMUM ENTROPY HARMONIC GRAMMAR**, which is shown to succeed in learning both grammatical *and* item-specific behavior by encoding general constraints as fixed effects and lexical constraints as a random effect. The learner treats the grammar and lexicon differently, in that granular effects of the vocabulary are subordinated in the learning process to broad effects in the grammar. Mixed models are used widely in linguistics experiments and across scientific fields, and have proven to be highly effective in modeling datasets displaying variation both within morphophonology and other fields of linguistics. Here I present an argument that adopting the mixed-effects logistic regression model as a theory of the language learner is a crucial step toward capturing the capabilities of language learners.

1.1. Structure of the dissertation

This dissertation is structured as follows. Chapter 2 presents a literature review of major investigations into variation and lexical propensities, and models of how they are learned and represented in the linguistic system. It highlights a series of experiments that suggest that language learners frequency match to quantitative trends across the lexicon. Chapter 3 presents an investigation into Slovenian palatalization, showing that suffixes display distinct lexical propensities to trigger palatalization while stems display distinct propensities to undergo it, even after taking into account phonological factors previously found to condition the variation. Chapter 4 presents the corpus investigation into French liaison, also showing that individual Word1s display distinct lexical propensities to undergo liaison, even after a host of other factors previously found to condition the variation are taken into account. Chapter 5 presents a series of nonce probe studies into the intuitions of French speakers, with results suggesting that speakers internalize word-specific contrasts observed in corpus investigations into French liaison. Chapter 6 compares models for learning a frequency-matching grammar together with lexical propensities. Though models couched in Maximum Entropy Harmonic Grammar are characterized by overfitting lexical constraints to the data, the mixed-effects logistic regression model—which incorporates a generality bias by privileging general constraints over lexical constraints—surmounts the overfitting problem, succeeding to learn a frequency-matching grammar together with lexical propensities. Chapter 7 presents a series of previous investigations into an apparent learner bias to generalize constraints across morphosyntactic domains—some of which implicate the bias hypothesis, and others of which contradict it. I add another case suggesting that learners can acquire a highly morphosyntactically specific process: a backness dissimilation alternation in Malagasy that lacks a counterpart generalization in phonotactics. The section closes by raising a set of questions concerning how generality bias should be represented most broadly in the system, such that it can accommodate learner generalization while sustaining specificity at various levels of structure, whether it be morphemes, domains, or grammatical categories.

Chapter 2:

Literature review: lexical idiosyncrasies and aggregate generalizations in phonological variation

The enterprise of incorporating variability into phonological theory dates back at least as far as Chomsky & Halle (1968), who used minor rules and exception features to capture variation across words. Variation received considerable attention within sociolinguistic research in generative phonology (Labov 1973, 1989, 1994, 2001; Wolfram 1969; Cedergren & Sankoff 1974; Trudgill 1974; Guy 1980, 1991, 1997; *et seq*), but beyond seemingly only a handful of well-known works dating before the nineties (e.g., Chomsky & Halle 1968; Kenstowicz & Kisseberth 1977; Zonneveld 1978), variation has only received substantial attention from phonological theorists in the past twenty or so years (Anttila 1997, Boersma 1997, Boersma & Hayes 2001, Pater 2000, Zuraw 2000, Hayes & Londe 2006, Becker 2009, Coetzee & Kawahara 2013, *inter alia*).

To set up the discussion on lexical propensities, this chapter provides a brief survey of a few of the leading works on phonological variation. We begin with empirical background on variable secondary stress in English, as well as Pater (2000)'s classic treatment of the phenomenon using lexically specific constraints in Optimality Theory. I then review a series of prior corpus and experimental investigations that reveal language learners' ability to frequency match to aggregate trends in the lexicon. I review Zuraw (2000)'s frequency-matching model, which learns and represents both idiosyncratic pronunciations of individual words as well as aggregate trends across the lexicon. Finally, I review a more recent set of works suggesting that individual words or morphemes can display idiosyncratic lexical propensities to participate in a variable process. I follow up with a discussion of the current state of models for the learning and representation of these propensities.

2.1 Pater (2000): lexical variation in English secondary stress, and an early OT account

Pater (2000) addresses within Optimality Theory (Prince & Smolensky 1993/2004) a classic case of lexical variation, the assignment of English secondary stress. The following data illustrate the puzzle:

(2a) infórm ìnformátion
(2b) impórt ìmpòrtátion

Pater's approach to these data is to say that secondary stress on the second syllable is an idiosyncratic property of certain words (e.g., *ìmpòrtátion*). This arbitrariness is represented in the grammar using lexically indexed constraints, which target only a subset of the lexicon.

Primary stress occurs consistently on the rightmost nonfinal syllable, which Pater accounts for using constraints aligning heads to the right of the prosodic word as well as constraints against non-finality. Initial secondary stress also occurs consistently, accounted for by constraints that require non-final bimoraic syllables to be parsed into feet. The central challenge is accounting for lexical variation in the occurrence of pretonic secondary stress, in which stress is present in (2a) but absent in (2b). In a first pass, Pater posits the following constraints (definition in (3b) modified for purposes of readability; *cf*. Pater (2000), p. 252-254 for further discussion):

(3a) CLASH-HEAD: no stressed syllable can flank the head syllable of a prosodic word.

(3b) IDENT-OO(stress): if a syllable in base form *x* is stressed, then the corresponding syllable in the derived form *f*(*x*) is stressed.

CLASH-HEAD prefers destressing of the pretonic syllable to avoid clash, while IDENT-OO(stress) prefers stress preservation in (2a-b), as their base forms *infórm* and *impórt* have stress on the second syllable. Ranking CLASH-HEAD above IDENT-OO(stress) accounts for pretonic stresslessness in *ìnformátion* (see below), but fails to predict preservation in *ìmpòrtátion*. Pater thus posits the lexically-specific faithfulness constraint IDENT-OO(stress)-S, a version of IDENT-OO(stress) that only targets words belonging to some set S. By specifying that the underlying form *importation*, but not *information*, is included in S, their two surface forms can be accounted for as follows:

information base form: *infórm*	IDENT-OO(stress)-S	CLASH-HEAD	IDENT-OO(stress)
[ìn][fòr][má]tion		*!	
→ [ìnfor][má]tion			*
importation-S base form: *impórt*	IDENT-OO(stress)-S	CLASH-HEAD	IDENT-OO(stress)
→ [ìm][pòr][tá]tion		*	
[ìmpor][tá]tion	*!		*

Table 1: *tableaux with lexically specific constraints (adapted from Pater 2000, p. 259)*

High-ranked IDENT-OO(stress)-S targets only **ìmportátion*, since its underlying form is specified as belonging to S—thus, stress is preserved so that *ìmpòrtátion* surfaces. Otherwise, *ìnformátion* is simply the result of constraints on stress clash and faithfulness. Lexically variable alternation is thus represented as the conflict between lexically-specific constraints, or constraints that target subsets of lexemes, and constraints motivated independently by the rest of the grammar.

2.2 Frequency matching in experimental investigations into phonological variation

Just what do speakers internalize when they internalize a system of phonological variation? A number of experiments dating from the early 2000's suggest that when speakers learn such a system, they acquire not only the idiosyncratic pronunciations of each form (and, in particular, whether or not any given form will idiosyncratically alternate), but also frequency match to aggregate trends across the lexicon (Frisch, Broe, & Pierrehumbert 1996; Coleman & Pierrehumbert 1997; Eddington 1998, 2004; Berkley 2000; Zuraw 2000, 2010; Bailey & Hahn 2001; Frisch & Zawaydeh 2001; Pierrehumbert 2002; Albright 2002; Albright & Hayes 2003; Ernestus & Baayen 2003; Hayes & Londe 2006; *et seq*). In particular, a number of experiments reveal that when speakers of a language with variation are tested on novel items eligible to undergo the variable process, their responses match lexical frequencies in the aggregate. We review a few case studies below, as well as approaches to modeling knowledge of lexical trends together with knowledge of lexical idiosyncrasy.

2.2.1 Zuraw (2000, 2010): knowledge of lexical idiosyncrasy and lexical trends in Tagalog

Zuraw (2000, 2010) presents a case study of frequency matching by Tagalog speakers to quantitative trends in the lexicon, and proposes a model that jointly represents listed information together with lexical trends. The central case study is variable nasal assimilation and nasal substitution in Tagalog. When a set of prefixes attach to stems with initial segments which lack place, they surface as *paŋ*-, *maŋ*-, and *naŋ*- (e.g. *hukbo* ‘army’, *paŋ-hukbo* ‘military’); but when the initial segment has place, the prefix-final nasals either assimilate in place or the nasal and the subsequent segment fuse into a nasal whose place is identical to that of the segment. Data showing variation of outcome are provided below.

(4)	*b*:	mag-**b**igaj, ‘to give’	ma-**m**igaj, ‘to distribute’	*Substitution*
		bigkas, ‘to pronounce’	ma**m-b**i-bigkas, ‘reciter’	*Assimilation*
	d:	**d**alaːŋin, ‘prayer’	ʔi-pa-**n**alaʔin, ‘to pray’	*Substitution*
		dinig, ‘audible’	pa**n-d**inig, ‘sense of hearing’	*Assimilation*
	g:	**g**indaj, ‘unsteadiness on feet’	pa-**ŋ**i-**ŋ**indaj, ‘unsteadiness on feet’	*Substitution*
		gaːwaj, ‘witchcraft’	ma**ŋ-g**a-gaːwaj, ‘witch’	*Assimilation*

Note that forms like [ma**m-b**i-bigkas] and [pa-**ŋ**i-**ŋ**inday] have a reduplicative morpheme; if the form undergoes substitution, then both the first segment of the reduplicative morpheme and the stem-initial segment are substituted.

Zuraw gathered a set of 1,736 words that had an obstruent-initial stem and substituting/assimilating prefixes, obtaining two trends: first, substitution is most likely with a front obstruent (*p* and *b*) and least likely with a back obstruent (*k* and *g*); second, substitution is more likely when the obstruent is voiceless rather than voiced. The bar chart below shows affix behavior of the two most common affixes, *paŋ*- and *maŋ*-, when they come before different stem-initial obstruents. The bars show both the place trend and the voicing trend. Zuraw performed a contingency table analysis, finding that both place of articulation and voicing were significant predictors of substitutability.

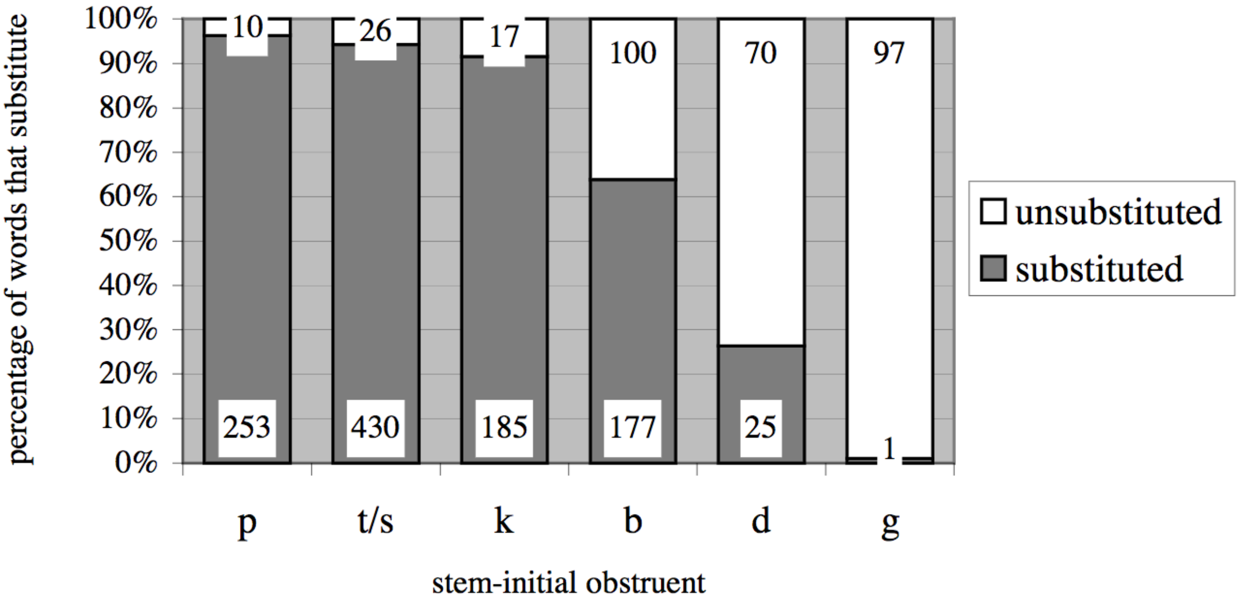


Figure 2: *trends in Tagalog nasal substitution (Zuraw 2000, p. 23)*

Zuraw raises the question of whether speakers internalize these trends. She notes that, despite these trends, the behavior of individual words is unpredictable, and moreover even the behavior of derivatives of the same stem is unpredictable. She gives the following examples demonstrating the latter:

(5) *bigaj*, 'gift'
pa**m-b**igaj, 'gifts to be distributed'
pa-**m**i-migaj, 'act of giving away'
maː-**m**i-migaj, 'distributor'
ma-**m**igaj, 'to distribute (actor focus)'

bugbog, 'wallop'
pa-**m**ugbog, 'wooden club used to pound clothes during washing'
pa**m-b**u-bugbog, 'act of clubbing or pounding; assault'
ma**m-b**ugbog, 'to wallop'

bulos, 'harpoon'
pa-**m**ulos 'harpoon'
ma**m-b**u-bulos 'harpooner'

buʔoʔ, ‘whole’
pa**m**-**b**uʔoʔ, ‘something used to produce a whole’
pa-**m**u-muʔoʔ, ‘becoming whole; coagulation’
ma-**m**uʔoʔ, ‘to solidify; to clot’

In light of the above, it could very well be that speakers simply list each construction in the lexicon, without internalizing nasal substitution, or any kind of lexical trend associated therewith.

Hence she administered a nonce probe study to nine native speakers of Tagalog, testing for the following: (i) whether speakers internalized nasal substitution; (ii) whether speakers internalized the lexical trends within nasal substitution. In the acceptability task of her study, for example, Zuraw designed a set of sentences probing at whether speakers internalized these facts for *maŋ*-RED$_{\text{CV}}$-STEM constructions—where *maŋ*- is a prefix that forms professional or habitual nouns (as in the English suffx *–er*). Speakers were presented with a series of sentence pairs. In both sentences, the speaker is first presented with a *pag*-RED$_{\text{CV}}$-STEM construction, followed by a *maŋ*-RED$_{\text{CV}}$-STEM construction, where the stem is novel. One sentence in the pair displays nasal substitution to the latter, while the other sentence displays only assimilation. Speakers were asked to rate each sentence of each pair on a scale from 1 (bad) to 10 (good).

The participants’ results mimicked the lexical frequencies found in her corpus, as illustrated in the table below (note that in the table below the *y*-axis is the difference between the acceptability scores for a stem substituted and the same stem unsubstituted—positive values mean that substituted was rated higher than unsubstituted, and lower values mean that substituted was rated lower than unsubstituted). Overall, speakers rated nasal substitution as a broadly applicable process to obstruent-initial nonce words, in that none of the acceptability scores approached -5. Furthermore, on average, speakers replicated the lexical trends observed in the corpus data on nasal substitution: they rated substitution of voiced obstruents as more acceptable than assimilation before them; and they generally rated the substitution of more front obstruents as more acceptable than that of more back obstruents (except for *p*, which was associated with lower acceptability ratings relative to coronals).

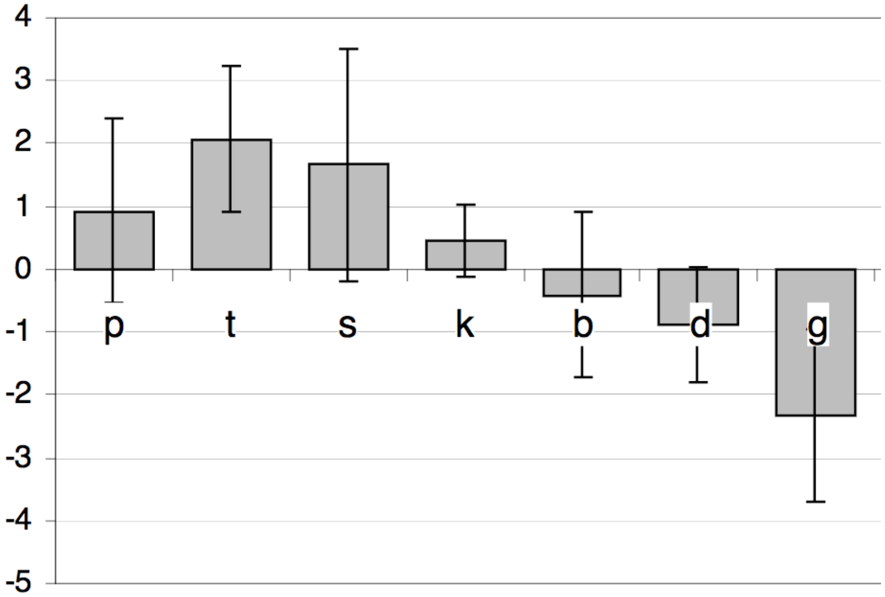


Figure 3: *acceptability rating differences across nine Tagalog speakers (Zuraw 2000, p. 43)*

The grammar proposed aims to capture two both the idiosyncratic behavior of attested words together with the broad applicability of nasal substitution and the trends displayed across the lexicon—both the voicing effect and place effect. The approach reviewed below is adapted from Zuraw (2010), which is based upon the Zuraw (2000) dissertation. The grammar and the vocabulary are accounted for in Optimality-Theoretic system with stochastic constraint ranking (Hayes & MacEachern 1998, Boersma 1997, Boersma & Hayes 2001, Hayes 2000, among others). Zuraw notes a potential three-way distinction that must be captured: words lexicalized as undergoing nasal substitution (e.g., *ma-ma-mahala?*, ‘responsibility’, with 81 tokens in her corpus, related to *bahala?*, ‘manager’); words lexicalized as not undergoing nasal substitution (e.g., *mam-babasa*, ‘reader’, with 725 tokens in her corpus, related to *basa*, ‘reading’); and words whose behavior is not lexicalized. The behavior of idiosyncratic, lexicalized words must be protected by lexical information, and not be subject to the broader grammar. Thus a certain set of whole words which behave idiosyncratically—that is, various {*ma[+nas]*, *pa[+nas]*}(+RED$_{CV}$)+STEM constructions (Zuraw assumes the triggering suffixes have final nasal features in her 2010 paper, rather than *ŋ*)—are stored in the lexicon, with IO-FAITH and OO-FAITH governing their behavior. For example, [ma-migaj], 'to distribute', which originated from *ma[+nas]*- and the stem *bigaj*, is therefore stored as /mamigaj/, and is captured as in the table below. IO-FAITH ranks above OO-FAITH, such that stored /mamigaj/ surfaces faithfully, rather than as a form with nasal assimilation, which is faithful to the base of the related form [mag+bigaj], 'to give'.

/mamigaj/ related to [mag+bigaj]	IO-FAITH	OO-FAITH
→ mamigaj		*
mambigaj	*!	

Table 2a: *IO-FAITH trumps in OO-FAITH determining the output of stored constructions with substitution (Zuraw 2010, p. 444)*

On the other hand, the word [mam$_1$b$_2$abasa] is lexicalized as undergoing assimilation rather than substitution, and thus is stored as /mam$_1$b$_2$abasa/. High-ranking MAX-IO protects against deletion of the obstruent, while UNIFORMITY-IO protects against the fusing of m_1 and b_2; low-ranking NOCODA, militating against codas, is violated, such that /mam$_1$b$_2$abasa/ surfaces faithfully.

/mam$_1$b$_2$abasa/	MAX-IO	UNIFORMITY-IO	NOCODA
→ mam$_1$b$_2$abasa			*
mam$_1$abasa	*!		
mam$_{1,2}$abasa		*!	

Table 2b: *High-ranking IO-FAITH trumps in* NOCODA *in determining the output of stored constructions with assimilation (Zuraw 2010, p. 445)*

Words with behaviors not yet established are not stored, and so IO-FAITH constraints are irrelevant. *ASSOCIATE, a constraint violated when the prefix-final nasal feature associating to the subsequent obstruent, militates against nasal substitution, and is therefore ranked variably with DEP-C, which is violated by forms displaying assimilation, where a full nasal segment is spelled out in the prefix. The two constraints, variably ranked, result in variable surfacing of substitution and assimilation in forms which are not yet stored, as in the table below:

/ma[+nas]$_1$+b$_2$log/	*ASSOCIATE	DEP-C
→ ma-m$_{1,2}$log	*	
→ mam$_1$-b$_2$log		*

Table 2c: *variable ranking of *ASSOCIATE and DEP-C resulting in variable substitution and assimilation in forms not yet stored (adapted from Zuraw 2010, p. 445)*

High-ranking faithfulness constraints preserve lexical information, while the constraints determining the behavior of unstored forms are lower-ranked. To capture the voicing effect, Zuraw posits *NC̥, ranked variably with *ASSOCIATE to generate higher likelihood of substitution to forms with stem-initial voiced obstruents. To capture the place effect, Zuraw posits *[ŋ, *[n, and *[m, also variably ranked. In Optimality Theory with stochastic constraint ranking, constraints are assigned ranking values on a continuous scale. Via the Gradual Learning Algorithm (Boersma 1997), her system learns from whole, listed words the ranking values for these constraints in such a way that *[ŋ has a stronger effect in the grammar than *[n, which in turn has a stronger effect than *[m, in preventing substitution from applying to unstored forms. This derives higher rates of substitution before labials, medial rates before coronals, and lower rates before velars. The system also learns a modestly high ranking value for *NC̥—higher than that of *ASSOCIATE—ensures that unstored forms with stems with initial voiceless obstruents undergo substitution more readily than forms with stems with initial voiced obstruents. Finally, the system also learns a very high ranking value for INTEGRITY-IO and UNIFORMITY-IO, thereby deriving idiosyncrasies in individual words—that is, it learns whether some word is stored as having undergone substitution or not. The figure below shows that the learner applies nasal substitution variably, and acquires the lexical trends associated therewith, having learned from the sample of stored forms.

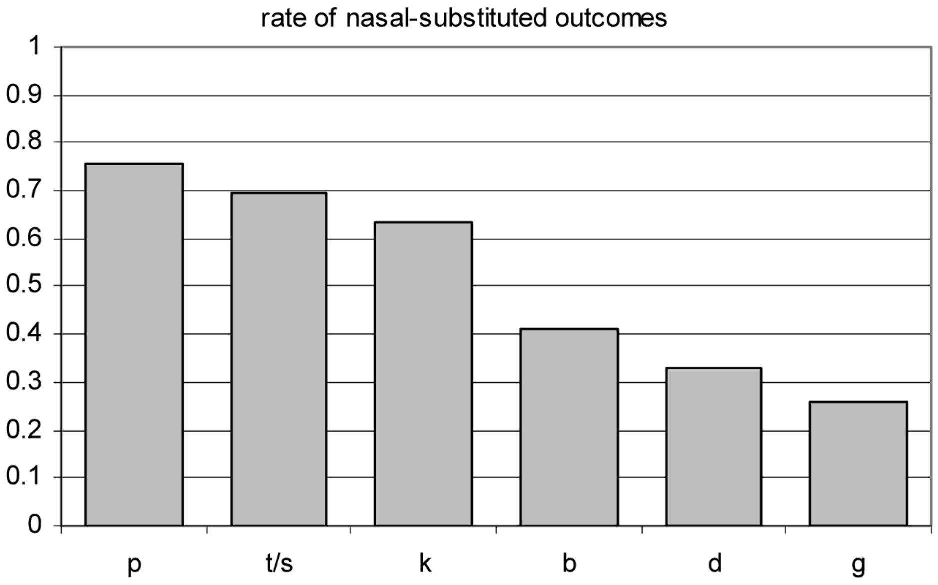


Figure 4: *variable nasal substitution and the voicing and place effects learned from stored forms (Zuraw 2010, p. 450)*

Zuraw's model thus captures the dualism of Tagalog speakers' knowledge of phonological variability: both the idiosyncrasy of stored whole words, as well as the phonological trends observed in the aggregate, across the entire lexicon. High-ranking faithfulness constraints together with USELISTED ensure that lexical information is protected, whereas the lower-ranking constraints—applicable only when the learner encounters an unstored form—and their relative rankings derive the variable tendency towards substitution as well as the voicing and place effects. For another approach to modeling of idiosyncrasy together with lexical trend occurring around this time, see Becker (2009), on a model of lexical variation and trends in Turkish voicing alternations that builds on Pater (2000, 2007)'s concept of constraint cloning.

Before moving forward, I note that Zuraw's (2000, 2010) model does not overcome the general problem of how to model learning of a frequency-matching grammar together with lexical idiosyncrasy. As we will see in Chapters 3, 4, and 5, *individual morphemes* can vary across a propensity spectrum to undergo a phonological process, rather than stored whole forms. Moreover, more recent works (Zuraw & Hayes 2017, Smith & Pater 2017) have found that stochastic OT captures only a proper subset of paradigms displaying variation relative to probabilistic Harmonic Grammar, which is based in weighted constraints. Hence a central task of the dissertation (in particular, Chapter 6) is to find a model that learns a frequency-matching grammar in the face of lexical idiosyncrasy down to the level of morpheme in a more encompassing, constraint weight-based framework.

2.2.2 Other investigations into frequency matching to lexical trends

Ernestus & Baayen (2003) present a simple case of frequency matching by speakers to alternation trends observed in the Dutch lexicon. The language features contrastive obstruent voicing word-internally as well word-final obstruent devoicing. For example, the two infinitives *verwijden* [vɛrʋɛi**d**ən] 'widen'-INF and *verwijten* [vɛrʋɛi**t**ən] 'reproach'-INF form a minimal pair that establishes the voicing contrast, while the suffixless forms *verwijd* [vɛrʋɛi**t**] 'widen' and *verwijt* [vɛrʋɛi**t**] 'reproach' reveal word-final voicing neutralization. Interestingly, the lexicon displays a tendency that relates the likelihood of a voiced word-internal obstruent to its place and manner. In particular, as the table below illustrates, Ernestus & Baayen's study of the CELEX corpus uncovers that stops are less likely to be voiced relative to fricatives, and fronter obstruents are less likely to be voiced relative to backer obstruents. Moreover, they find that when speakers are presented with a novel form with a word-final voiceless obstruent, speakers' judgments about whether the same form coming before a vowel-initial suffix should have a voiced obstruent match fairly closely with the frequencies given in the lexicon.

Obstruent	*#voiced/total in lexicon*	*%voiced in lexicon*	*%voiced in experiment*
p/b	20/230	9%	4%
t/d	177/719	25%	9%
f/v	151/451	33%	23%
s/z	116/166	70%	49%
x/ɣ	127/131	97%	80%

Table 3: *CELEX statistics for intervocalic obstruent voicing*

The findings suggest that Dutch speakers not only possess knowledge about which words in their lexicon feature a voiced obstruent word-internally, but knowledge about the broad trends across the lexicon illustrated above. The investigators present a series of models, one being an analogical model and another being based in stochastic OT, all of which succeed in capturing the data.

Hayes & Londe (2006) further uncover that Hungarian speakers frequency match to lexical trends in vowel harmony. The investigators show that, in a large corpus of data on Hungarian vowel harmony compiled by querying Google (see Hayes & Londe 2006 for specifics), whether a back vowel in the stem triggers backness harmony on the dative suffix vowel *–nɛk* depends on the height and number of intervening neutral vowels, with higher neutral vowels correlated with higher application rates and more neutral vowels correlated with lower application rates. The examples below illustrate the two trends; Figure 5 gives proportions in the corpus.

	UR	Dative form	Gloss
	Intervening i, iː		
(6a)	/ɔpoʃt**o**li+n**ɛ**k/	[ɔpost**o**li-n**ɔ**k]	‘apostolic’
	/b**u**li+n**ɛ**k/	[b**u**li-n**ɔ**k]	‘party’
	/m**a**ːrtiːr+n**ɛ**k/	[m**a**ːrtiːr-n**ɔ**k]	‘martyr’
	Intervening eː		
(6b)	/f**ɔ**seːn+n**ɛ**k/	[f**ɔ**seːn-n**ɛ**k]	‘charcoal’
	/ɔd**ɔ**leːk+n**ɛ**k/	[ɔd**ɔ**leːk-n**ɔ**k]	‘datum’
	/g**ɔ**lleːr+n**ɛ**k/	[g**ɔ**lleːr-n**ɔ**k]	‘collar’
	Intervening ɛ		
(6c)	/komp**o**nɛns+n**ɛ**k/	[komp**o**nɛns-n**ɛ**k]	‘component’
	/hɔmb**u**rgɛr+n**ɛ**k/	[hɔmb**u**rgɛr-n**ɛ**k]	‘hamburger’
	/kr**ɔ**pɛk+n**ɛ**k/	[kr**ɔ**pɛk-n**ɔ**k]	‘dude’

	UR	Dative form	Gloss
	Zero neutral vowels		
(7a)	/ɔbl**ɔ**k+n**ɛ**k/	[ɔbl**ɔ**k-n**ɔ**k]	‘window’
	/bir**o**ː+n**ɛ**k/	[bir**o**ː-n**ɔ**k]	‘judge’
	/kommunizm**u**s+n**ɛ**k/	[kommunizm**u**s-n**ɔ**k]	‘Communism’
	One neutral vowel		
(7b)	/f**ɔ**seːn+n**ɛ**k/	[f**ɔ**seːn-n**ɛ**k]	‘charcoal’
	/ɔpoʃt**o**li+n**ɛ**k/	[ɔpost**o**li-n**ɔ**k]	‘apostolic’
	/m**a**ːrtiːr+n**ɛ**k/	[m**a**ːrtiːr-n**ɔ**k]	‘martyr’

	Two neutral vowels		
(7c)	/d**o**ktrine:r+n**ɛ**k/	[d**o**ktrine:r-n**ɛ**k]	‘doctrinaire’
	/k**ɔ**libe:r+n**ɛ**k/	[k**ɔ**libe:r-n**ɛ**k]	‘caliber’
	/b**o**ri:te:k+n**ɛ**k/	[b**o**ri:te:k-n**ɔ**k]	‘envelope’

The investigators furnish evidence that the count and height effects present in Hungarian vowel harmony are phonologically productive. When speakers were presented with novel stems in a written two-alternative forced choice task, in accumulation they replicated both effects: application of harmony to *–nɛk* depended on how distant the triggering back vowel was from the suffix, and on the height of the stem-final vowel. In particular, speakers matched to a significant extent the rates found in the corpus, as illustrated in Figure 5 below.

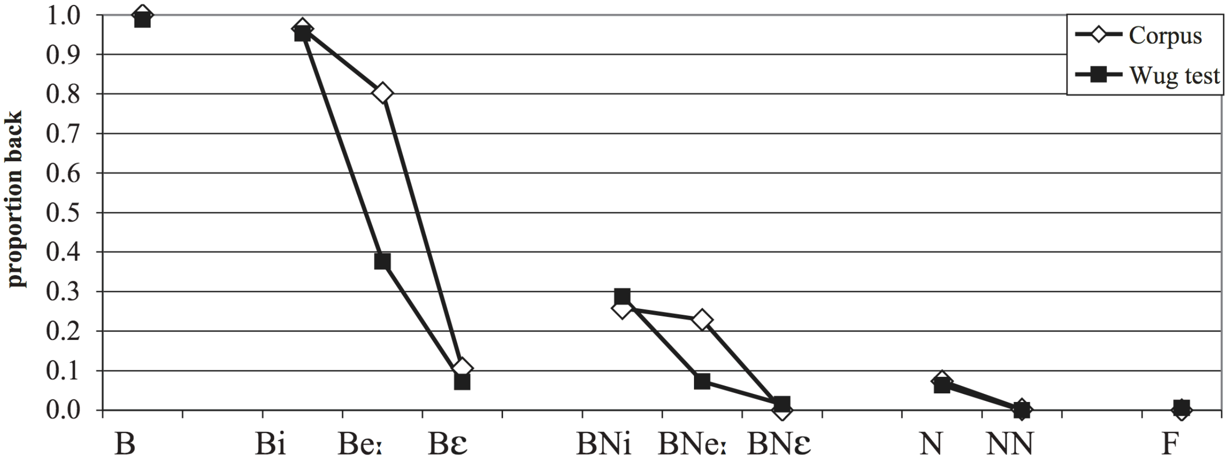


Figure 5: *frequency matching in wug test (B = back vowel; N = neutral vowel belonging to* [i e: ɛ]*; F = front rounded vowel) (Hayes, Zuraw et al. 2009, p. 832)*

The results suggest that Hungarian speakers are frequency matching to lexical trends in vowel harmony, and thus that these gradient effects need to be treated in the model of phonology.

Before moving forward, I mention here that not all experiments in the field have fully yielded the frequency matching result (Becker, Ketrez & Nevins 2011, Becker, Nevins & Levine

2012, Hayes, Zuraw et al. 2009, Hayes & White 2013, Jarosz 2017, Jarosz & Rysling 2017). In particular, Becker, Ketrez & Nevins (2011) found that while some regularities in subset of the Turkish lexicon displaying laryngeal alternations are productively extended in their nonce probe study, not all of them are. They attribute the incomplete learning of the lexicon to a deviation from frequency matching guided by Universal Grammar: though the phonologically well-motivated generalizations were productively extended by speakers, those presumed to be accidental generalizations within the lexicon were not. Moreover, Becker, Nevins & Levine (2012) found that while a laryngeal alternation in the English plural (*lea**f** ~ lea**v**es*) applies more regularly to attested monosyllables in the language than polysyllables, a wug test reveals no such preference (at least between monosyllables and disyllabic iambs, where final stress is held constant), while a series of artificial language learning studies of English speakers uncovers a universal bias towards protecting initial syllables. The results of the artificial language learning studies in particular suggest the presence of an initial syllable faithfulness bias in universal grammar, which thereby gives an explanation to the wug test results: frequency matching is inhibited in a case where the generalization runs counter to a universal bias. More generally, while frequency matching seems to be a general capability of language learners, these studies suggest that phonological generalizations running counter to formal or naturalness biases in learning occasionally go unlearned, or are learned relatively poorly.

2.3 Token variation and lexical propensities

Zuraw (2009) covers a case of variation in Tagalog in which a tapping rule exhibits not only lexical variation, but token variation too. Consider the data below:

Tapping: d → ɾ / V__V

(8a) *Categorical application*

[dunoŋ], 'knowledge'	[ma-ɾunoŋ], 'intelligent'
[dinig], 'heard'	[ma-ɾinig], 'to hear'
[dupok]	[ma-ɾupok], 'fragile'

(8b) *Categorical non-application*

[daʔig], 'beaten'	[ma-daʔig], 'beaten'
[dulas], 'slipperiness'?	[ma-dulas], 'slippery'
[daʔan] 'road'	[ma-daʔan-an], 'passable'

(8c) *Token variation*

[duŋis], 'dirt on face'	[ma-ɾuŋis ~ ma-duŋis], 'dirty (face)'
[dumi], 'dirt'	[ma-ɾumi ~ ma-dumi], 'dirty'

While tapping displays lexical variation across forms in (8a-b), it applies stochastically even across tokens of the same word in (8c). Zuraw (2009) plotted in a histogram counts of the different rates of the words in her corpus. The counts of application rates across words exhibit a *U-shaped distribution*: the great majority of words fall towards the poles (around 2,500 words that never tap, and 2,000 that always tap), but in fact over 400 words tap with rates between 2% and 98%. The following histogram excludes the ends of the scale:

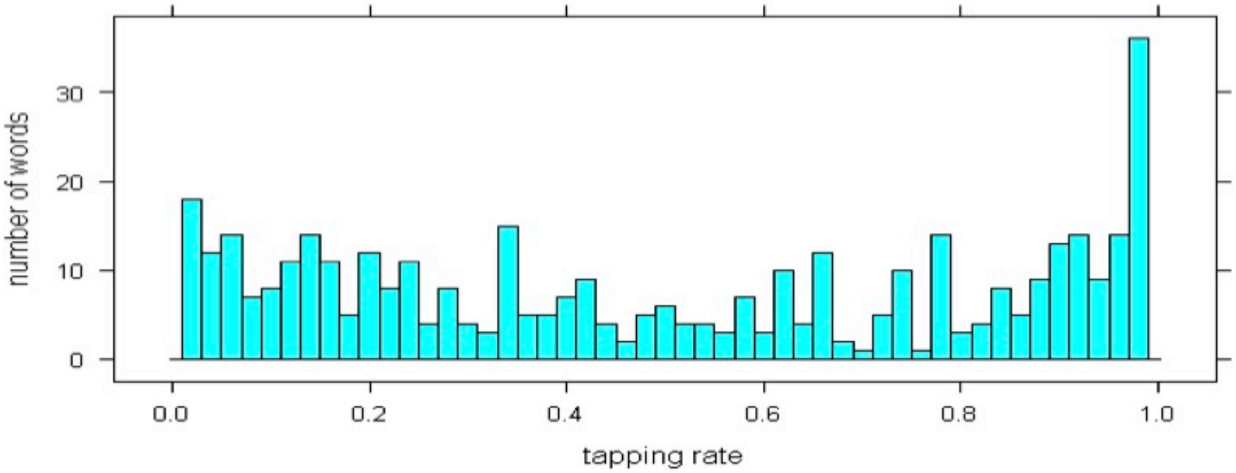


Figure 6: *histogram of medial tapping rates (Zuraw 2013, p. 9)*

If it were the case that token variation is encoded in the grammar as a rule applying at a fixed rate across words in the lexicon, then one might expect the distribution of tapping rates to resemble roughly a tall and thin bell curve, its mean centered around the fixed rate. For instance, if the rate of token variation were in proportion with the counts at the poles—2,500 untapped words and 2,000 tapped words—then one might expect the bell curve to be closely confined to and centered at a rate of 100*(2/4.5) = 45%. In fact, what is observed above is that words display lexical propensities: idiosyncratic rates that range fully between ends of the scale.

More recent research has likewise uncovered lexical propensities to participate in a phonological process. Linzen, Kasyanenko & Gouskova (2013), Gouskova & Linzen (2015) show that, in Russian prepositions showing a vowel-zero alternation, the identity of the root following preposition conditions the rate of application (and, even further, that some suffixes, when coming after these roots, regularize application rate—see Gouskova & Linzen 2015 in particular). Smith & Moore-Cantwell (2017) likewise find that in the English comparative paradigm, both phonological factors and the identity of the adjective play a part in determining whether that adjective takes *–er* as a suffix versus *more* in a paraphrastic construction (see

Section 2.3.1 below). Rosen (2016) shows that, in Japanese noun-noun compounds displaying variable rendaku, the identities of individual roots seem to influence the rate at which the process applies. In particular, both the identity of the first noun in the compound as well as that of the second play a part in determining the rate at which the compound as a whole undergoes the process (see also Rosen 2001, Irwin 2016). Tanaka (2017) similarly finds that the second root in compound surnames in Japanese influences the likelihood of rendaku in these names, in addition to phonological factors; moreover, the compound name as a whole can behave idiosyncratically, even after taking into consideration the phonological factors and the identity of the second root in such names.

Previous research into the modeling and learning of variation has focused on how speakers might come to learn lexical variation, in which a process applies stochastically across words, but categorically across tokens of the same word. But as the distribution shown above in Figures 6 suggests, the use of traditional phonological diacritics like [+Rule X] and [-Rule X], and their updated versions in OT, conceal large-scale stochastic systems in which individual morphemes are coded for specific levels on continuous probabilistic scales of propensity to trigger or undergo processes. The fact that these optionally tapped words in fact distribute across a spectrum to tap raises important questions: Do speakers have knowledge of idiosyncratic rates across morphemes? Provided they do, how is such knowledge acquired and represented?

2.3.1 Modeling lexical propensities and trends in Maximum Entropy Harmonic Grammar

Though the first question remains open, Moore-Cantwell & Pater (2016) make strides in answering the second, and in particular address how token fixedness and variation can be represented in phonological theory. Though current models of gradient productivity in language such as Maximum Entropy Grammar capture frequency matching in nonce word experiments, Moore-Cantwell and Pater observe that they fail to capture the fixed pronunciations of existing words in a language. Their MaxEnt-based approach captures fixed pronunciations in the lexicon but variable behavior and trend-matching in wug tests by including in the grammar lexically specific constraints along with analogous general constraints. This approach predicts an inverse correlation between the number of exceptions and the degree of productivity, which is borne in productivity studies and tendencies to regularization (Peperkamp *et al* 2010).

Suppose it were the case that English applied penultimate stress to half of its words, and antepenultimate stress otherwise. The authors first consider a MaxEnt approach to this pattern that lacks lexically specific constraints. Consider the tableaux below:

/bætækæ/	*P*	*H*	ALIGN-R *w* = 2	NONFIN *w* = 2
bə(ˈtækə)	50%	-2		-1
(ˈbætə)kə	50%	-2	-1	

Table 4a: *Variable stress without lexical specificity (Moore-Cantwell & Pater, p. 56)*

This model predicts that all like words share the same rate of penultimate stress. /bætækæ/ is essentially in free variation, with penultimate stress surfacing half the time and antepenultimate stress surfacing otherwise. In reality, though the lexicon as a whole displays variation across words with respect to stress location, words like *banána* and *Cánada* are fixed with respect to stress location. Thus the model below is a more accurate characterization, in which words are moderated by lexically specific constraints that mirror general constraints on stress:

	P	*H*	ALIGN-R-*banana* *w* = 5	NONFIN-*Canada* *w* = 5	ALIGN-R *w* = 2	NONFIN *w* = 2
bə(ˈnænə)	99%	-2				-1
(ˈbænə)kə	1%	-7	-1		-1	
kə(ˈnædə)	1%	-7		-1		-1
(ˈkænə)də	99%	-2			-1	

Table 4b: *Variable stress with lexically specific constraints (Moore Cantwell & Pater, p. 57)*

Here stress in the above words is primarily determined by weighted lexically specific constraints, which can model individual idiosyncratic propensities of particular words. Nonetheless, general constraints still receive positive weight, fitting general tendency across the lexicon.

Given certain parameter settings of their MaxEnt model, their approach of using general constraints for trends and lexically indexed constraints for idiosyncrasies generally succeeds in capturing Ernestus & Baayen (2013)'s corpus data on variation in Dutch voicing alternations. The general constraints *VTV (forbidding intervocalic voiceless consonants) and *VDV (forbidding intervocalic voiced consonants) as well as *VbV, *VdV, *VfV, etc. were used to fit to overall lexical trends, while lexically specific versions of *VTV and *VDV were used to regulate idiosyncratic rates within words. Batch Gradient Descent was used to learn appropriate weights for these constraints. As the table below reveals, the learning simulation eventually arrived at weights for these constraints that predicted rates that, to a significant degree, mimic those observed in the lexicon and experiment—though they further note that their model predicts rates that are relatively exaggerated towards the poles.

		% voiced		% voiced	%voiced
	Trend	Lexicon	no. forms	Experiment (EB)	Simulation
p/b	voiceless	9%	230	4%	**1%**
t/d	voiceless	25%	719	9%	**9%**
s/z	voiceless	33%	451	23%	**18%**
f/v	voiced	70%	166	49%	**84%**
x/ɣ	voiced	97%	131	80%	**99%**

Table 5: *Moore-Cantwell & Pater (2016)'s modeling results (p. 60) when trained on Ernestus & Baayen (2013)'s corpus data*

In sum, the general constraints cum lexically-indexed constraints approach has appeared to furnish a promising model of lexical trends and idiosyncrasy within MaxEnt. Moore-Cantwell and Pater themselves nevertheless note that changing the parameter settings, in particular those of the MaxEnt regularization term, can affect the outcome. Tanaka (2017), who also takes up Moore-Cantwell and Pater's approach to model lexical variation in Japanese surnames, further finds that the approach overfits lexical constraints to his dataset without a sufficiently strong regularization bias—a concern that we will explore further in Chapter 6. For other recent approaches to lexical trends and gradient idiosyncrasy, see Rosen (2016) on a Gradient Symbolic Computation account of Japanese rendaku, and Linzen, Kasyanenko & Gouskova (2013), Gouskova & Linzen (2015) on an account of the variable vowel-zero alternation in Russian prepositions using lexical scaling factors.

Smith & Moore-Cantwell (2017) cover lexical trends and idiosyncrasy in the morphological/paraphrastic alternation in English comparatives, and propose a MaxEnt model that closely resembles Moore-Cantwell & Pater (2016)'s approach. In their data, adjectives can be modified with *-er* (*happier*) or *more* (*more happy*), depending on phonological, frequency-related, and lexical factors. They observe that individual adjectives are idiosyncratic in their rates to take *–er* versus *more*—a striking example of lexical propensities. This is illustrated in the figure below:

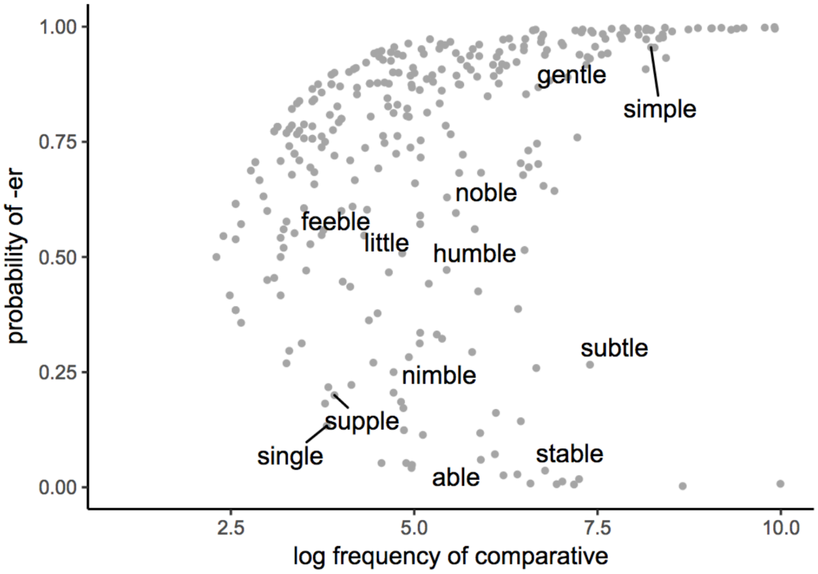


Figure 7: *idiosyncratic propensities for adjectives to take –er in the Corpus of Contemporary American English (COCA: Davies 2008) (Smith & Moore-Cantwell 2017, p. 5)*

In a corpus study, they first compare the results of a fixed effects logistic regression model containing a variety of phonological and frequency-related factors to a mixed effects logistic regression model, which includes adjective identity as a random intercept—i.e., a coefficient for the identity of each adjective. They find that the latter obtains a much better fit to the corpus data—hence lexical identity appears to condition variation, even when other factors are controlled for (including frequency). The purpose of this comparison was to assess whether lexical idiosyncrasy plays a role in conditioning variation (but not to propose mixed-effects logistic regression as a new model of idiosyncrasy and trend knowledge, as far as I am aware). As for the modeling and the learning theory, they posit a MaxEnt-based model that uses UR constraints (see Pater, Staubs, Jesney & Smith 2012, Smith 2015) to represent lexical idiosyncrasy, rather then lexically indexed constraints. UR constraints are situated within an online error-driven learner in which learning data are sampled according to lexical frequency, and UR constraints are induced only when needed, and decay when they are not used. The model predicts that high frequency lexical items are more likely to diverge from overall grammatical generalizations, reflecting findings that processing of novel expressions relies upon abstract knowledge, while reliance upon direct experience increases with increased exposure to an expression (Morgan & Levy 2016).

2.4 Summary and prospectus

We have seen that phonological processes can vary seemingly arbitrarily in whether or not they apply to individual words or morphemes. In addition to internalizing whether a variable process applies on a word-by-word basis, language learners can frequency match to trends observed over the entire lexicon, as has been observed and replicated repeatedly in nonce probe studies. Thus models of phonology and the lexicon must be able to learn and represent both lexical idiosyncrasy as well as gradient, probabilistic trends across the lexicon. A very recent spate of research finds that individual words or morphemes can even display distinct lexical propensities—idiosyncratic, gradient rates at which a word or morpheme participates in the variable process. It is an open question how pervasive lexical propensities are in the world's phonologies, and whether they are internalized by language learners. In the following sections, I take up these questions by investigating lexical variation in Slovenian palatalization and French liaison, building off of prior studies. It will be shown that models of the two processes are improved when factors encoding lexical propensities are included, even after controlling for other phonological and frequency-related factors (Chapters 3 and 4). Moreover, an experimental investigation into the intuitions of French speakers suggests that they acquire these idiosyncratic propensities to undergo liaison (Chapter 5). Recent MaxEnt-based models of lexical variation appear to have promise in capturing of lexical propensities together with lexical trends, but recent investigators have found that they only succeed under particular parameter settings. This potential weakness is explored in Chapter 6: learning simulations reveal that the MaxEnt approach eventually overfits lexical constraints to idiosyncrasies in the course of learning—a result that in fact appears to be general across parameter settings. In turn, the MaxEnt approach fails to capture broad, grammatical trends under broad assumptions. We explore an approach to lexical variation couched in a similar model, mixed-effects logistic regression, which circumvents the overfitting problem by privileging general constraints over lexical constraints in a way that that current MaxEnt models cannot. The mixed-effects model is found to capture both lexical propensities as well as broad trends across the lexicon.

Chapter 3:

Lexical propensities in Slovenian palatalization

We first investigate lexical idiosyncrasy as it occurs in Slovenian velar palatalization, whereby stem-final velars palatalize before a certain set of suffixes.

(9)	obla**k**-a	‘cloud’-GEN	obla**tʃ**-it͡sa	‘cloud’-DIM
	dow**g**-a	‘long’-GEN	dow**ʒ**-ina	‘length’

A large set of corpus data reveals that palatalization varies both on a morpheme-by-morpheme basis and on a token-by token basis (Jurgec 2016; *cf.* Bajec 2000, Toporišič 2001). Some suffixes essentially always trigger palatalization, while other suffixes never do. In (10a), for example, the stem /dowg/, ‘long’ always palatalizes before the nominalizing suffix /-ina/, but it never does before the suffix /-in/, ‘tall’. A suffix’s status as a palatalization trigger cannot merely be reduced to whether it begins with a front vocoid, as suffixes like /-k/ and /-n/ can also trigger palatalization, while front vocoid-initial suffixes like /-in/ and /-i/ do not.

(10a) **Lexical variation**

Some suffixes trigger palatalization, while others do not

Stem		Triggers		Non-triggers	
dow**g**-a	‘long’-GEN	dow**ʒ**-ina	‘length’	dow**g**-in	‘tall male’
du**x**	‘smell, ghost’	du**ʃ**-k-a	‘breath-DIM-GEN	du**x**-i	‘smell’-PL
baro**k**	‘baroque’	baro**tʃ**-n-i	‘baroque’-ADJ-DEF	baro**k**-ist	‘baroque’-PER

Certain suffixes trigger palatalization across only some eligible stems, but not others, as in (10b). For example, stems such as /oblak/, ‘cloud’ obligatorily palatalize before /-it͡sa/, while stems such as /kokoʃk/, ‘hen’ obligatorily do not.

(10b) **Lexical variation**

Before suffixes that tend to trigger palat'n, some stems undergo it while others do not

Stem		Stem before diminutive -it͡sa		Status
obla**k**-a	'cloud'-GEN	obla**tʃ**-it͡sa	'cloud'-DIM	*Undergoer*
ʋra**g**-a	'devil'-GEN	ʋra**ʒ**-it͡sa	'devil'-DIM	*Undergoer*
pe**g**-a	'spot'-GEN	pe**g**-it͡sa	'spot'-DIM	*Non-undergoer*

Finally, particular stems can vary in whether or not they undergo palatalization before certain suffixes — that is, these stems vary on a token-by-token basis, as in (10c). For example, /nɔg/ 'leg' can surface faithfully *or* as palatalized when it comes before /-it͡sa/.

(10c) **Token variation**

Stem		Stem before suffixes tending to trigger palat'n	
nɔ**g**-a	'leg'-GEN	nɔ**g**-it͡sa ~ nɔ**ʒ**-it͡sa	'leg'-DIM
bɾe**g**-a	'river bank'-GEN	bɾe**g**-nat ~ bɾe**ʒ**-nat	'river bank'-ADJ
gra**x**-a	'pea'-GEN	gra**x**-k-a ~ gra**ʃ**-k-a	'pea'-DIM-GEN

This section investigates corpus data on Slovenian palatalization, reaching the following conclusions: (i) variable palatalization is conditioned phonologically, confirming Jurgec (2016)'s corpus investigation; (ii) nevertheless, we find that different morphemes gradiently participate in the process: a statistical model of corpus data that encode lexical propensities on suffixes (e.g., [0.7 Palatalization]) outperforms one that encodes suffixes merely as ([+ Palatalization]) (*cf.* Jurgec 2016); (iii) further modeling investigation reveals that propensities are associated both with triggering suffixes and undergoing stems.

3.1 Review of Jurgec (2016)

Jurgec (2016) conducted a corpus study of Slovenian palatalization, finding that both lexical and phonological factors affect whether palatalization applies to a particular stem-suffix pair. He provides an analysis the data couched in Maximum Entropy Harmonic Grammar; the primary objective was to capture the phonological factors conditioning variation, but, as I will argue, he does not adequately capture the degree to which lexical factors conditions variation.

Jurgec extracted data from the *Dictionary of Standard Slovenian* (Bajec 2000) and the *Slovenian Orthographic Dictionary* (Toporišič 2001), two online dictionaries that contain 110,000 and 130,000 word types, respectively. In particular, he extracted stems ending in velars that were followed by any one of several suffixes noted by Toporišič (2001) to trigger palatalization. To obtain token rates of palatalization associated with the different words, he then looked each word up in the *Gigafida* (Logar-Berginc et al. 2012), a text corpus containing 1.2 billion tokens from a variety of written sources, published between 1990–2011. Investigating variable palatalization in written data is possible because palatalization is reflected in orthography (e.g., [obla**k**-a], <*obla**k**a*>; [obla**tʃ**-iʦa], <*obla**č**ica*>).

Jurgec primarily focused on 9 commonly occurring suffixes in the corpora, obtaining that each of them triggered palatalization to some degree. He provides the table below. Before the diminutive suffix /-ʦ/, for example, 32 out of 86 velar-final stems obtained from the dictionary underwent palatalization in more than 50% of the tokens extracted from the *Gigafida*; across all stems, there were roughly 7,500 tokens that underwent palatalization before /-ʦ/ out of roughly 38,100. /-k/, on the other hand, triggered palatalization the majority of the time in 81 out of 91 stems, and in 292,100 tokens out of 300,800 overall.

	-/ʦ/	-/k/	-/n/	-/itʃ/	-/iʦa/	-/ina/	-/je/	-/nat/	-/oʋje/
Number of stem types	86	92	169	20	107	36	59	17	26
> 50% palatalized across stem tokens	32	81	151	20	49	34	58	12	10
Number of tokens	38.1	300.8	3916.3	63.5	313.6	840.9	174.8	3.4	4.4
(in 1000s) palatalized	7.5	292.1	3233.9	63.4	242.2	840.1	174.8	2.2	0.8

Table 6: *Jurgec's table showing suffix-specific rates of palatalization (Jurgec 2016, p. 7)*

Jurgec addressed less thoroughly the role of the stem in conditioning variation. He provided a small handful of stems that appear to undergo palatalization at different token rates before the same suffix. Before the diminutive suffix /-iʦa/, for example, /sux/, ‘dry’ palatalized at a rate of 65% across tokens; /smrek/, ‘spruce’ palatalized at a rate of 85%; and /oblak/, ‘cloud’ palatalized in all tokens.

Jurgec finds that phonological factors can affect whether a stem-suffix pair undergoes palatalization. For example, though some suffixes that do not begin with a front vocoid can trigger palatalization, its presence is associated with higher overall rates of palatalization. And though palatalization can apply to stem-final velars of all kinds, it targets the stops *k* and *g* more regularly than it does the fricative *x*. In addition, palatalization applies regularly if surfacing faithfully would otherwise produce a geminate: *k/g*-final stems palatalize nearly categorically before diminutive *–k*. Finally, palatalization is gradiently blocked by velars occurring earlier in the stem, and is categorically blocked by postalveolars occurring earlier in the stem.

Jurgec ran a linear mixed effects model on his corpus data, which included consonant identity and suffix as fixed factors, and stem identity as random intercept and slope. The model results reveal that a variety of suffixes trigger at rates distinct from –ʦ, even when consonant identity is controlled for; moreover, based on their *t*-values and estimates, it would seem that the suffixes are associated with distinct propensities to trigger palatalization.

Jurgec accounts for some of the variation within Maximum Entropy Harmonic Grammar (Smolensky 1986, Goldwater & Johnson 2003, Hayes & Wilson 2008), using the MaxEnt Grammar Tool (http://linguistics.ucla.edu/people/hayes/MaxentGrammarTool/; Wilson & George 2009). In particular, he accounts primarily for the phonological factors that affect palatalization. To account for the *lexical* factors conditioning variation, the nine suffixes analyzed were merely encoded with as [+ Palatalization]. Such machinery separates the nine suffixes that can trigger palatalization *whatsoever* from those that *never* trigger it, but does not distinguish between idiosyncratic rates of any of the palatalizing suffixes (suggested by Table 6), and does not account for any potential stem-specific idiosyncrasies either (see Section 3.2.2). The precise account of lexical variation was left for future investigation.

The central constraint driving palatalization in Jurgec's analysis is PAL/_{palatalizing suffixes}, with IDENT violated by palatalized forms. To account for the tendency for stems to palatalize before front vocoid-initial suffixes, PAL/_{i j} is also used. Geminate avoidance in palatalization is mediated by $*C_iC_i$, while the long-distance effects are mediated by *POSTALV … POSTALV$_{stem}$, *VEL … POSTALV$_{stem}$, and *POSTALV … ALVAFFR$_{stem}$. The corpus frequencies and violation profiles were fed into the MaxEnt learning tool, which outputted constraint weights that maximized fit to the frequencies given the inputted constraints.

Some aspects of Jurgec's model were successful: for example, the combined effect of PAL/_{palatalizing suffixes} and PAL/_{i j} resulted in the greater tendency for front vocoid-initial suffixes to trigger greater rates of palatalization. In the tableau below, the faithful candidate violates both PAL/_{palatalizing suffixes} and PAL/_{i j}; their weights sum up to 5.0, and so the Harmony of the faithful candidate is -5.0. The palatalized candidate only violates IDENT, which has a weight of 0, and so its Harmony is -0.0. Since the Harmony of the palatalized candidate is much higher than that of the faithful candidate (- 0.0 > - 5.0), the model predicts that the palatalized candidate surfaces the vast majority of the time.

/breg-ina/	Observed rate	Predicted rate	Harmony	PAL/_{i j} $w = 3.4$	PAL/_{palat. suffix} $w = 1.6$	IDENT $w = 0$
bregina	0%	0%	-5.0	- 3.4	- 1.6	
breʒina	100%	100%	-0.0			- 0

Table 7a: *tableau showing palatalization with –ina (Jurgec 2016, p. 23)*

According to Jurgec, MaxEnt valued the weight of IDENT so low because palatalization generally obtains in the data, and environments where it does not are better explained by the other phonological factors (e.g., *POSTALV … POSTALV$_{stem}$). Problematically, this results in a mediocre fit to forms with suffixes that do not begin with a front vocoid, i.e., those forms in which palatalization rate may be regulated by lexical factors such as stem or suffix identity. In the tableaux below, palatalization is predicted to take place to tokens of /breg-nat/ far more regularly than the observed rate.

/breg-nat/	Observed rate	Predicted rate	Harmony	PAL/_{i j} *w* = 3.4	PAL/_{palat. suffix} *w* = 1.6	IDENT *w* = 0
bregnat	50%	17%	-1.6		- 1.6	
breʒnat	50%	83%	-0.0			- 0

Table 7b: *tableau showing palatalization with –nat (Jurgec 2016, p. 23)*

Jurgec states that with these constraints alone the fit to the data is imperfect—presumably due to the *lexical idiosyncrasies in the behavior of individual morphemes*, which he leaves to further research. In this section, I expand upon Jurgec's pioneering study of variation in Slovenian palatalization. The aim is to test whether the system would be accounted for more comprehensively by encoding the 9 suffixes on a spectrum to trigger palatalization (e.g., [0.7 Palatalization])—rather than with [+Palatalization]—and by encoding individual stems on a spectrum as well.

3.2 Corpus investigation into Slovenian palatalization

Following Jurgec, words consisting of velar-final stems and one of the nine palatalizing suffixes were extracted from the *Dictionary of Standard Slovenian*. Words were annotated for what suffix they ended in, and whether or not they underwent palatalization. Recall that Jurgec fed each word obtained from the dictionary into the *Gigafida*, the massive written text corpus, to obtain token palatalization rates for each word. I departed from Jurgec's methods by not simply feeding each word from the dictionary into *Gigafida*; rather, I concatenated each stem extracted from the dictionary with each of the nine suffixes, creating hypothetical stem-suffix pairs that may or may not be attested in the *Gigafida*; then, to obtain token rates of palatalization for these pairs, I fed each of them into the *Gigafida.* This yielded a data set containing about 3,000,000 tokens of stem-suffix pairs that either did or did not undergo palatalization.

With these token rates, I calculated the average palatalization rates for each suffix—for any given suffix -*x*, I calculated for each stem *s* the token palatalization rate of *s*-*x* (obtained from the *Gigafida*), and then averaged over these rates. For example, the stem /ag/ palatalizes before suffix /-je/ 22% of the time in my data, but /kak/ palatalizes before it 99% of the time; averaging over stems, the average palatalization rate of /-je/ is 85%. The rates are given in the table below. We see that suffixes pattern along an **entire palatalization propensity spectrum**: /-ʦ/, for example, triggers palatalization across 42% of the stems that precede it; /-iʦa/ triggers the process after about 70% of them stems; and /-k/ triggers it after about 95% of the stems. The rates in the figure below resemble those obtained by Jurgec.

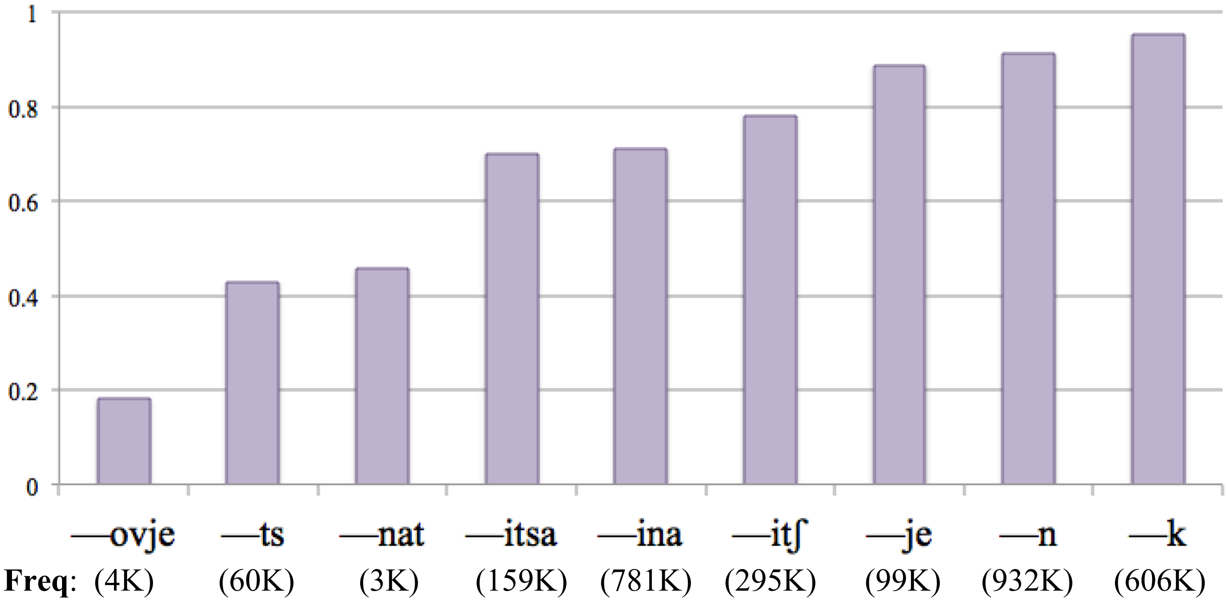


Figure 8a: *palatalization rates for different suffixes, averaged across stems*

Here I show that the stems in my data distribute across the propensity spectrum too. I extracted 260 stems that had an allomorph occurring before at least four suffixes, and calculated their average palatalization rate: for any given stem *s*, I calculated for each suffix -*x* the token palatalization rate of *s*-*x*, and then averaged over these rates. Provided in the histogram below is the number of stems associated with a particular average palatalization rate. 113 stems are clear palatalizers, having palatalized in the vast majority of tokens. 19 stems were non-palatalizers, having palatalized in nearly none of the tokens. Finally, a healthy minority of stems—128 in total—fall between extreme rates. This suggests that **stems are coded on an entire spectrum to palatalize**, just as suffixes are.

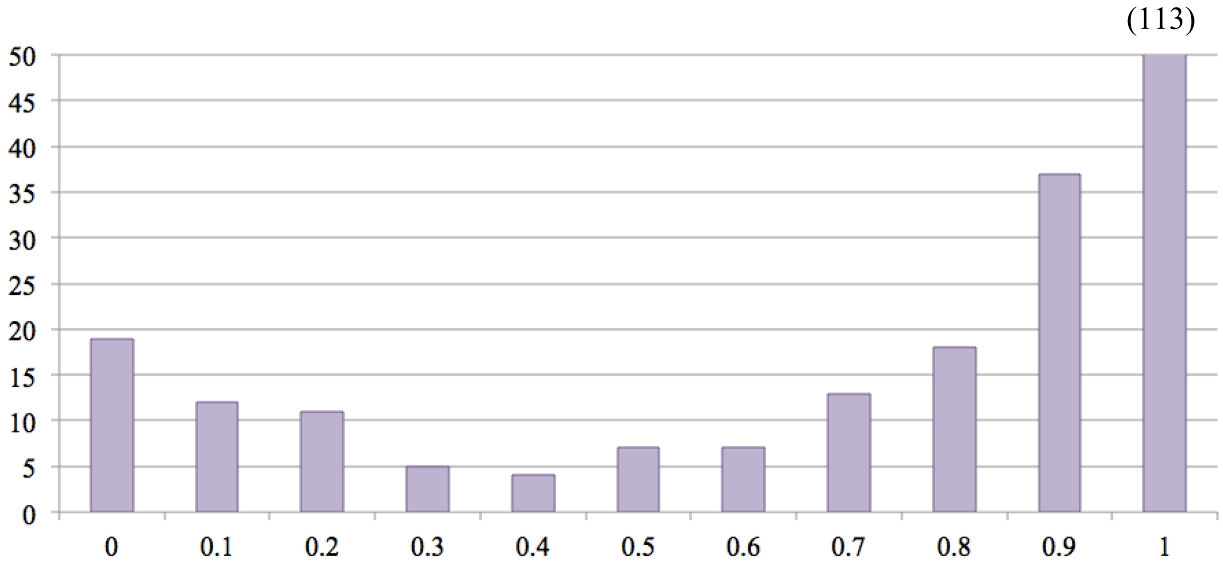


Figure 8b: *Histogram of stem palatalization rate frequencies*

The investigation so far suggests that variation in Slovenian palatalization must be accounted for by coding individual morphemes on an entire propensity spectrum. We now compare a series of mixed logistic regression models to further demonstrate the value of lexical propensities for modeling phonological variation.

3.2.1 Binary versus gradient palatalization triggers: a comparison of logistic regression models

The `glmer` function of the *lme4* package (Bates & Maechler 2011) in R (R Core Development Team 2014) was used to fit various logistic regression models to the palatalization data. We first compare models of the palatalization triggers—namely, the nine suffixes studied in Jurgec (2016). We run a *Baseline Model*, which incorporates Jurgec's phonological factors as well as whole word (stem + suffix) as random intercept, but treats the nine suffixes as having the same triggering status ([+ Palatalization]). We compare the Baseline Model to a second model, the *Suffix Propensity Model*, which contains Jurgec's phonological factors but which also allows the different suffixes to take on different propensities to trigger palatalization ([0.7 Palatalization]). In particular, the Baseline Model includes the following factors proposed in Jurgec (2016): stem-final consonant identity, whether the suffix begins with a front vocoid, whether the stem contains a post-alveolar distant from the target, whether the suffix contains a post-alveolar affricate; I also add to this a factor for log word frequency. The Baseline Model lacked the distant velar factor since had a rather small effect in Jurgec's study; moreover, an examination of my corpus revealed that palatalization rates were no lower with distant velars than without. Moreover, investigation of consonant identities in my data reveals that *g* undergoes palatalization less overall than *k*—70% versus 85%, respectively; *x* undergoes palatalization at an 80% rate, but *x*-final stems constituted a small minority of the stems overall. As Jurgec points out, the distinction between *k* and *g* is not surprising, considering that while *k* changes only in place, *g* changes in both place and continuancy:

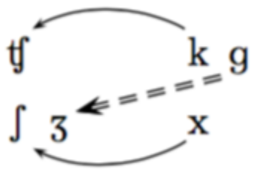


Figure 9: *Graph of mappings (Jurgec 2016, p. 13)*

The Baseline Model serves as a null hypothesis regarding whether different suffixes condition palatalization at different rates: the model regresses only over suffixes noted by Toporišič (2001) and Jurgec (2016) to condition palatalization, and so if a model that can encode suffix-specific rates performs better than the baseline, then that suggests that these palatalization-triggering suffixes in fact trigger the process at different rates. A now standard way to encode item-specific idiosyncrasies in statistical models is to use mixed effects logistic regression (Fruehwald 2012, Shih & Inkelas 2016, Zuraw & Hayes 2017, Smith & Moore-Cantwell 2017, Shih 2018). Broad statistical generalizations can be captured together with item-specific idiosyncrasies through encoding general constraints as fixed effects and item-specific idiosyncrasies as a random intercept. As we will see in Chapter 6, this model of will be of great significance to the theory of lexical variation and its learning. For now we simply use the mixed model to assess whether lexically specific effects are present in the system (Zuraw & Hayes 2017, Smith & Moore-Cantwell 2017). The Suffix Propensity Model contains all factors in the Baseline Model, plus an additional factor for suffix identity, coded as a random intercept—that is to say, every individual suffix (9 in total) is allowed to be associated with an idiosyncratic propensity to trigger palatalization. If the Suffix Propensity Model outperforms the baseline, then that would suggest that suffixes trigger palatalization at distinct rates, even after other phonological factors proposed to condition palatalization are controlled for.

Any row in the dataset consisted of: a token of a stem+suffix pair; whether the token underwent palatalization; the natural logarithm of word frequency; the identity of the stem; the identity of the suffix; the identity of the stem-final consonant (*k, g, x*); whether suffix begins with a front vocoid; whether the input contained a velar geminate[1]; whether the output contained a palatalized geminate; whether stem contains a post-alveolar distant from the target; and whether the suffix contains a post-alveolar affricate.

The Baseline Model contained six factors, all coded as main effects: `logwf`, `cons`, `gem`, `frontvocoid`, `S...S]`, and `k...S]`, and `[...ts`, defined below:

[1] None of the outputs contained a {k, g}+k geminate, as –k always undergoes *yer*-insertion (Jurgec 2016). The insensitivity of the geminate constraint to yers in Slovenian is a topic I leave for further research.

(11) `logwf`: log of the frequency of stem+suffix pair

`cons`: identity of the stem-final consonant (*k*, *g*, *x*)

`kk`: 1 if the stem ended in *k* or *g* and the suffix was –*k*; 0 otherwise.

`tStS`: 1 if the stem ended in *tʃ*{*k*, *g*}; 0 otherwise.

`frontvocoid`: 1 if the suffix begins with a front vocoid (*i*, *j*); 0 otherwise.

`S...S]`: 1 if the stem contains a postalveolar that is not adjacent to the target; 0 otherwise.

`[...ts`: 1 if the suffix contains an alveolar affricate (–/iʦa/, –/ʦ/); 0 otherwise.

The dataset consists of stem+suffix *tokens*, yet we wish for the model to treat each whole word as an observation, rather than word tokens themselves. Hence we also encode in the Baseline Model, as well as in all subsequent models tested, whole word as a random intercept.

The results of the Baseline Model are given below. The factors were compared against a reference intercept group of whole words that had a *g* as stem-final consonant, had a log(word frequency) of 0, lacked an input geminate, had a suffix that did not begin with a front vocoid or contain an alveolar affricate, and had a stem that did not contain an earlier postalveolar. The results given below indicate that the factors given above are mostly significant predictors of palatalization, in line with Jurgec (2016), with the exception of `frontvocoid`. Consonant identity significantly influences palatalization rate, with `consk` receiving a positive coefficient relative to baseline `consg`. The coefficients for the geminate constraints, `S...S]` and `[...ts` were all in the direction predicted by Jurgec: we find higher palatalization rates before –*k*, suggesting that the process assists in geminate avoidance; distant sequences of postalveolar sequences are also avoided; and morphemes with *ʦ* trigger at lower rates. Moreover, word frequency exerts a small, negative effect, though significant.

```
Random effects:
 Groups Name        Variance Std.Dev.
 word   (Intercept) 101.9    10.1
Number of obs: 2940918, groups:  word, 4822

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept:    3.95       0.47      8.29  <0.001   ***
 consg)

logwf         -0.08       0.03     -2.67   0.007   **

consx          1.59       0.66      2.41   0.015   *
consk          1.96       0.47      4.11  <0.001   ***

kk             4.94       0.80      6.12  <0.001   ***
tStS          -4.53       1.96     -2.31   0.020   *

frontvocoid   -0.52       0.39     -1.32   0.183

S...S         -1.67       0.78     -2.12   0.033   *

[...ts        -3.53       0.46     -7.55  <0.001   ***
```

Table 8: *Baseline Model results for Slovenian palatalization*

The Akaike Information Criterion (AIC; Akaike 1973; *cf*. Kullback-Leibler 1951) scores models based on fit to the data and number of parameters, with a lower score being better. See Bolker et al. (2009) for discussion on and justification for using the AIC to compare mixed logistic regression models to assess whether a random intercept is to be included. The Baseline Model's AIC value is **8767.8**.

The Baseline Model assumes that the nine different suffixes are equal in their propensity to trigger palatalization—that is to say, they are all simply [+Palatalization]. We thus compare its performance to that of the Suffix Propensity Model, which contains suffix identity as a random intercept. The results are presented in output below. The results are similar to those of the Baseline Model, except for the following: `S...S]` is now a weak trend, and `[...ts` is no longer a significant predictor—probably because the random intercept encoding suffix identity subsumes this factor. The variance of the suffix intercept is far from zero, suggesting that much of the variation not explained by the main effects or the random intercept for whole word can be explained by the suffix identities.

```
Random effects:
 Groups Name        Variance Std.Dev.
 word   (Intercept) 95.55    9.77
 suffix (Intercept) 35.69    5.97
Number of obs: 2940918, groups:  word, 4822; suffix, 9

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept:    0.11       2.92      0.03   0.969
 consg)

logwf         -0.10       0.03     -3.09   0.001    **

consx          2.15       0.70      3.05   0.002    **
consk          1.92       0.48      3.99  <0.001    ***

kk             7.53       1.47      5.08  <0.001    ***
tStS          -4.51       2.01     -2.24   0.024    *

frontvocoid    2.98       4.05      0.73   0.462

S...S]        -1.43       0.79     -1.80   0.070    .

[...ts        -1.90       4.82     -0.39   0.692
```

Table 9: *Suffix Propensity Model results for Slovenian palatalization*

While the Baseline Model's AIC is **8767.8**, the Suffix Propensity Model's AIC value is **8283.7**—a substantial reduction of about 500 points. Between any two models *A* and *B* of the same dataset, *B* outperforms *A* if its AIC is lower by at least 10 points (Burnham & Anderson 2004). Moreover, a likelihood ratio test between the two models suggests that the Suffix Propensity Model substantially outperforms the Baseline Model ($p < 0.001$). The results suggest that these suffixes indeed trigger palatalization at different rates—even after controlling for all of Jurgec's factors—thereby corroborating Jurgec’s observations about suffix behavior. These results suggest that individual morphemes are encoded on a spectrum—that is, with lexical propensities—rather than on a binary scale.

3.2.2 Extending the model: gradient palatalization undergoers

Does encoding undergoing stems improve model performance relative to the baseline? We first compare the performances the Baseline Model against the *Stem Propensity Model*, which encodes stem identity as a random intercept, allowing for each stem (2,720 in total) to be associated with different propensities.

Given below are the results for the *Stem Propensity Model*. The results are similar to the Suffix Propensity Model, except now `S...S]` is no longer a significant predictor contra Jurgec (2016)—this is to say, distant postalveolars do not appear to condition rate once we take into consideration lexical idiosyncrasies of both stems and suffixes. Furthermore, `[...ts` is significant, in particular because this model does not encode suffix propensities, the random effect that subsumed this main effect in the Suffix Propensity Model.

```
Random effects:
 Groups Name        Variance Std.Dev.
 word   (Intercept) 48.22    6.94
 stem   (Intercept) 78.24    8.84
Number of obs: 2940918, groups:  word, 4822; stem, 2720

Fixed effects:
              Estimate  Std. Error z value Pr(>|z|)
(Intercept:     4.08       0.65      6.27  <0.001   ***
 consg)

logwf          -0.09       0.03     -2.84   0.004   **

consx           2.00       1.00      2.00   0.045   *
consk           2.63       0.70      3.73  <0.001   ***

kk              6.09       0.77      7.86  <0.001   ***
tStS           -4.49       2.19     -2.05   0.040   *

frontvocoid     0.40       0.37      1.08   0.280

S...S]         -1.40       1.14     -1.23   0.217

[...ts         -2.82       0.41     -6.82  <0.001   ***
```

Table 10: *Stem Model results for Slovenian palatalization*

The Stem Propensity Model's AIC value is **8128.9**. This is a substantial improvement over the Baseline Model, whose AIC value is **8767.8**—a drop of about 640 points. A likelihood ratio test confirms that the Stem Propensity Model is superior (p < 0.001).

Does encoding undergoing stems *as well as* triggering suffixes with propensities improve model performance beyond the three prior models? We compare the performances the Suffix Propensity Model the *Stem and Suffix Propensity Model*, which encodes stem identity and suffix identity as distinct random intercepts, allowing for each stem and suffix to be associated with different propensities. In particular, the Stem and Suffix Propensity Model contains all factors in the Baseline Model—in particular, Jurgec's phonological factors—plus two factors for stem identity and suffix identity, both coded as random intercepts. Here, every suffix and every stem can be associated with idiosyncratic propensities to trigger and undergo palatalization, respectively.

Given below are the results for the *Stem and Suffix Propensity Model*. The results are similar to the *Suffix Propensity Model*, except now `S...S]` is no longer a significant predictor contra Jurgec (2016)—this is to say, distant postalveolars do not appear to condition rate once we take into consideration lexical idiosyncrasies of both stems and suffixes. Moreover, the intercept is not significant—under the conditions defined by the reference level, the baseline rate does not depart substantially from chance rate. This presumably would not be the case if we were to regress over a broader dataset that, for example, also included suffixes that *never* trigger palatalization—inclusion of such suffixes would reduce the baseline rate of palatalization across the data.

```
Random effects:
 Groups Name        Variance Std.Dev.
 word   (Intercept) 49.40    7.02
 stem   (Intercept) 68.06    8.25
 suffix (Intercept) 19.54    4.42
Number obs: 2940918, groups:  word, 4822; stem, 2720; suffix, 9

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept:     1.15       2.24      0.51   0.608
 consg)

logwf          -0.10       0.03     -3.22   0.001    **

consx           2.36       1.00      2.35   0.018    *
consk           2.59       0.69      3.75  <0.001    ***

kk              7.94       1.32      6.01  <0.001    ***
tStS           -4.58       2.16     -2.11   0.034    *

frontvocoid     2.72       3.01      0.90   0.366
S...S          -1.20       1.12     -1.06   0.284
morph.with.ts  -1.88663    3.58     -0.52   0.598
```

Table 11: *Stem+Suffix Model results for Slovenian palatalization*

The Stem and Suffix Propensity Model's AIC value is **7801.5**. We can compare the AIC's of all three models measured thus far:

(12)	Baseline Model AIC:	**8767.8**
	Suffix Propensity Model AIC:	**8283.7**
	Stem Propensity Model AIC:	**8128.9**
	Stem and Suffix Propensity Model AIC:	**7801.5**

The Stem and Suffix Propensity Model substantially outperforms both the Baseline Model, the Suffix Propensity Model, and the Stem Propensity Model—a series of likelihood ratio tests confirms these facts (all yielding $p < 0.001$)—which strikingly suggests that both triggering suffixes *and* undergoing stems are associated with idiosyncratic lexical propensities to participate in palatalization. The coefficients of the levels of the random intercepts for suffix and stem further reveal morphemic gradience, as the tables below illustrate:

Suffix	Rate
-ovje	-4.05
-ina	-1.27
-nat	-0.40
-itʃ	-0.38
-ts	-0.16
-itsa	0.16
-k	0.58
-je	1.48
-n	4.03

Stems (sample)	Rate
trak-	-5.34
tramik-	0.00
transcendenk-	0.05
tradicionalistik-	0.55
tragikomik-	1.14
travmatik-	1.30
tragik-	2.31

Table 12: *coefficients for stems and suffixes in Slovenian*

Note that the Stem Propensity Model's AIC value is lower than the Suffix Propensity Model's AIC value; moreover, in the Stem and Suffix Propensity Model, the variance of the random intercept for stem ($\sigma^2 = 68.06$) is greater than that of the random intercept for suffix

(σ^2 = 19.54). This indicates that undergoing stems explain more variance in the dataset than triggering suffixes do—over three times as much. Though further crosslinguistic investigation must be undertaken to confirm whether the following hypothesis is plausible, there may exist a bias such that undergoing morphemes across languages explain more lexical variation in the relevant paradigms than triggering morphemes do. How this bias should be implemented if it does exist, and what the implications of this bias are for the perceiver-learner, are left to further research.

To ensure that our model is making reasonable predictions, we assess whether the Stem and Suffix Propensity Model is in fact predicting the rates of significant phonological conditioners of palatalization as well as the propensities at which different morphemes trigger or undergo palatalization. Below, I show that the rates of the two significant phonological trends—consonant identity and geminate avoidance—are predicted by the model. I used the `predict` function in the *lme4* package in R to obtain rates under different phonological conditions. For the consonant identity trend, the model predicts that *k*-final stems palatalize at nearly a 95% rate in words with mean frequency (log(word frequency) = 9.5), other phonological factors notwithstanding; on the other hand, the model predicts that *g*-final stems palatalize at roughly a 50% rate under those same conditions. The corpus gives a 99% rate of palatalization to words with *k*-final stems and with log(word frequency) between 9 and 10, other phonological factors notwithstanding; moreover, the corpus gives a 52% palatalization rate to words with *g*-final stems that otherwise satisfy the same conditions. In addition, the model-predicted rates of geminate avoidance in words of mean frequency also the analogous rates in the corpus.

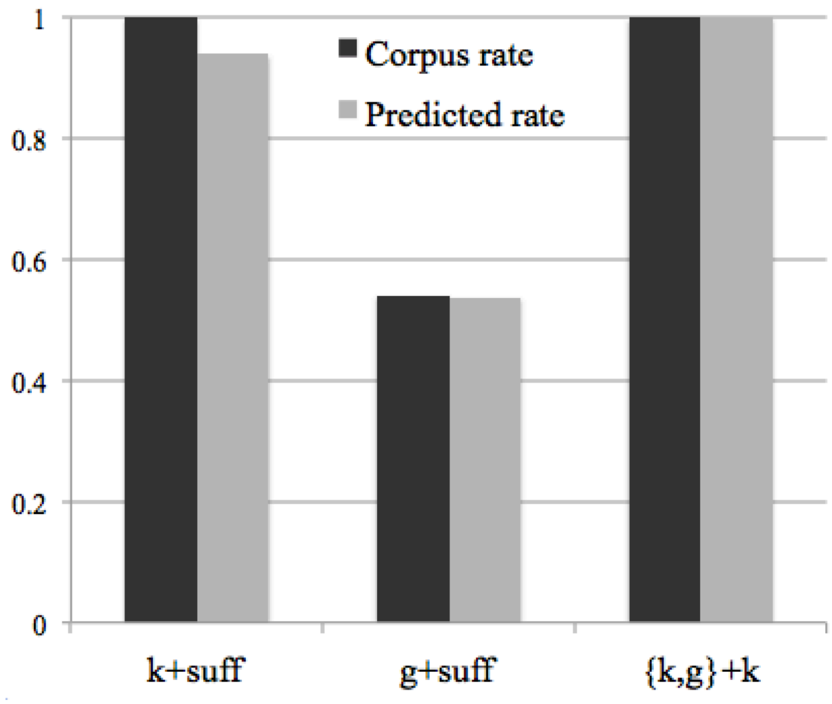
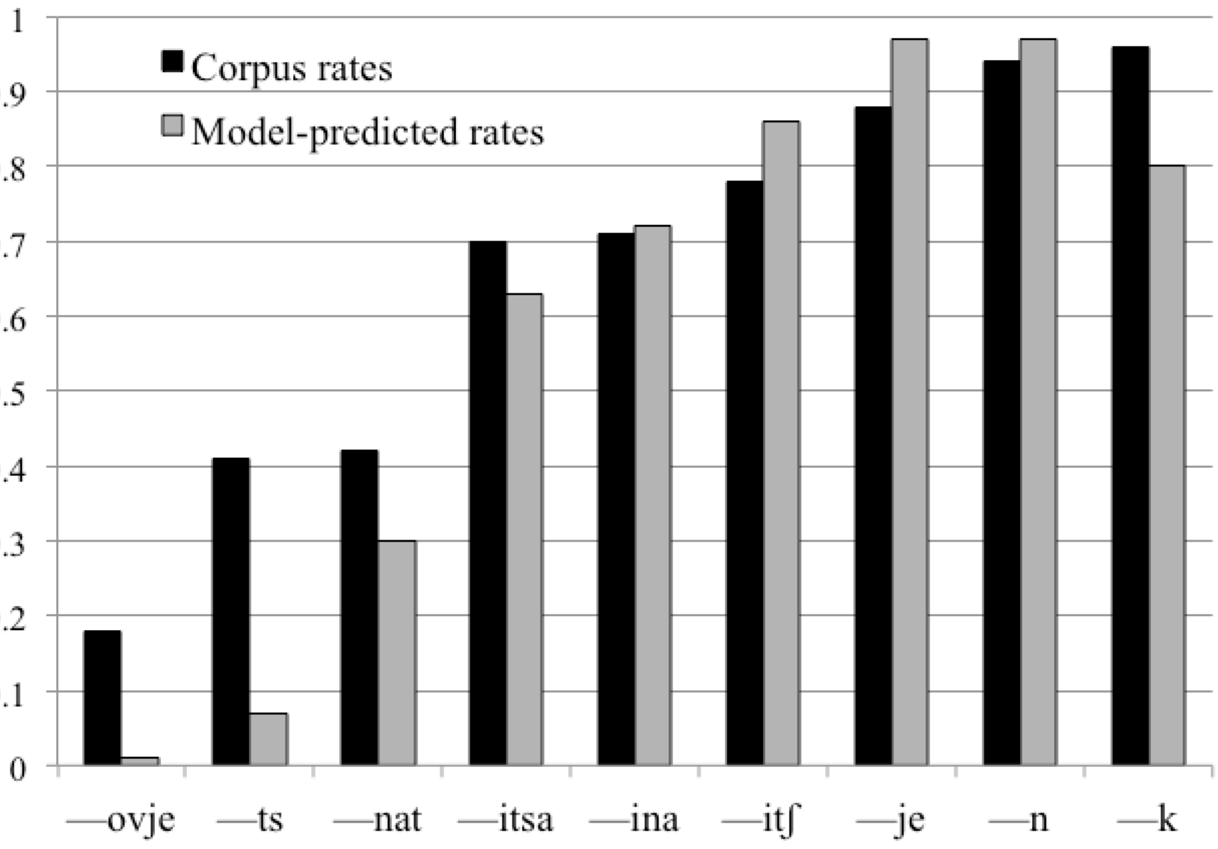


Figure 10a: *model succeeds in predicting phonological trends*

Finally, the model makes good predictions about the propensities of the nine different suffixes, generally fitting to the average palatalization rates for each suffix given in the corpus, as illustrated in the table and figure below:

Suffix:	-ovje	-ʦ	-nat	-iʦa	-ina	-itʃ	-je	-n	-k
Average rate across stems:	18%	41%	42%	70%	71%	78%	88%	94%	96%
Predicted rate:	1%	7%	30%	63%	72%	86%	97%	97%	80%

Table 13: *model-predicted suffix rates generally match corpus rates*



Freq: (4K) (60K) (3K) (159K) (781K) (295K) (99K) (932K) (606K)

Figure 10b: *model-predicted suffix rates generally match corpus rates*

3.3. Discussion and summary

Overall, these findings strongly favor theories that encode morphemes' participation on a spectrum (e.g., [0.7 Rule X]; *cf.* Moore-Cantwell & Pater 2016, Smolensky & Goldrick 2016, Zuraw 2016, Zuraw & Hayes 2017), and disfavor theories that encode a morpheme's status on merely a binary scale ([+/- Rule X]; esp. Walther & Wiese 1999; Anttila 1997, Pater 2000, Becker 2009, Jurgec 2016, *inter alia*). These results would challenge any claims that propose to merely group morphemes together arbitrarily or based on semantic profile, or to refer to stored whole words or phrases to capture the bulk of variation (Zuraw 2000, 2010; Bybee 2001, 2002).

Morphemes—both the undergoing stems and triggering suffixes in Slovenian—distribute across a spectrum, and so at the very least we need a theory that is capable of referring to idiosyncrasies of individual morphemes, whether it be through lexical indexation, UR constraints, or partitioning into very fine sublexica or cophonologies (Indexation/UR constraints: Pater 2000, Pater 2010, Pater, Staubs, Jesney, Smith 2012, Smith 2015, *inter alia*; Sublexical Phonology: Becker & Gouskova 2016 *et seq*; Cophonology Theory: Anttila 2002; Inkelas & Zoll 2005; *inter alia*).

Chapter 4:

Lexical propensities in French liaison

This section aims to show that lexical propensities significantly improve model performance on a phenomenon that has constituted a long-standing puzzle in phonology: variation in French liaison.

4.1 Some previous results for French liaison

4.1.1 Côté (2011)

French liaison has been investigated extensively in prior research (Delattre 1951, 1966; Schane 1968; Dell 1973/1985; Ågren 1973; Selkirk 1974; Klausenburger 1978; Morin & Kaye 1982; Morin 1986; de Jong 1994; Tranel 1981, 1996; Fougeron 2001a, b; Walker 2001; Boula de Mareüil et al. 2003; Durand & Lyche 2008; Mallet 2008; Côté 2011; Barreca & Christodoulides 2017; Kilbourn-Ceron 2017; Zuraw & Hayes 2017; *inter alia*). This section summarizes some major highlights from Côté (2011), who provides a review of prior literature concerning French liaison, variation in its application, and the factors that influence this variation. French liaison is the pronunciation of a consonant between two words (below categorized as Word1 and Word2), the latter being vowel- or glide- initial, in a variety of triggering contexts:

(13)	*un cordeau* ‘a line’	[œ̃ kɔʁdo]	*un homme* ‘a man’	[œ̃ **n** ɔm]
	vous voulez ‘you want’	[vu vule]	*vous allez* ‘you go’	[vu **z** ale]

grand prix	[gʁɑ̃ pʁi]	*grand oiseau*	[gʁɑ̃ **t** wazo]
‘grand prize’		‘great bird’	
très facile	[tʁɛ fasil]	*très actif*	[tʁɛ **z** aktif]
‘very easy’		‘very active’	

The majority of liaison consonants originated as word-final consonants, but between the twelfth and sixteenth century most of them progressively eroded away, but were retained in a prevocalic environment between two words exhibiting a high degree of cohesion (Morin 1986). A simplified, schematic account for the data in (13) would state that the final consonant emerges to prevent hiatus: the consonant-final allomorph of Word1 arises to prevent two adjacent vowels from occurring in the output.

Liaison consonants are, for the most part, restricted to a small subset of regularly occurring consonants in French, namely [z n t]. The consonants [ʁ p] occasionally surface in context, but [ʁ] occurs typically only after a small number of pronominal adjectives (e.g., *premier* ‘first’), and [p] only after the adverbs *trop*, ‘too much’ and *beaucoup*, ‘a lot’.

Whether liaison applies nearly categorically, optionally, or is blocked altogether is conditioned by morphosyntactic environment (Schane 1968; Selkirk 1974; Klausenburger 1978; Morin & Kaye 1982; Boula de Mareüil et al. 2003; Durand & Lyche 2008). Given below are environments previously found to be associated with categorical application (data from Côté 2011):

(14)	a. Determiner + adjective/noun		c. Verb/enclitic + enclitic	
	les enfants	[le **z** ɑ̃fɑ̃]	*allez-y*	[ale **z** i]
	‘children’		‘go ahead’	
	un autre enfant	[œ̃ **n** otʁ ɑ̃fɑ̃]	*allez-vous-en*	[ale vu **z** ɑ̃]
	‘another child’		‘go away’	

b. Proclitic + proclitic/verb		d. Compounds and fixed phrases	
vous en avez	[vu **z** ã **n** ave]	*mesdames et messieurs*	[medam **z** e mesjø]
‘you have some’		‘ladies and gentlemen’	
on arrive	[ɔ̃ **n** aʁiv]	*comment-allez-vous*	[kɔmã **t** ale vu]
‘we arrive’		‘how are you’	

In contrast, application apparently never applies between subject and verb, singular noun and adjective, or after conjunctive *et*. The following asterisked forms are purely hypothetical, but they would look as follows if they were to undergo liaison:

(15)	*l’enfant a réussi*	[l ãfã a ʁeysi], *[l ãfã **t** a ʁeysi]	Côté (2011)
	‘the child has succeeded’		
	un repas italien	[œ̃ rœpa italjɛ̃], *[œ̃ rœpa **z** italjɛ̃]	
	‘an Italian meal’		
	lui et elle	[lɥi e ɛl], *[lɥi e **t** ɛl]	
	‘him and her’		

Furthermore, the following environments were found to condition only optional application:

(16)	a. Adjective-PL + noun		d. Adverb + X	Côté (2011)
	beaux outils	[bo **z** uti] ~ [bo uti]	*mieux intégré*	[mjø **z** ɛ̃ntegʁe]
	‘beautiful tools’		‘better integrated’	~ [mjø ɛ̃ntegʁe]
	b. Adjective-SG.MASC + noun		e. Verb + X	
	gros effort	[gro **z** ɛfɔʁ]	*il est arrivé*	*[il ɛ **t** aʁive]*
	‘big effort’	~ [gro ɛfɔʁ]	‘he arrived’	~ *[il ɛ aʁive]*
	c. Preposition/conjunction + X		f. Noun-PL + adjective	
	quand elle arrive	[kã **t** ɛl aʁiv]	*soldats italiens*	[sɔlda **z** italjɛ̃]
	‘when she arrives’	~ [kã ɛl aʁiv]	‘Italian soldiers’	~ [sɔlda italjɛ̃]

Notice here that the data exhibit free variation within the same sequence, rather than simply variation over different sequences.

Applicability of liaison is also affected by phonological factors. A corpus study given in Mallet (2008) reveals that liaison applies more readily if the liaison consonant is *n* rather than *t* or *z*, or if Word1 is monosyllabic rather than polysyllabic. Previous research also argues that length of the sequence following Word1 is a significant factor, with shorter sequences triggering liaison more regularly. Morin & Kaye (1982) offer the following contrast:

(17a) *Ils travaillent d'abord et mangen*[t] *après.* Côté (2011)
'They work first and eat after'

(17b) ?*Ils mangen*[t] *après qu'ils aient fini leur travail.*
'They eat after that they have-SUBJ finished their work.'

In particular, liaison was suggested to be more natural in the former context than the latter, where the sequence following Word2 is longer.

Finally, some researchers have suggested that corpus-based propensities *might fluctuate based purely on lexical factors* (Ågren 1973, de Jong 1994, Boula de Mareüil et al. 2003, Mallet 2008, Barreca & Christodoulides 2017). Consider the proportion of realized liaison after four monosyllabic adverbs in the two corpora provided below (rates reported from Côté 2011):

(18)	*très*, 'very'	*plus*, 'more'	*bien*, 'well'	*pas*, 'not'
Mallet (2008)	97%	64%	43%	1%
de Jong (1994)	99%	96%	82%	7%

The data above suggest that liaison is in part conditioned by individual lexemes, and cannot be reduced to independent structural factors. Moreover, liaison has been found to be positively correlated with Word1 frequency (de Jong 1994; Fougeron 2001a, b; Kilbourn-Ceron 2017). Liaison has been evolving for nearly a millennium (Morin 1986), and overall it appears that synchronic knowledge of the variation constitutes knowledge of segmental, prosodic, morphosyntactic, lexical, and stylistic conditioning. I now turn to Zuraw & Hayes (2017), which focuses in particular on variation in a co-conspirator process of French liaison, namely *élision*.

4.1.2 Zuraw & Hayes (2017): Variation in French liaison/élision

Zuraw & Hayes (2017) investigate a related kind of allomorphy occurring in French, which seemingly too militates against hiatus. A variety of function words and adjectives in the language have two allomorphs, and which one gets chosen is determined both phonologically and lexically—we refer to this below as *liaison*/*élision* (though note that some authors reserve the term liaison for words with a single spelled form, as we have seen above). Consider the following data from Zuraw & Hayes (2017) on function word allomorphy:

Word	Example of CV allomorph	Gloss	Example of C/CVC allomorph	Gloss
‘the-*fem*’:	***la** courgette* [**la** kuʁʒɛt]	‘the zucchini’	***l’**aubergine* [**l** obɛʁʒin]	‘the eggplant’
‘of’:	***de** jonquilles* [**də** ʒɔ̃kij]	‘of daffodils’	***d’**iris* [**d** iʁis]	‘of irises’
‘of the-*masc*’:	***du** petit* [**dy** pətit]	‘of the small one’	***de l’**enfant* [**də l** ɑ̃fɑ̃]	‘of the child’
‘at/to the-*masc*’:	***au** lac* [**o** lak]	‘at the lake’	***à l’**étang* [**a l** etɑ̃]	‘at the pond’

Table 14: *French liaison/élision (Zuraw & Hayes 2017, p. 519)*

In this case, the consonant-final allomorph is employed when the following word is vowel-initial; otherwise, the vowel-final allomorph surfaces.

Some vowel-initial Word2s fail to trigger liaison/élision. Many of these are called *h-aspiré* words, owing to the fact that they are spelled with an initial <h>:

(19) Vowel-initial Word2s that behave as though they are consonant-initial

la hache	[**la** aʃ], *[**l** aʃ]	‘the axe’
du haricot	[**dy** aʁiko], *[**də la**ʁiko]	‘of the bean’
un homard	[œ̃ omaʁ], *[œ̃ **n** omaʁ]	‘a lobster’
un héros	[œ̃ eʁo], *[œ̃ **n** eʁo]	‘a hero’

These words take the CV allomorph despite the resulting hiatus. *h-aspiré* words’ behavior is a vestige of an earlier stage of the French language in which they bore an initial consonant (Zuraw & Hayes 2017, p. 520). Nevertheless, blocking by <h>-initial words is variable: some <h>-initial words are non-*h-aspiré* words in the sense that they fail to block liaison (e.g., *u*[n] *homme*). In addition to *h-aspiré words*, words with initial glides also display variable blocking behavior, as the data below illustrate:

Non-elided allomorphs			*Elided allomorphs*		
le *yodle*	[**lə** jɔdl]	‘yodels it’	***l’****iode*	[**l** jɔd]	‘the iodine’
le *yaourt*	[**lə** jauʁt]	‘the yogurt’	***l’****yeuse*	[**l** jøz]	‘the oak’
la *hiérachie*	[**la** jeʁaʁʃi]	‘the hierarchy’	***l’****hiatus*	[**l** jatys]	‘the hiatus’
la *huée*	[**la** ɥe]	‘the booing’	***l’****huître*	[**l** ɥitʁ]	‘the oyster’
le *huitième*	[**lə** ɥitjɛm]	‘the eighth’	***l’****huile*	[**l** ɥil]	‘the oil’
			l’*huissier*	[**l** ɥisje]	‘the bailiff’
le *ouistiti*	[**lə** wistiti]	‘the marmoset’	***l’****ouest*	[**l** wɛst]	‘the west’

Table 15: *Variation in French élision before glide-initial words (Walker 2001, p. 105-106)*

A pattern of variation thus arises: some, but not all, vowel-initial words block liaison/élision, regardless of whether or not the word is <h>-initial. Zuraw & Hayes (2017) analyze variable blocking of liaison/élision in a set of 358 Word1-Word2 sequences in which Word2 is glide-initial or spelled with an initial <h>. The data were extracted from the Google Ngrams corpus for French (https://books.google.com/ngrams).

Consider the table below, which features different Word2s and their propensity to trigger liaison across Word1s in the corpus:

Word2	*liaison rate*	*gloss*
habituel	96.9%	‘habitual-*masc*’
habituelle	99.0%	‘habitual-*fem*’
habitus	97.1%	‘habitus’
hache	0.1%	‘axe’
hachette	0.0%	‘hatchet; moth sp.
hacienda	78.6%	‘hacienda’
haddock	0.0%	‘haddock’
Hadès	85.3%	‘Hades’
hadji	0.0%	‘haji’
Hadrien	98.6%	‘Hadrian’

Table 16: *Individual liaison rates across h-aspiré words (Zuraw & Hayes 2017, p. 525)*

The analyst might conclude from these data that liaison exhibits both lexical and free variation: lexical in the sense that cases like *habituelle* trigger liaison the great majority of the time, but cases like *hadji* wholly block it; and free in the sense that the majority of words do not feature a strictly categorical effect, but rather display propensities only biased towards the extremes — and, as can be observed above, items like *hacienda* are associated with strikingly medial rates. Zuraw & Hayes plotted in a histogram, shown below, the different triggering propensities of the various Word2s, where the propensity of a particular Word2 was taken to be the average over all Word1s it cooccurs with:

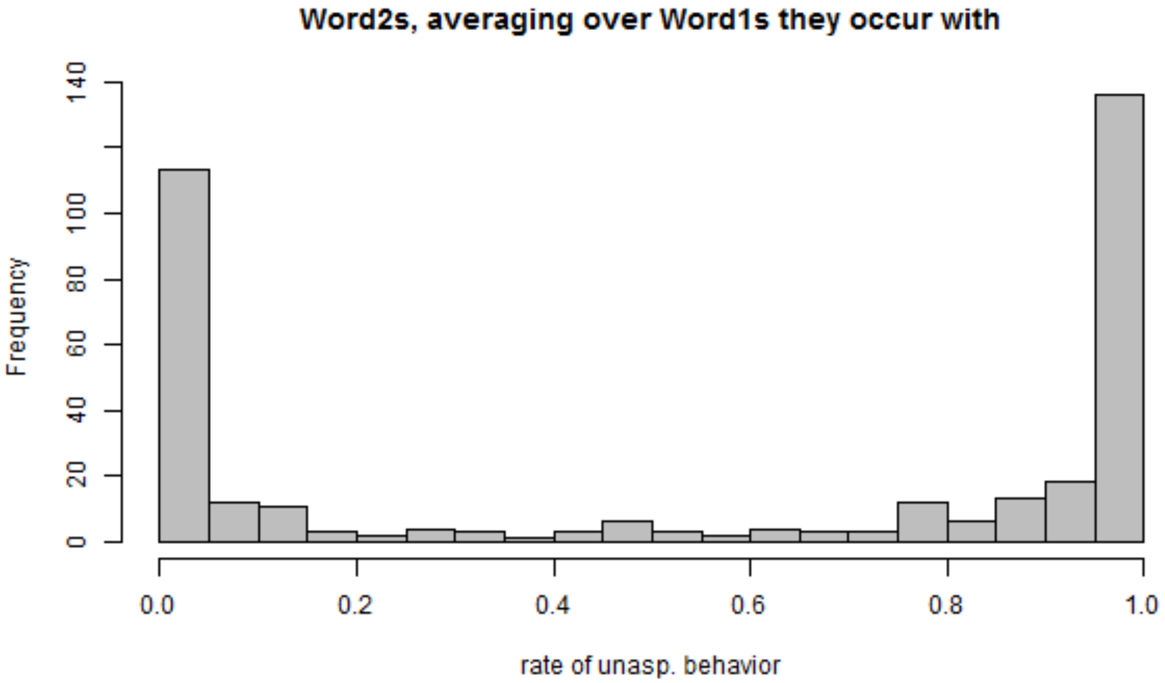


Figure 11: *Zuraw & Hayes's histogram showing number of Word2's with particular rate (Zuraw & Hayes 2017, p. 524)*

The different rates form a U-shaped distribution: though most propensities associated to Word2 cluster around the poles, quite a few Word2s exhibit medial propensity to trigger liaison. These data suggest that the allomorphy here is not encoded as categorically applying across words, or even as stochastically applying across words at a fixed rate. Rather, the allomorphy occurs across words at idiosyncratic rates.

In light of the above, we now turn to a corpus investigation of French liaison. This is different from Zuraw & Hayes's study in that the phenomenon they investigate is written (***la*** *courgette* ~ ***l'****aubergine*), and they rely on a corpus of written forms; my study, on the other hand, draws upon a corpus of spoken *liaison*, which in what follows should be understood to be the phenomenon whereby one spelled form can be pronounced two ways depending on whether the following form begins with a vowel (*très facile* [tʁɛ fasil] ~ *très actif* [tʁɛ **z** aktif]). The corpus study is conducted to assess whether words in French are associated with idiosyncratic liaison rates. As I will show below, at least the identities of a variety of different Word1s are significant predictors of liaison—that is to say, various Word1s undergo liaison at different rates.

4.2 Variation in a corpus of French liaison

I give in this section a brief description of the *Phonologie du Français Contemporain* database, the corpus from which I drew. Additionally, I compare a series of logistic regression models fitted to the data, revealing that individual words play a role in predicting rates of liaison, even after other factors previously found to affect liaison are controlled for.

4.2.1 Extracting from the PFC corpus

The *Phonologie du Français Contemporain* (PFC; Durand, Laks & Lyche 2002, 2009; Durand & Lyche 2008; http://www.projet-pfc.net/) is a large online database of spoken French which contains a sub-database of around 54,000 Word1-Word2 sequences having the graphical form <...C#V...>, categorized for whether they undergo liaison. Sequences are also classified for whether Word1 was monosyllabic or polysyllabic, for whether the juncture exhibits a pause or glottal stop in speech, and, provided the sequence does undergo liaison, for the particular liaison consonant. Potential liaison contexts were defined as those that Delattre (1951, 1966) defined as potential contexts—Durand and Lyche exclude cases where Word1 is a singular noun, or *et*, following Delattre (1951, 1966)'s findings that for such words liaison is categorically forbidden. See Durand & Lyche (2008) for further discussion of the PFC protocol.

I extracted all of these sequences from the PFC database, excluding the very few cases

(< 1% of the data) marked in the PFC as *liaison non-enchainée*, uncertain liaison, and epenthetic liaison. The extracted PFC data were already classified for factors previously found to significantly affect application of liaison, including the identity of the liaison consonant and whether the word was mono- or polysyllabic. I further classified the data for parts of speech of Word1 and Word2,[2] and log frequencies of Word1 and Word2. Log frequencies were further scaled for purposes of achieving model convergence: each value for a Word1 was divided by the maximal log frequency for Word1 in the corpus; Word2 received the same treatment. This brought frequency values for each word into a range between 0 and 1.

Following Zuraw and Hayes, I calculated for each of the 184 Word1s occurring 100 or more times in the corpus its mean liaison rate, averaging over the Word2s they cooccur with. A histogram of the resulting propensities is shown below. 85 Word1's are categorical undergoers and 30 Word1's are categorical nonundergoers, while 69 Word1's undergo liaison at medial rates.

[2] Part of speech was annotated by retrieving part of speech information of each word from the *Lexique* database (New et al. 2001). In words where the part of speech was ambiguous, the most commonly reported part of speech was taken. Words with 20 and more tokens were hand-checked to verify that that they had the correct part of speech given the construction that they were situated in.

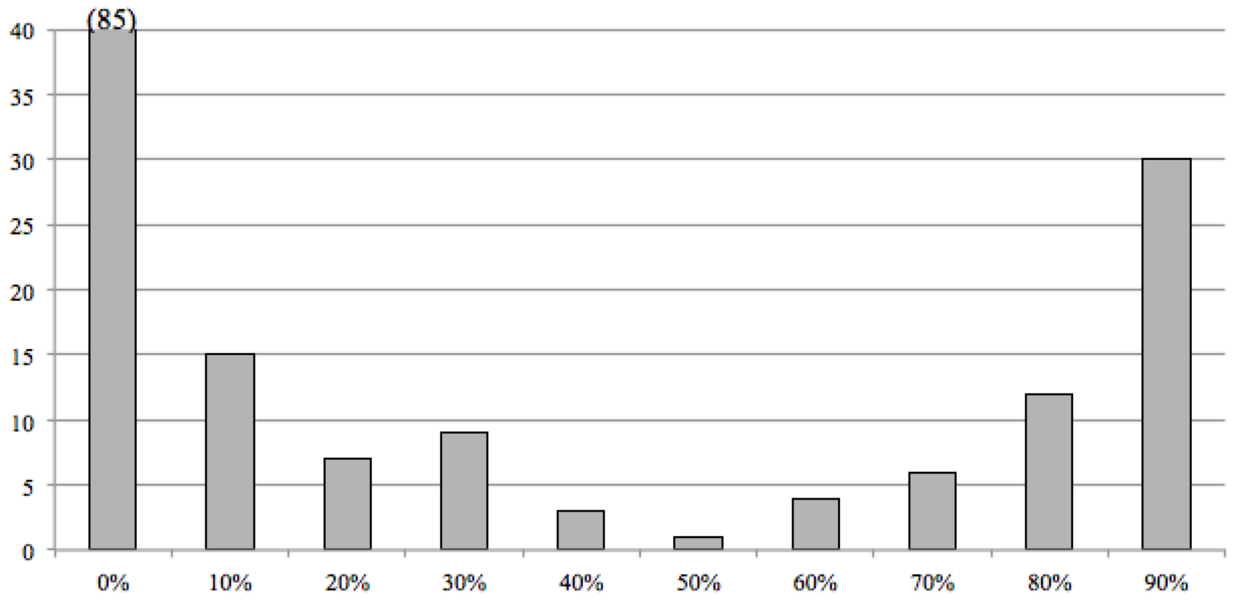


Figure 12a: *Histogram of Word1 rate frequencies*

Analogous calculations were made for the 115 Word2s occurring 100 or more times in the corpus, plotted in the histogram below:

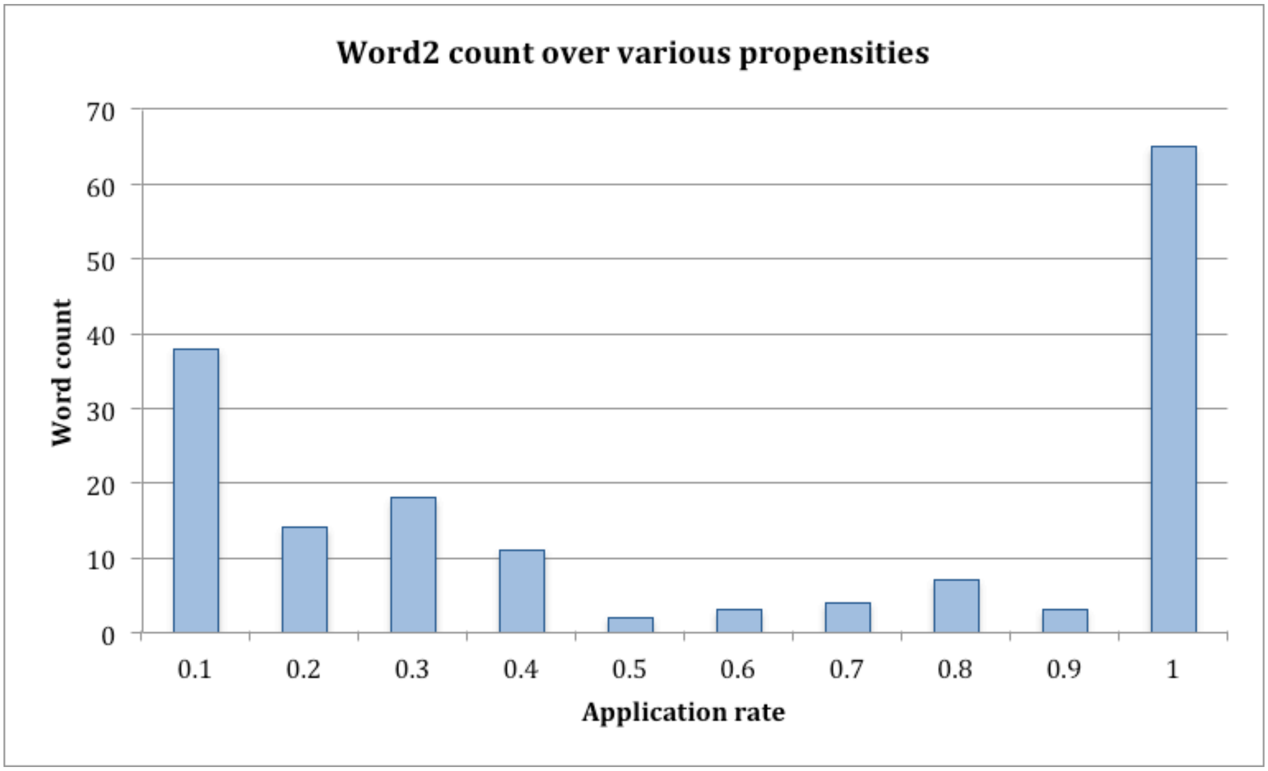


Figure 12b: *Histogram showing number of Word2s with particular rate*

As can be observed in the tables above, my findings replicate Zuraw (2016) and Zuraw & Hayes (2017)'s findings, in that the morpheme-specific propensities exhibit an apparent U-shaped distribution. The majority of the rates occur at the polls of the scale, though a healthy minority of words occur between 10% and 90% on the scale of liaison rates.

4.2.2 A baseline logistic regression model of corpus data on French liaison

This section presents a statistical analysis of the factors that condition liaison, with the aim of showing that models of liaison are significantly improved by encoding distinct Word1's with different propensities to participate in liaison. I used the `glmer` function of the *lme4* package (Bates & Maechler 2011) in R (R Core Development Team 2014) to fit two logistic regression models to the liaison data extracted from the PFC corpus: a *Baseline Model*, which lacked factors referring to word identity, but which contains the aforementioned factors previously found to condition liaison, including grammatical context, liaison consonant, Word1 syllable count, and Word1 frequency; and a *Word1 Propensity Model*, where Word1 identity is coded as a random intercept, i.e., where individual Word1's (3,462 in total) can be associated with idiosyncratic propensities. As I will show, the propensity model performs substantially better according to well-established modeling metrics.

The logistic regression models here regress over bigram types. Each row of the data set includes a unique bigram, its liaison rate calculated as the number of liaised tokens of that bigram divided by the total number of tokens of that bigram in the corpus. In all models, the dependent variable is the liaison rate of the bigram. Each row also contains information about the identity of the liaison consonant, whether Word1 is mono- or polysyllabic, part of speech of

Word1, part of speech of Word2, frequency information about Word1, frequency information about Word2, the identity of Word1, and the identity of Word2.

The Baseline Model contained six factors, all coded as main effects: cons, syls, W1POS, W2POS, W1freq, and W2freq. cons takes on the five possible values for the identity of the liaison consonant given in the corpus: *n, p, r, t, z*. syls takes 0 or 1 depending on whether Word1 is mono- or polysyllabic, respectively. W1POS and W2POS takes on the following values:

(20a) W1POS: ADV: adverb
CON: conjunction
DET: determiner
NOM: noun
NUM: number
PRE: preposition
PRO: pronoun
VER: verb

(20b) W2POS: ADV: adverb
CON: conjunction
DET: determiner
NOM: noun
NUM: number
ONO: name (person or location)
PRE: preposition
PRO: pronoun
UTT: utterance (coded “euh”, “oe” in the PFC)
VER: verb

The factors in the Baseline Model are thus defined as follows:

(21) cons: identity of the liaison consonant

syls: 0 if Word1 is monosyllabic, else 1

W1POS: part of speech of Word1

W2POS: part of speech of Word2

W1freq: log of Word1 token frequency across the corpus, divided by the maximum log value of token frequency ranging over all Word1’s in the corpus

`W2freq`: log of Word2 token frequency in the corpus, divided by the maximum log value of token frequency ranging over all Word2's in the corpus

The results of the Baseline Model are given below. The factors were compared against a baseline (intercept) of Word1-Word2 pairs that had liaison consonant *n*, were monosyllabic, had Word1 and Word2 as adjective part of speech, and had a scaled log-frequency of 0. The results given below indicate that the factors given above are significant predictors of liaison rate, in line with prior research that found significant effects for these factors (de Jong 1994, Mallet 2008), with the exception of `W2freq`—see below for further discussion.

The graph below suggests a substantial effect of liaison consonant identity on liaison. We include *t*, *z*, *n*, and *r*, each of which occurs across more than 100 bigram types in the dataset. The *x*-axis gives consonant and, in parenthesis, number of bigram types with Word1 ending in that consonant.

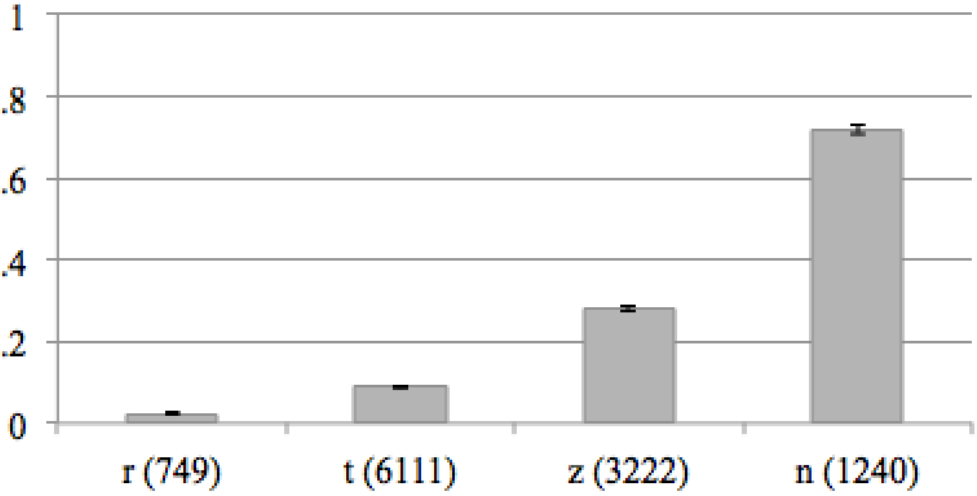


Figure 13a: *different liaison rates based on final consonant of Word1*

The effect of consonant identity is significant in the Baseline Model: `const` and `consz`, for example, receive negative coefficients relative to baseline `consn`, confirming that liaison rates associated with [t] and [z] are associated with overall lower rates than [n] (Mallet 2008).

```
Coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept:    3.64     0.29     12.36   <0.001   ***
consn)

consr         -0.73     0.34     -2.10    0.035   *
const         -1.11     0.17     -6.50   <0.001   ***
consp         -1.23     0.40     -3.04    0.002   **
consz         -1.27     0.13     -9.49   <0.001   ***
```

Table 17a: *Baseline Model results: consonant identity factor*

Word1 syllable count also substantially affects liaison rate in my data, as the graph below reveals—monosyllabic Word1s are associated with higher liaison rates than polysyllabic Word1s.

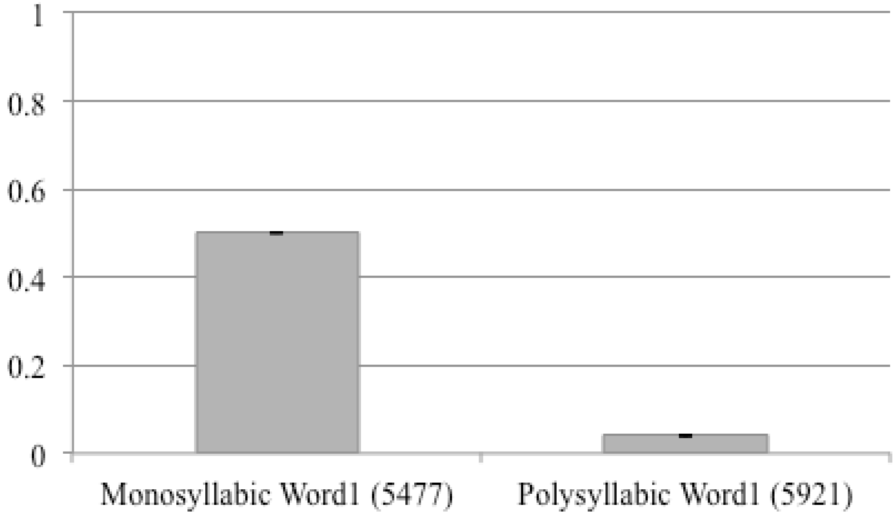


Table 13b: *different liaison rates based on Word1 syllable count*

In the Baseline Model, `syls` receives a negative coefficient, confirming Mallet (2008)'s finding that polysyllabic words are associated with lower levels of liaison.

```
Coefficients:

               Estimate Std. Error z value Pr(>|z|)
(Intercept:      3.64      0.29      12.36   <0.001   ***
syls0)

syls1           -1.67      0.11     -14.68   <0.001   ***
```

Table 17b: *Baseline Model results: Word1 syllable count factor*

Word1 and Word2 part of speech are significant predictors of liaison, in line with prior research as well (Durand & Lyche 2008). Table Xa gives PFC rates for grammatical contexts stated by Durand & Lyche (2008) to be associated with categorical liaison—liaison rates are high in these contexts, but nevertheless not at ceiling (see Mallet (2008), who also finds variable liaison in contexts traditionally considered to be associated with categorical application). Table Xb gives PFC rates for grammatical contexts stated by Durand & Lyche (2008) to be associated with variable liaison, confirming their findings.

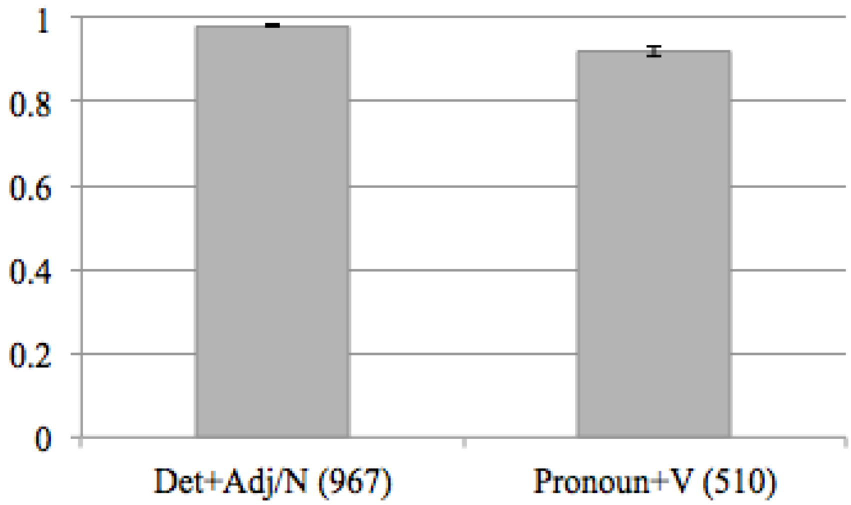


Figure 13c: *PFC rates for grammatical contexts stated to be associated with categorical application in Durand & Lyche (2008)*

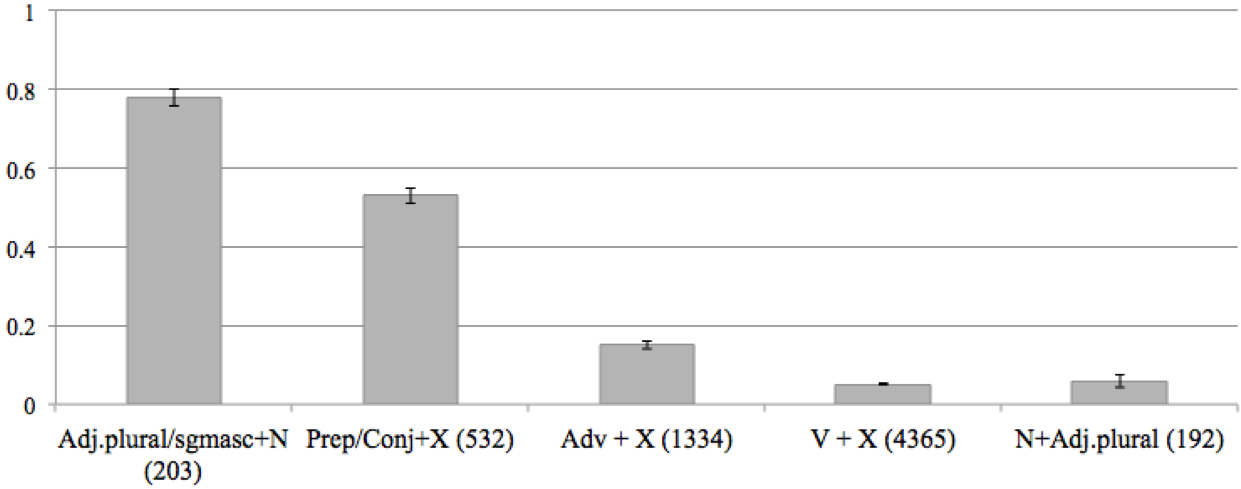


Figure 13d: *PFC rates for grammatical contexts stated to be associated with variable application in Durand & Lyche (2008)*

In the Baseline Model, for example, we find that determiner status, relative to adjective status, significantly increases liaison rate by observing the positive coefficient, `1.1529`, of the significant factor `W1POSDET`; on the other hand, we find that conjunction status significantly decreases liaison rate by observing the negative coefficient, `-2.9465`, of the significant factor `W2POSCON`. The coefficients span a range of values, suggesting that different grammatical categories are associated with different liaison rates distribute across a spectrum. This can be observed below (see (20) above for codes for the different levels):

```
Coefficients:

             Estimate Std. Error z value Pr(>|z|)
(Intercept:   3.64      0.29     12.36   <0.001   ***
W1POSADJ,
W2POSADJ)

W1POSPRO      1.19      0.19      6.05   <0.001   ***
W1POSDET      1.15      0.19      5.94   <0.001   ***
W1POSNUM      0.73      0.25      2.85    0.004   **
W1POSPRE      0.40      0.24      1.66    0.095   .
W1POSCON     -1.52      0.29     -5.16   <0.001   ***
W1POSADV     -1.70      0.17     -9.53   <0.001   ***
W1POSNOM     -2.00      0.17    -11.22   <0.001   ***
W1POSVER     -2.59      0.18    -14.11   <0.001   ***

W2POSNOM      0.75      0.13      5.55   <0.001   ***
W2POSVER     -0.47      0.12     -3.83   <0.001   ***
W2POSNUM     -1.55      0.66     -2.34    0.019   *
W2POSONO     -2.05      0.42     -4.84   <0.001   ***
W2POSPRE     -2.35      0.22    -10.37   <0.001   ***
W2POSUTT     -2.41      0.35     -6.74   <0.001   ***
W2POSDET     -2.46      0.33     -7.31   <0.001   ***
W2POSCON     -2.94      0.33     -8.89   <0.001   ***
W2POSADV     -3.01      0.24    -12.36   <0.001   ***
W2POSPRO     -3.40      0.31    -10.67   <0.001   ***
```

Table 17c: *Baseline Model results: part of speech factors for Word1 and Word2*

Lastly, `W1freq` is a significant predictor of liaison, at least in the Baseline Model. The graph below shows a trend toward higher liaison rates when Word1 is more commonly occurring in the corpus. Values of the *x*-axis in the table consist of intervals: 0.1, for example, represents liaison rate across Word1's with a scaled log frequency between 0 and 0.1; 0.2 represents the rate across Word1's with a scaled log frequency between 0.1 and 0.2; and so on.

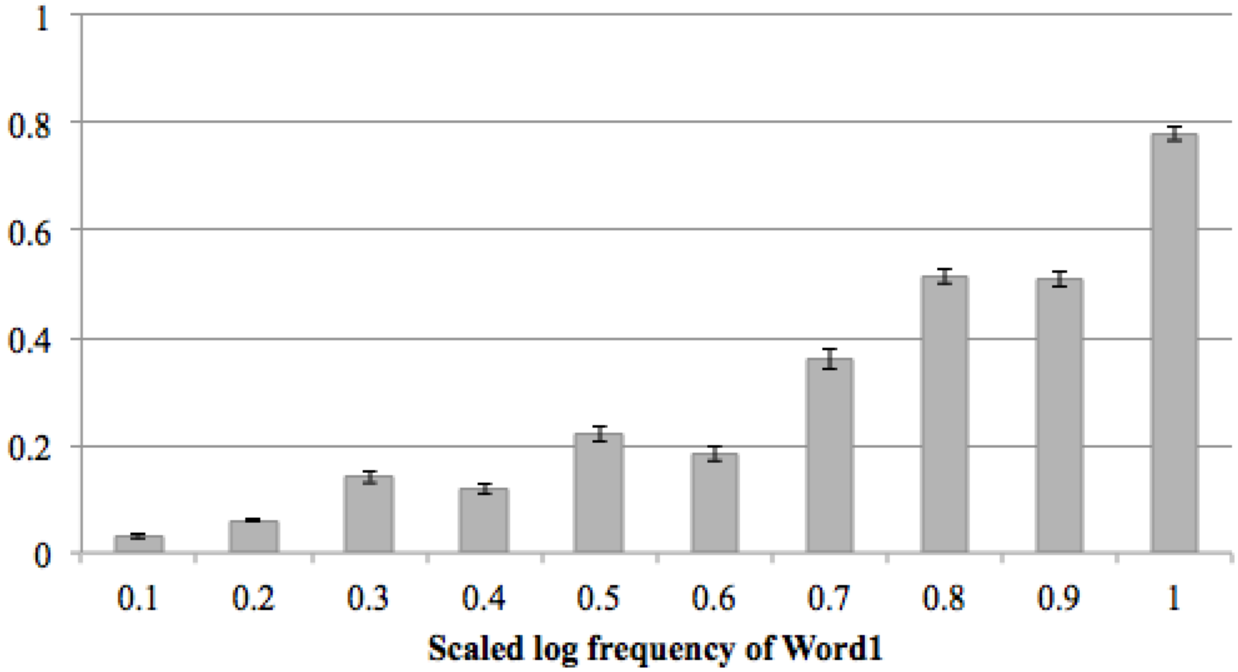


Figure 13e: *liaison rate based on Word1's scaled log frequency*

In the Baseline model, we observe from the positive coefficient for `W1freq(0.7322)` that higher frequencies of Word1 encourage liaison, confirming prior findings (de Jong 1994; Fougeron 2001a, b; Kilbourn-Ceron 2017); `W2freq`, on the other hand, is not a significant predictor according to the complete corpus and the model.[3]

[3] Though Kilbourn-Ceron (2017) finds that higher Word2 frequencies are associated with higher rates of liaison in the plural noun-adjective and adjective-noun contexts, we do not find a Word2 frequency trend across the PFC dataset overall (at least when we consider frequency of Word2 occurring within the PFC specifically). This should not suggest that the former results are invalid—though I do not take up this work here, more research should be conducted to assess whether Word2 frequency conditions variation within particular grammatical contexts.

```
Coefficients:

            Estimate Std. Error z value Pr(>|z|)
(Intercept:   3.64     0.29      12.36   <0.001    ***
W1freq0,
W2freq0)

W1freq        0.73     0.19       3.80   <0.001    ***
W2freq       -0.20     0.20      -1.00    0.317
```

Table 17d: *Baseline Model results: factors for scaled log frequency*

Recall from Section 3.2 that the Akaike Information Criterion scores models based on fit to the data and number of parameters, with a lower score being better. The current model's AIC score is **4119.3**.

Taking stock, we have constructed a logistic regression model of liaison that incorporates both phonological and frequency-based conditioning factors previously found to condition liaison, finding that almost all of them are highly significant predictors. We now turn to fitting a mixed-effects logistic regression model with Word1 as a random intercept. If the following model were to substantially outperform the Baseline Model—the latter of which already contains a variety of factors that significantly condition liaison—then this would lends strong support to the hypothesis that variation in liaison is at least in part lexically conditioned, even after taking into account other potential conditioning factors.

4.2.3 A mixed-effects logistic regression model of French liaison with lexical propensities

The Word1 Propensity Model contains the factors in the Baseline Model as fixed effects—namely, `cons`, `syls`, `W1POS`, `W2POS`, `W1freq`, and `W2freq`—and Word1 identity, coded as a random intercepts for the different Word1s in the corpus (e.g., Word1=*très*). The results for Word1 Propensity Model indicate that various levels of most of the main effects are significant predictors of liaison rate, in line with prior findings. The effect of consonant identity survives even when Word1 is taken to be a random effect: *r*, *t*, and *z* are found to significantly decrease liaison rate relative to *n*, the consonant identity in the intercept condition.

```
Random effects:
 Groups Name        Variance Std.Dev.
 W1     (Intercept) 3.618    1.902
Number of obs: 11398, groups:  W1, 3158

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept:     4.02     0.55     7.20    <0.001   ***
consn)

consp          -1.06     1.34    -0.78     0.431
consz          -2.12     0.34    -6.15    <0.001   ***
consr          -2.13     0.71    -2.99     0.002   **
const          -2.16     0.42    -5.12    <0.001   ***
```

Table 18a: *Word1 Propensity Model results: consonant identity factor*

Syllable count for Word1 was also a significant predictor, with polysyllabicity being associated with lower rates of liaison:

```
Random effects:
 Groups Name        Variance Std.Dev.
 W1     (Intercept) 3.61     1.90
Number of obs: 11398, groups:  W1, 3158

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept:     4.02     0.55       7.20  <0.001   ***
syls0)

syls           -1.82     0.20      -8.87  <0.001   ***
```

Table 18b: *Word1 Propensity Model results: Word1 syllable count factor*

Various parts of speech of Word1 and Word2 were also significant predictors of liaison:

```
Random effects:
 Groups Name        Variance Std.Dev.
 W1     (Intercept) 3.61     1.90
Number of obs: 11398, groups:  W1, 3158

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept:     4.02     0.55       7.20  <0.001   ***
POS1ADJ,
POS2ADJ)

POS1NUM         2.26     0.57       3.92  <0.001   ***
POS1PRO         1.27     0.50       2.52   0.011   *
POS1DET         0.17     0.51       0.33   0.736
POS1PRE         0.14     0.60       0.24   0.806
POS1CON        -0.71     0.91      -0.77   0.438
POS1ADV        -2.54     0.48      -5.30  <0.001   ***
POS1NOM        -2.67     0.31      -8.48  <0.001   ***
POS1VER        -3.19     0.36      -8.71  <0.001   ***

POS2NOM         1.03     0.17       5.97  <0.001   ***
POS2VER        -0.00     0.16      -0.05   0.954
POS2NUM        -1.19     0.88      -1.35   0.175
POS2UTT        -2.38     0.42      -5.56  <0.001   ***
POS2PRE        -2.42     0.28      -8.57  <0.001   ***
POS2ONO        -2.53     0.52      -4.84  <0.001   ***
POS2DET        -2.65     0.43      -6.11  <0.001   ***
POS2ADV        -2.79     0.29      -9.59  <0.001   ***
POS2CON        -3.19     0.41      -7.75  <0.001   ***
POS2PRO        -3.46     0.38      -9.08  <0.001   ***
```

Table 18c: *Word1 Propensity Model results: part of speech factors for Word1 and Word2*

Finally, the Word1 frequency measure from the Baseline Model is *not* a significant predictor of liaison when Word1 identity is taken as a random effect. As in the Baseline Model, Word2 frequency does not have a significant effect either:

```
Random effects:
 Groups Name        Variance Std.Dev.
 W1     (Intercept) 3.618    1.902
Number of obs: 11398, groups:  W1, 3158

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept:    4.02       0.55     7.20   <0.001   ***
W1freq0,
W2freq0)

W1freq         0.28       0.41     0.67    0.500
W2freq        -0.23       0.24    -0.94    0.343
```

Table 18d: *Word1 Propensity Model results: factors for scaled log frequency*

The Word1 Propensity Model explains a good deal of the variation found in liaison: its AIC is **3566.7**. On the other hand, the Baseline Model had an AIC of 4119.3. Recall that, in comparing two models *A* and *B* of the same data set, *B* outperforms *A* if *B* receives an AIC that is at least ten points lower than *A*'s (Burnham & Anderson 2004). The Word1 Propensity Model scored over ***550*** points lower than the Baseline Model. Thus the propensity model performs vastly better than both the Baseline Model and the binary-scale model. This suggest that lexical factors do indeed play a role in conditioning variation (Ågren 1973, de Jong 1994, Mallet 2008) even after other phonological factors previously proposed to condition liaison has been taken into account. Moreover, it validates the usage of lexical propensities to account for lexical idiosyncrasies in French liaison.

Corpus investigations conducted thus far have not uncovered an effect of Word2 identity on liaison, at least for the PFC data. I conjecture that the null result is merely an artifact of the smalls size of the corpus—after all, Zuraw & Hayes (2017) found in their data a gradient effect of the identities of glide-initial and <h>-initial words on the likelihood of *liaison*/*élision*. Further investigation should be conducted to assess whether the null result is general across a broader set of data.

4.3 Summary

Here we have not only confirmed prior findings that liaison is conditioned phonologically, but have obtained corpus-based evidence that it is also conditioned lexically. In particular, our PFC investigation uncovered that different Word1s idiosyncratically condition the rate at which a bigram undergoes liaison—even after controlling for other phonological and frequency-based factors previously found to condition liaison—with rates spanning across a liaison propensity spectrum. While Zuraw & Hayes (2017) show that *liaison/élision* allomorphy is conditioned by the identities of different Word2s in their written data, I have shown that Word1 conditions liaison in a *liaison* corpus spoken by French adults, strengthening the case for morpheme-specific propensities. In light of these findings, a number of questions arise. Do speakers internalize these idiosyncratic rates? Moreover, are they capable of internalizing rates associated with particular Word1s, or do they internalize rates solely on a bigram-by-bigram basis? We now turn to a nonce probe experiment to address these questions.

Chapter 5:

Experimental evidence for speaker internalization of lexical propensities

Typically a variable process is said to vary freely in some set of words—that is, apply optionally to those words—or vary lexically: whether the process applies or not differs word-by-word, but any given word has a fixed pronunciation. This picture differs from what we observe in French liaison. Recall from the PFC corpus investigation that a healthy variety of words display gradient *lexical propensities* to undergo French liaison, patterning along an entire spectrum ranging from complete non-undergoers to complete undergoers. The effect seemingly cannot be chalked up to factors previously found to affect liaison, considering that a large effect of word identity was found in the corpus study even when these factors were controlled for. Hence, one can ask: do learners internalize the gradient effect of word identity? That is, do speakers internalize word-specific rates of liaison? The answer would have significant ramifications for phonological theory: it would challenge theories in which individual morphemes are coded with values on a binary scale (e.g., [+/- Rule]; Walther & Wiese 1999; Anttila 1997, Pater 2000, Becker 2009, Jurgec 2016, among others); but it would support theories in which the triggering or undergoing status of individual morphemes are implemented on an entire spectrum (e.g., [0.7 Rule]; Moore-Cantwell & Pater 2016, Smolensky & Goldrick 2016, Zuraw 2016, Zuraw & Hayes 2017). Furthermore, while speakers may very well possess knowledge about the behavior of whole, listed bigrams, evidence that French speakers track liaison on a word-by-word basis would suggest that they acquire lexical idiosyncrasies on a word-by-word basis, and not only as a set of whole, listed bigrams (Zuraw 2000, 2010; Bybee 2001, 2002; *cf.* Steriade 1994, 1999 *et seq*).

The present study assesses whether speakers can internalize word-specific rates of variable liaison. Consider the idiosyncratic behavior of various adverbs when they come before vowel-initial adjectives. de Jong (1994) conducted a corpus study of liaison using the Orléans corpus (Lonergan et al. 1974, Blanc & Biggs 1971, Bergounioux et al. 1992, Baude & Dugua 2011, *inter alia*), with data coming from speakers located in Orléans, France. According to Lonergan et al.'s catalog, there are 487 recordings of interviews, conferences, spontaneous speech from the street, and more, totaling to 315 hours of recorded speech. From this corpus, de Jong lists the following rates for *très*, *plus*, *bien*, and *pas*, which appear to pattern across a spectrum of propensity to undergo liaison:

Word1	Gloss	Rate
très	'very'	0.99
plus	'more'	0.96
bien	'very'	0.82
pas	'not'	0.07

Table 19a: *Liaison rates for four different adverbs as reported in de Jong (1994)*

Mallet (2008), drawing from PFC liaison data in or before 2008, list the following rates for *très*, *plus*, *bien*, and *pas* in his study:

Word1	Gloss	Rate
très	'very'	0.97
plus	'more'	0.64
bien	'very'	0.43
pas	'not'	0.01

Table 19b: *Liaison rates for four different adverbs as reported in Mallet (2008)*

Furthermore, Côté (2013), drawing from a subset of annotated liaison data in the PFC in or before 2013, lists the following rates for six adverbs:[4]

Word1	**Gloss**	**Tokens**	**Rate**
très	'very'	25	0.84
plus	'more'	22	0.59
trop	'too'	7	0.14
pas	'not'	163	0.01
mieux	'better'	10	0.00
moins	'less'	19	0.00

Table 19c: *Liaison rates for six different adverbs as reported in Côté (2013)*

These tables illustrate that words with very similar profiles—all of them monosyllabic adverbs, most of them with liaison consonant *z*—can bear very different liaison rates. *Bien* and *trop* have different liaison consonants—*n* and *p*, respectively—but their rates could not differ from the others based on consonant identity alone: *n* is associated with liaison rates higher than *z* (Mallet 2008), and yet *très* and *plus* liaise more than *bien* in both de Jong (1999) and Mallet (2008)'s investigations; and *p* rarely liaises, but of the seven tokens of *trop* given in Côté (2013), one of them liaises, yet *mieux* and *moins*, with even more tokens, never liaise in her investigation. Finally, in my investigation of the PFC corpus, five adverbs show different rates when they come before adjectives:

[4] In particular, Côté draws from liaison data that were at the time of the study not only transcribed, but annotated for various phonological, morphosyntactic, and sociolinguistic properties by the PFC creators themselves.

Word1	Tokens	Rate (%)
très	282[5]	0.90
bien	129	0.85
plus	117	0.79
moins	19	0.26
pas	246	0.00

Table 19d: *PFC liaison rates for five different adverbs*

The rates from the PFC corpus are overall similar to those found in other corpora, with a couple exceptions: though in my investigation *bien* liaises somewhat more than *plus*, in de Jong (1994) and Mallet (2008)'s studies *plus* liaises more than *bien*, and in fact liaises nearly as much as *très* in de Jong (1994); in my investigation *moins* liaises more than *pas*, but rates for *moins* are near zero in the other studies.

Considering how similar the rates are across corpora, it seems eminently plausible that speakers *internalize* the idiosyncratic rates of the different adverbial words discussed above. We thus investigate the psychological reality of liaison rates for the five words in Table 19d. Native speakers of French living in Paris were targeted to participate in a nonce probe task (Berko 1958) to assess whether speakers internalize these rates.

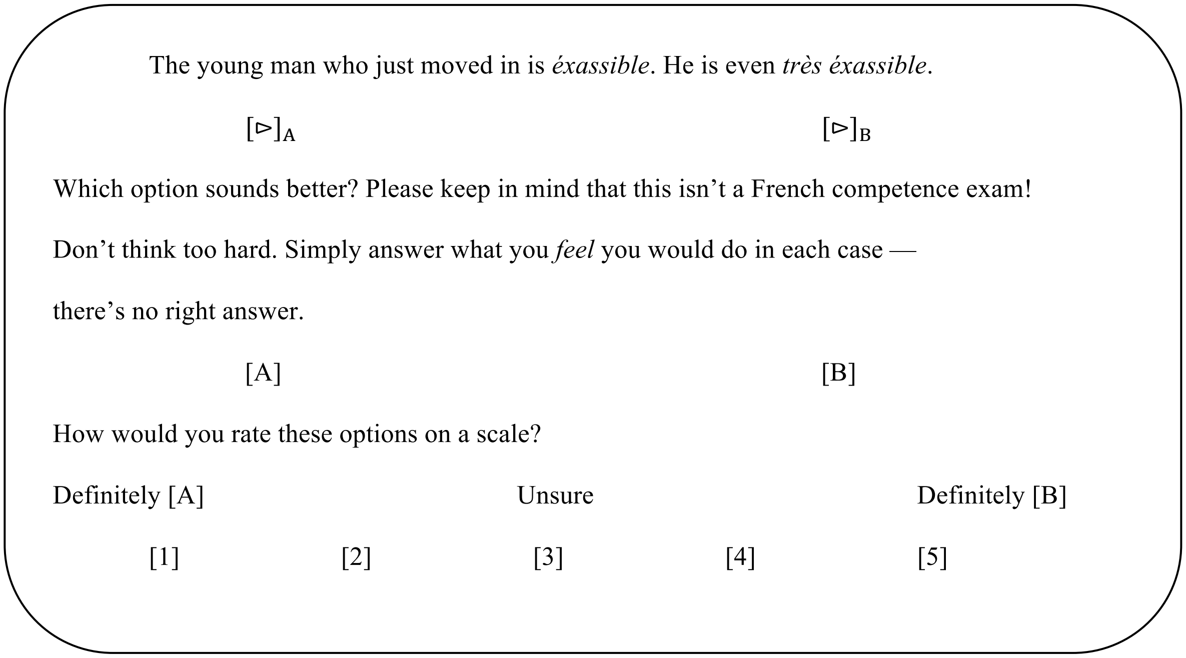
5.1 Experiment 1: Testing for the internalization of lexical propensities associated with *très*, *plus*, and *pas*

5.1.1 Task

Critical experiment trials consisted of adverb-adjective bigrams, where the adverb was taken from one of *très* ('very')*, plus* ('more')*,* and *pas* ('not'), and the adjective was a vowel-initial

[5] Another 1,960 of tokens with *très* were followed by the same Word2, *inquiet*, as in the bigram *très inquiet*, 'very worried'. (1,875 of these tokens liaised, for a bigram rate of roughly 0.95.) If we exclude cases of *très inquiet* from the token count for *très*, then we get 262 tokens total, roughly the same as *pas*.

nonce form. During each trial, participants read two carrier sentences, the first one introducing the nonce adjective and the second one introducing the whole bigram. They were then exposed two recordings of the bigram: one with the liaison form of the adverb, and one with the non-liaison form. The participant was then asked if they preferred the liaison form or the non-liaison form of the adverb, and then was asked to rate the two recordings on a scale from 1 to 5 (Scholes 1966; see also Zuraw 2000 for a similar design). The order in which the two pronunciations were given was randomized across trials. A snapshot of a single trial is given below, both in English and French:



The young man who just moved in is *éxassible*. He is even *très éxassible*.

[▻]$_A$ [▻]$_B$

Which option sounds better? Please keep in mind that this isn't a French competence exam!

Don't think too hard. Simply answer what you *feel* you would do in each case —

there's no right answer.

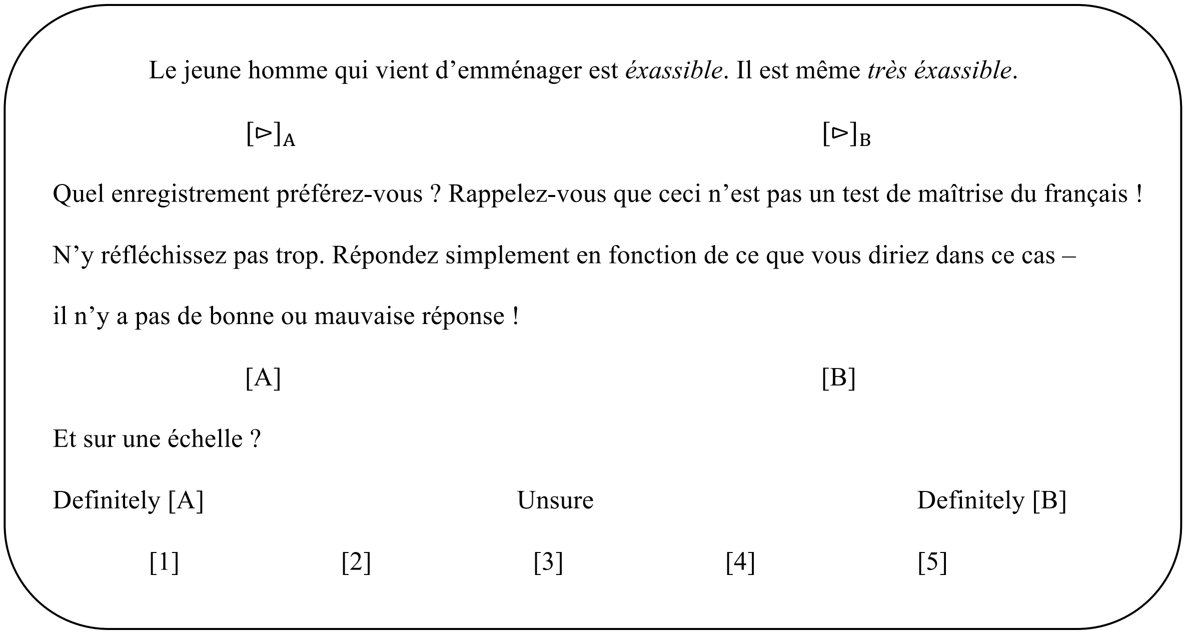
[A] [B]

How would you rate these options on a scale?

Definitely [A] Unsure Definitely [B]

[1] [2] [3] [4] [5]

Figure 14a: *trial snapshot (English translation)*



Le jeune homme qui vient d’emménager est *éxassible*. Il est même *très éxassible*.

[⊳]$_A$ [⊳]$_B$

Quel enregistrement préférez-vous ? Rappelez-vous que ceci n’est pas un test de maîtrise du français ! N’y réfléchissez pas trop. Répondez simplement en fonction de ce que vous diriez dans ce cas – il n’y a pas de bonne ou mauvaise réponse !

[A] [B]

Et sur une échelle ?

Definitely [A] Unsure Definitely [B]

[1] [2] [3] [4] [5]

Figure 14b: *trial snapshot (French translation)*

If participants replicated in their assessments the distinctions found in the corpus rates, then this would suggest that speakers internalize word-specific rates of French liaison; that is, they acquire knowledge of individual lexical propensities to undergo a variable phonological process.

5.1.2 Stimuli

The experiment tests for whether speakers internalize the effect of word identity across five adjective modifiers, when they precede nonce adjectives. All stimuli were recorded by a native French speaker in her late twenties who lived in Paris for several years. The three items tested were *très* (‘very’)*, plus* (‘more’)*,* and *pas* (‘not’), which yield different rates in the corpus studies, as shown in the table and graph below:

	très, 'very'	*plus*, 'more'	*pas*, 'not'
de Jong (1994)	99%	96%	7%
Mallet (2008)	97%	64%	1%
My study, PFC	95%	79%	0%
AVERAGE:	**97%**	**80%**	**3%**

Table 20: *corpus liaison rates for très, plus, and pas*

The literature categorizes each of these words as adverbs (*cf.* Côté 2011, 2013). All three words are monosyllabic and end in the same liaison consonant, ruling out syllable count and consonant identity as explanations for the difference in rate. Word1 frequency and liaison rate cannot alone explain the three different rates: though *plus* is much less frequent than *très*, *pas* is much more frequent than *plus*, and yet has a much lower liaison rate. This leaves word identity as a viable hypothesis for differences in rate.

Participants heard each of these words before eighteen nonce adjectives, each of which had an adjectival suffix. Two of these nonce adjectives were consonant-initial, and were used only to assess whether the participant internalized the correct environment of liaison. The other sixteen were vowel-initial — four groups of four words with the same initial vowel. Each trial consisted of an adverb-nonce adjective bigram, totaling to 3 × 18 = 54 bigrams/trials. The stimuli are presented below.

Adv.	IPA
très	tʁɛ/tʁɛz
plus	ply/plyz
pas	pa/paz

Nonce adj.	IPA	Status
arvant	aʁvɑ̃	V-initial (critical)
agrivieux	agʁivjø	
amagné	amaɲe	
altimable	altimabl	
énantant	enɑ̃tɑ̃	
éprieux	epʁijø	
émoivré	emwavʁe	
écastable	ekastabl	
impergeant	ɛ̃pɛʁʒɑ̃	
invinieux	ɛ̃vinjø	
introché	ɛ̃tʁɔʃe	
ingroutable	ɛ̃grutabl	
autrillant	otʁijɑ̃	
auquieux	okijø	
aupristé	opriste	
auvissable	ovisabl	
carvassant	kaʁvasɑ̃	C-initial
persénible	pɛʁsenibl	

Table 21: *Experiment 1 stimuli*

5.1.3 Participants

51 participants were recruited from Amazon's *Mechanical Turk* (*cf.* Schnoebelen & Kuperman 2010, Sprouse 2011, Gibson, Piantadosi & Fedorenko 2011), a web application that provides access to survey-based experiments to participants located around the globe. The study was advertised as a French language survey, targeting speakers located exclusively in Paris, France. Each participant was directed to the survey, which was conducted over *Experigen* (Becker & Levine 2013), an online platform for survey-based experiments. Speakers were paid €2.75 for their participation. The average participant took about ten minutes to finish the survey. Other

than location, no prerequisites were put in place to exclude Mechanical Turkers from participating. After taking the survey, speakers were asked for their age, gender, whether they have hearing or any other language-related impairments, and whether they were a native speaker of French. *Post hoc* criteria for exclusion were as follows: the participant indicated non-native speaker status; the data reveal that the participant chose the liaison form before consonant-initial nonces on more than one trial; the participant chose the liaison form before the great majority of vowel-initial nonces — in particular, they chose the non-liaison form before at most two vowel-initial nonces. After applying these criteria, 2 participants were excluded for indicating non-native speaker status, 5 for indicating an impairment, 6 for choosing a liaison form before a consonant-initial nonce more than once, and 10 for choosing a liaison form in the vast majority of trials with vowel-initial nonces (≤ 2 exceptions out of 48). This left 28 participants who displayed evidence for variable liaison.

5.1.4 Results

We first give the different corpus rates given in three studies—de Jong (1994)'s study, Mallet (2008)'s study, and mine, covering the PFC corpus—followed by the results of forced-choice part of the experiment. We find that the speakers overall replicated the distinctions found in the corpora:

	très, 'very'	*plus*, 'more'	*pas*, 'not'
de Jong (1994)	99%	96%	7%
Mallet (2008)	97%	64%	1%
My study, PFC	95%	79%	0%
AVERAGE:	**97%**	**80%**	**3%**

Table 22: *corpus liaison rates for très, plus, and pas*

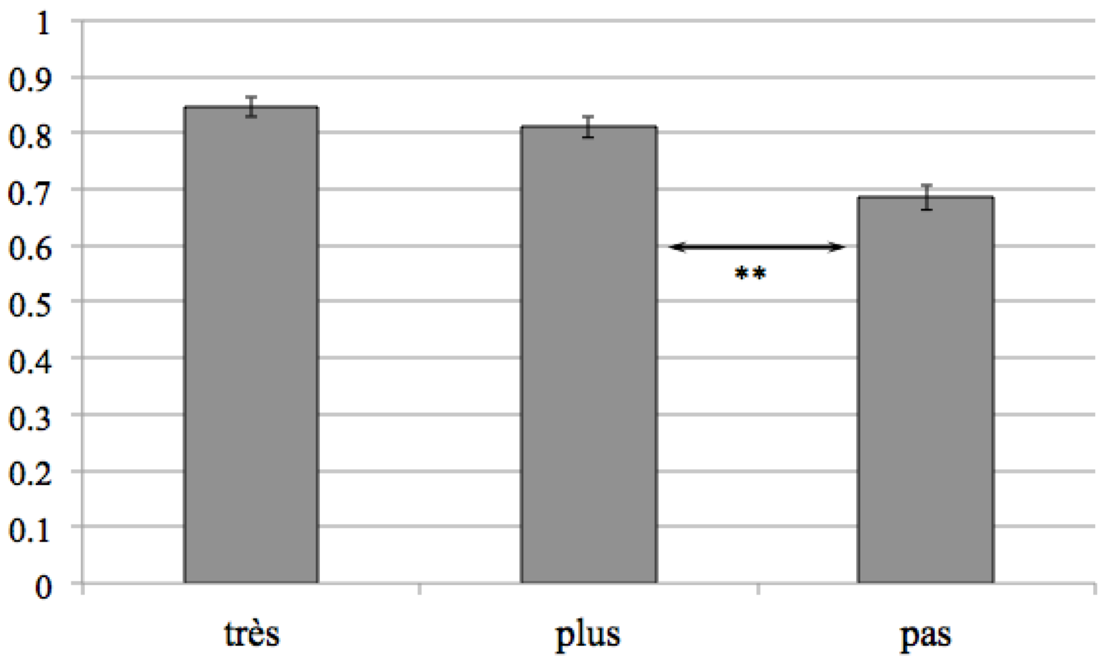


Figure 15: *Forced-choice Experiment 1 liaison rates for très, plus, and pas*

However, experiment rates were much higher than corpus rates overall. In particular, the experiment results reveal a gradient tendency for *pas* to liaise — a 69% liaison rate in the binary choice task — even though its corpus rates in the de Jong and Mallet studies and in the PFC stand between 0% and 7%.Why might this have occurred? There are a number of possibilities. I conjecture that this is due to the prescriptive pressure to maintain liaison in formal environments (e.g., Moisset 2000, Armstrong 2001 on style/register as a conditioner), because the stimuli were read by participants before any judgments were made[6], and because speakers knew that this experiment was targeting their knowledge of liaison. Moreover, the contrast between *très*, *plus*, and pas is more compressed in the experimental results than in the corpus. I conjecture that the compression effect here is due to the same reasons given in Hayes, Zuraw et al. (2009), who also obtained the this effect in their experiment. Hayes & Londe (2006) present results of a nonce probe study into Hungarian vowel harmony, showing that speakers closely frequency match to

[6] Thanks to Myriam Lapierre for raising this possibility.

statistical regularities found within the lexicon (see Section 2.2.2). In this experiment, speakers were asked to *volunteer* responses on their own. Hayes, Zuraw et al. (2009) reconduct the nonce probe experiment, and largely replicate Hayes & Londe (2006)'s findings, but obtain nonce rates that are compressed relative to the corpus rates and to the rates of the prior experiment. In Hayes, Zuraw et al.'s experiment, speakers were *offered* potential responses, and the investigators surmise that the compression effect is due to speakers being more likely to accept an unusual form when being presented with it, but less likely to volunteer it.

Are the differences statistically significant? To check this, a mixed-effects logistic regression model was fit to the results using the `glmer` function of the *lme4* package in R, containing Word1 coded as a fixed effect (*plus* as reference level), and random intercepts for nonce adjective and participant, and a random slope relating participant to Word1. Model output is given in the table below. Relative to reference level *plus*—which undergoes liaison significantly more than chance—*très* undergoes liaison only slightly more, while *pas* undergoes liaison significantly less. Though the contrast between *très* and *plus* are not significant in this experiment, the results nevertheless mirror the distinctions obtained in the corpus studies: liaison was preferred more with *très* (85%) than *plus* (81%), and more with *plus* than *pas* (69%). The variance of the random intercept for nonce was very close to zero, suggesting that nonce adjectives did not differ substantially in their propensity to trigger liaison. The variances of the random slope for participant were nonzero, suggesting variation in individual participants' contrasts between the five Word1's, as well as participant-specific variation in the results overall.

Random effects	*Variance*			
Nonce	0.01			
Participant: *plus*	2.84			
Participant: *très*	2.66			
Participant: *pas*	1.02			
Fixed effects	*Coef.*	*S.E.*	*z*	*p*
plus **(ref.)**	2.07	0.37	5.58	< 0.001
très	0.06	0.32	0.20	0.84
pas	-1.13	0.38	-2.93	0.003

Table 23: *output of a mixed-effects logistic regression model of Experiment 1 forced-choice task results*

The results of the five-point scale, given below, also generally confirm that speakers track lexical propensities for liaison:

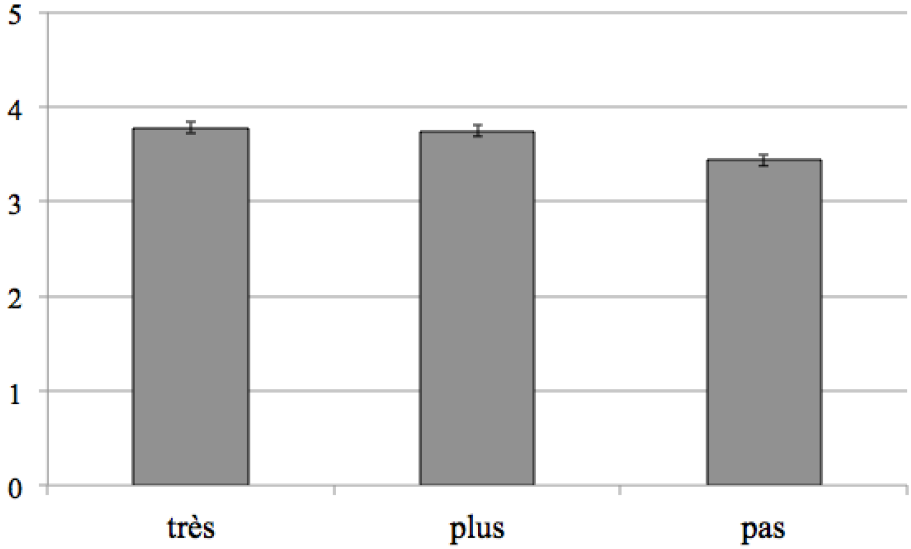


Figure 16: *Five-point scale Experiment 1 liaison scores for très, plus, and pas*

In particular, *très* and *plus* group together above *pas*. A mixed-effects linear regression model was fit to the results using the `lmer` function of the *lme4* package in R, containing Word1 coded as a fixed effect (with *plus* as reference level, to assess whether contrasts with neighboring words are significant), and random intercepts for nonce adjective and participant, and a random slope relating participant to Word1. Model output is given in the table below. Relative to reference level *plus*—whose liaison form is rated significantly higher than chance—*très* undergoes liaison only slightly more, while *pas* undergoes liaison significantly less. Note that *t*-values with absolute magnitude greater than 2 indicate significance of a predictor. The variance of the random intercept for nonce was, again, very close to zero, suggesting that nonce adjectives did not differ substantially in their propensity to trigger liaison. The variances of the random intercept and slope for participant were nonzero, again suggesting variation in individual participants' contrasts between the five Word1's, as well as participant-specific variation in the results overall.

Random effects	*Variance*
Nonce	0.01
Participant: *plus*	0.48
Participant: *très*	0.41
Participant: *pas*	0.13

Fixed effects	*Coef.*	*S.E.*	*t*
plus **(ref.)**	3.75	0.14	26.39
très	0.03	0.10	0.32
pas	-0.31	0.14	-2.16

Table 24: *output of a mixed-effects linear regression model of Experiment 1 ratings study results*

Overall, these results suggest that learners do not merely internalize fixed pronunciations together with general trends across the lexicon; rather, they also track rates in words with vacillating pronunciations. It is not particularly surprising that *très* and *plus* lack a significant contrast, considering that de Jong's (1994) study finds close corpus rates between the two words.

What is more surprising is the experimental result for *pas*— a 69% liaison rate in the binary choice task — even though corpus rates are very low. This further substantiates that speakers' knowledge of liaison applicability to different words is quantitative in nature (*cf*. Zuraw & Hayes 2017), despite the binary patterning of spoken language data across a subset of words (e.g., *pas*). We now probe further into French speakers' knowledge of lexical gradience, assessing whether they internalize the distinct propensities of five different adverbs: *très*, *plus*, *bien*, *moins*, and *pas*.

5.2 Experiment 2: Testing for the internalization of lexical propensities associated with *très*, *plus*, *bien*, *moins*, and *pas*

5.2.1 Task

The task was identical to that of the previous experiment.

5.2.2 Stimuli

The experiment tests for whether speakers internalize the effect of word identity, now across *five* different adjective modifiers when they precede nonce adjectives. The stimuli were recorded by the same speaker who recorded for Experiment 1. The three items tested were *très* ('very'), *plus* ('more'), and *pas* ('not'), which yield different rates in the corpus studies:

	très, 'very'	*plus*, 'more'	*bien*, 'well'	*moins*, 'less'	*pas*, 'not'
Mallet (2008)	97%	64%	43%	—	1%
de Jong (1994)	99%	96%	82%	—	7%
My study, PFC	95%	79%	85%	26%	0%
AVERAGE:	**97%**	**80%**	**70%**	**26%**	**3%**

Table 25: *corpus liaison rates for très, plus, bien, moins, and pas*

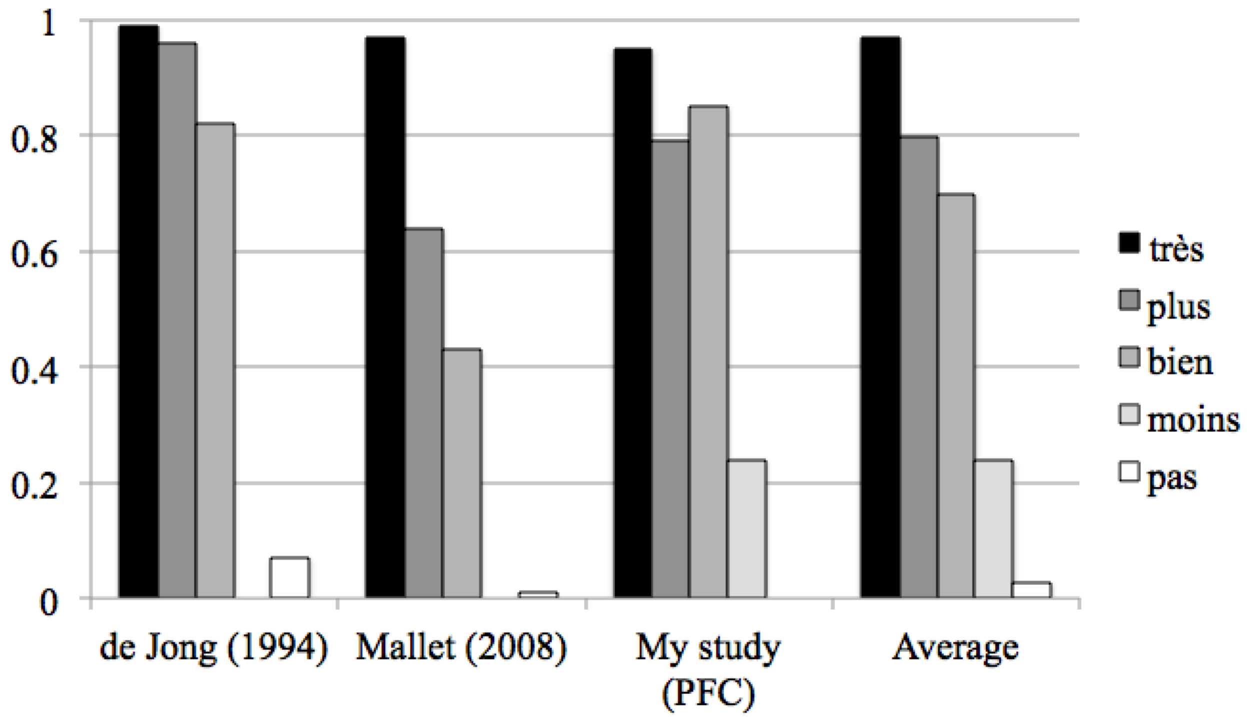


Figure 17: *corpus liaison rates for très, plus, bien, moins, and pas*

The literature categorizes each of these words as adverbs (*cf.* Côté 2011, 2013). All three words are monosyllabic and end in the same liaison consonant, ruling out syllable count and consonant identity as explanations for the difference in rate. Finally, the direct but nonetheless weak relationship between Word1 frequency and liaison rate cannot alone explain the three different rates: though *plus* is much less frequent than *très*, *pas* is much more frequent than *plus*, and yet has a much lower liaison rate. This leaves word identity as a viable hypothesis for differences in rate.

Participants heard each of these words before eighteen nonce adjectives, each of which had adjectival suffixes. Two of these nonce adjectives were consonant-initial, and were used only to assess whether the participant internalized the correct environment of liaison. The other sixteen were vowel-initial — four groups of four words with the same initial vowel. Each trial consisted of an adverb-nonce adjective bigram, totaling to 3 × 18 = 54 bigrams/trials. The stimuli are presented below: we have the five attested Word1s, plus the same stimuli for Word2 used in Experiment 1.

Adv.	IPA
très	tʁɛ/tʁɛz
plus	ply/plyz
bien	bjɛ̃/bjɛn
moins	mwɛ̃/mwɛ̃z
pas	pa/paz

Nonce adj.	IPA	Status
arvant	aʁvɑ̃	V-initial (critical)
agrivieux	agʁivjø	
amagné	amaɲe	
altimable	altimabl	
énantant	enɑ̃tɑ̃	
éprieux	epʁijø	
émoivré	emwavʁe	
écastable	ekastabl	
impergeant	ɛ̃pɛʁʒɑ̃	
invinieux	ɛ̃vinjø	
introché	ɛ̃tʁɔʃe	
ingroutable	ɛ̃grutabl	
autrillant	otʁijɑ̃	
auquieux	okijø	
aupristé	opriste	
auvissable	ovisabl	
carvassant	kaʁvasɑ̃	C-initial
persénible	pɛʁsenibl	

Table 26: *Experiment 2 stimuli*

5.2.3 Participants

72 participants located in Paris, France were recruited from Amazon's *Mechanical Turk*, and directed to the *Experigen* survey. Speakers were paid €3.25 for their participation. The average participant took about fifteen minutes to finish the survey. Again, other than location, no

prerequisites were put in place to exclude Mechanical Turkers from participating, and after they took the survey, they were asked for their age, gender, whether they have hearing or any other language-related impairments, and whether they were a native speaker of French. *Post hoc* criteria for exclusion were as in the first experiment. 6 participants were excluded for indicating non-native speaker status, 1 for indicating an impairment, 9 for choosing a liaison form before a consonant-initial nonce more than once, and 14 for choosing a liaison form in the vast majority of trials with vowel-initial nonces (≤ 2 exceptions out of 80). This left 42 participants who displayed evidence for variable liaison. For one participant, judgments were recorded in all but one frame, for unknown reasons; the rest of this participant's data were included in the analysis.

5.2.4 Results

We first give the different corpus rates given in three studies—de Jong (1994)'s study, Mallet (2008)'s study, and mine, covering the PFC corpus—followed by the results of forced-choice part of the experiment. We find, once again, that the speakers overall replicated the distinctions found in the corpora:

	très, 'very'	*plus*, 'more'	*bien*, 'well'	*moins*, 'less'	*pas*, 'not'
Mallet (2008)	97%	64%	43%	—	1%
de Jong (1994)	99%	96%	82%	—	7%
My study, PFC	95%	79%	85%	26%	0%
AVERAGE:	**97%**	**80%**	**70%**	**26%**	**3%**

Table 27: *corpus liaison rates for très, plus, bien, moins, and pas*

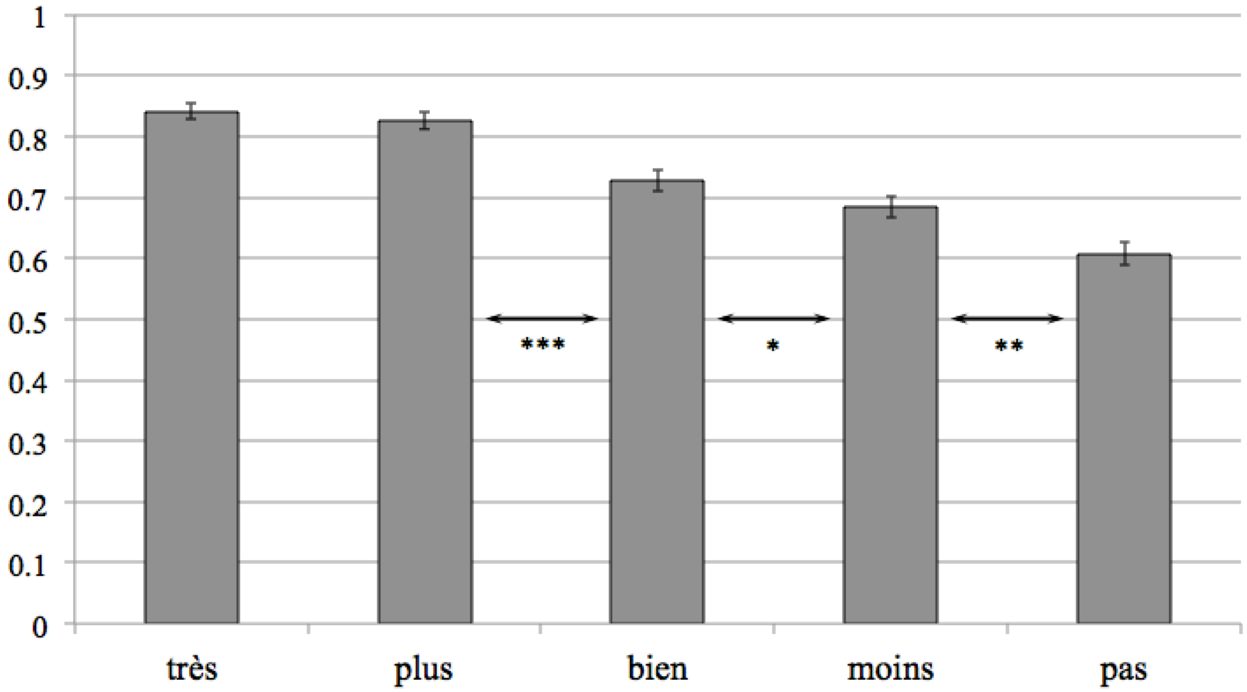


Figure 18: *Forced-choice Experiment 2 liaison rates for très, plus, bien, moins, and pas*

Again, experiment rates were overall higher and more compressed than corpus rates (see Section 5.1.4 for discussion as to why this might be the case). A mixed-effects logistic regression model was fit to the results using the `glmer` function of the *lme4* package in R, containing Word1 coded as a fixed effect, and random intercepts for nonce adjective and participant, and a random slope relating participant to Word1. I present results first with *plus* coded as reference level, and then with *moins* coded as reference level, to assess whether *plus* and *moins* differ significantly in their results relative to their neighboring words. Model output is presented in the tables below. Relative to *plus*—which undergoes liaison significantly more than chance—*très* undergoes liaison only slightly more, while *bien* undergoes liaison significantly less. The non-neighboring words *moins* and *pas* also undergo liaison significantly less than *plus*.

(*Random effects:* see Table 28b)

Fixed effects	*Coef.*	*S.E.*	*z*	*p*
plus **(ref.)**	2.16	0.27	7.94	< 0.001
très	0.17	0.25	0.70	0.49
bien	-0.76	0.20	-3.76	< 0.001
moins	-1.15	0.21	-5.53	< 0.001
pas	-1.59	0.22	-7.25	< 0.001

Table 28a: *output 1 of a mixed-effects logistic regression model of Experiment 2 forced-choice task results*

Relative to *moins*—which undergoes liaison significantly more than chance—*bien* undergoes liaison significantly more, while *pas* undergoes liaison significantly less. The non-neighboring words *très* and *plus* undergo liaison significantly more than *moins*. The results nevertheless mirror the distinctions from de Jong's, Mallet's, and my corpus investigations: liaison was preferred more with *très* (85%) than *plus* (83%), more with *plus* than *bien* (73%), more with *bien* than *moins* (68%), and more with *moins* than *pas* (61%). The variance of the random intercept for nonce was close to zero, suggesting that nonce adjectives did not differ substantially in their propensity to trigger liaison. The variances of the random slope for participant were nonzero, suggesting variation in individual participants' contrasts between the five Word1's, as well as participant-specific variation in the results overall.

Random effects	*Variance*
Nonce	0.05
Participant: *très*	0.93
Participant: *plus*	0.40
Participant: *bien*	0.29
Participant: *moins*	1.47
Participant: *pas*	0.15

Fixed effects	*Coef.*	*S.E.*	*z*	*p*
moins **(ref.)**	1.02	0.22	4.61	< 0.001
très	1.32	0.25	5.35	< 0.001
plus	1.14	0.21	-5.53	< 0.001
bien	0.38	0.18	2.29	0.028
pas	-0.44	0.15	-2.98	0.003

Table 28b: *output 2 of a mixed-effects logistic regression model of Experiment 2 forced-choice task results*

Finally, the results of the five-point scale, given below, also generally confirm that speakers track lexical propensities for liaison. *Très* and *plus* group together above *bien* and *moins,* and *bien* and *moins* group together above *pas*. Dashed horizontal lines are illustrated below to indicate these groupings.

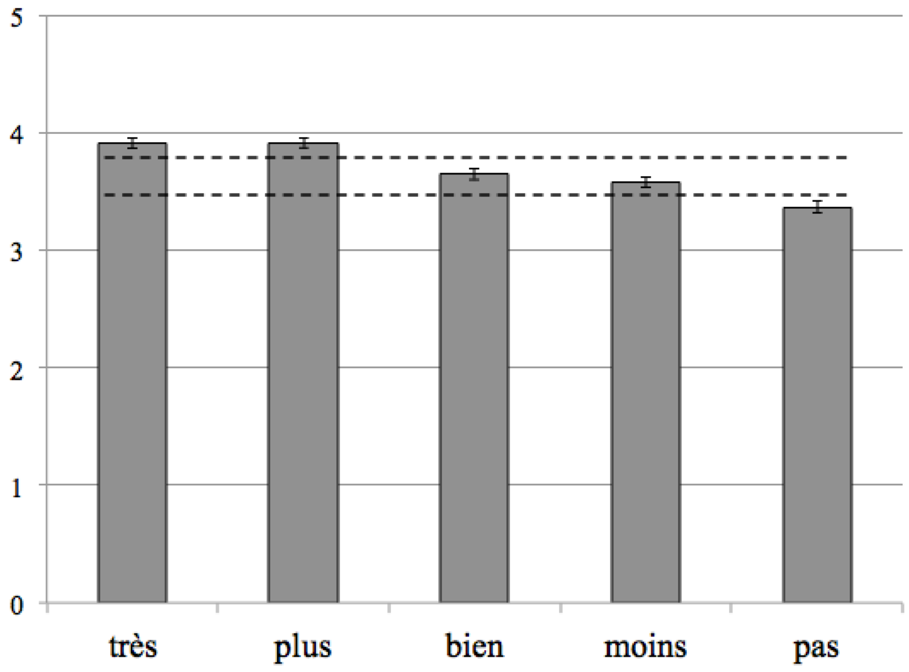


Figure 19: *Five-point scale experiment liaison scores for très, plus, bien, moins, and pas*

Once again, I present results first with *plus* coded as reference level, and then with *moins* coded as reference level, to assess whether *plus* and *moins* differ significantly in their results relative to their neighboring words. Relative to *plus*—whose liaison form is rated significantly higher than chance—*très* has an essentially identical liaison rating, while *bien* has a significantly lower liaison rating. The non-neighboring words *moins* and *pas* have liaison ratings significantly lower than *plus*.

(*Random effects:* see Table 29b)

Fixed effects	*Coef.*	*S.E.*	*t*
plus **(ref.)**	3.93	0.10	40.50
très	0.00	0.08	-0.02
bien	-0.26	0.08	-3.14
moins	-0.33	0.08	-4.26
pas	-0.54	0.11	-5.00

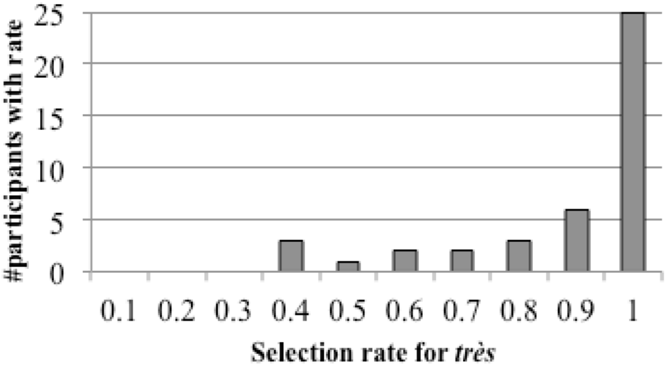
Table 29a: *output 1 of a mixed-effects linear regression model of Experiment 2 ratings study results*

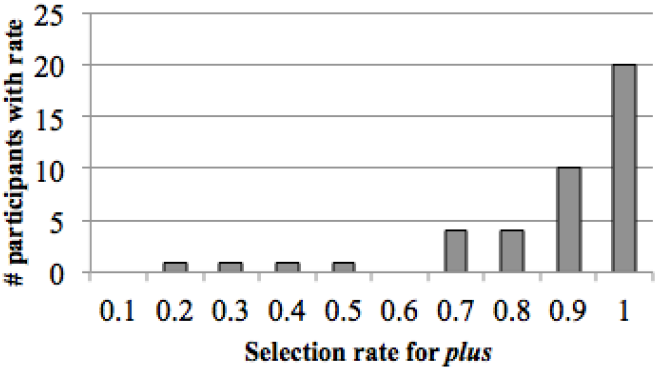
Relative to *moins*—whose liaison form is rated significantly higher than chance—*bien* has a non-significantly higher liaison rating, while *pas* has a significantly lower liaison rating. The non-neighboring words *très* and *plus* have liaison ratings significantly higher than *moins*. Once again, the variance of the random intercept for nonce was close to zero, suggesting that nonce adjectives did not differ substantially in their propensity to trigger liaison. The variances of the random slope for participant were nonzero, suggesting that variation emerges in individual participants' contrasts between the five Word1's.

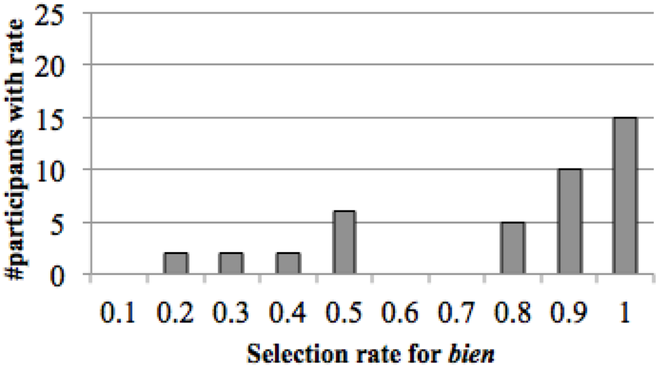
Random effects	*Variance*		
Nonce	0.01		
Participant: *très*	0.27		
Participant: *plus*	0.12		
Participant: *bien*	0.05		
Participant: *moins*	0.37		
Participant: *pas*	0.17		
Fixed effects	*Coef.*	*S.E.*	*t*
moins **(ref.)**	3.59	0.10	34.48
très	0.33	0.09	3.37
plus	0.33	0.08	4.26
bien	0.07	0.07	1.07
pas	-0.21	0.09	-2.45

Table 29b: *output 2 of a mixed-effects linear regression model of Experiment 2 ratings study results*

One might ask whether the gradience observed between words is simply the result of more participants choosing the liaison form categorically for one word over another. That is to say, it is in theory possible that the participants chose the liaison or non-liaison form of a word in a categorical manner—displaying no within-word variation—and that the gradience above (e.g., *très* receiving a higher rate than *bien*) is simply the result of *more participants* categorically choosing liaison form of one word over another. To assess whether this was the case, I composed histograms plotting liaison form selection rate in the binary choice task for a given word against the number of participants that selected the liaison form at that rate. Five histograms are given below for *très*, *plus*, *bien*, *moins*, and *pas*. We observe the following: 1) participants displayed considerable within-word variation, and did not merely choose the liaised or non-liaised form of a given word categorically; and 2) *très* and *plus* were associated with greater numbers of participants selecting liaison consistently, followed by *bien*, then *moins*, and finally *pas*. In particular, as we move from *très* and *plus* to *pas*, we find that participants preferred liaison with decreasing consistency.







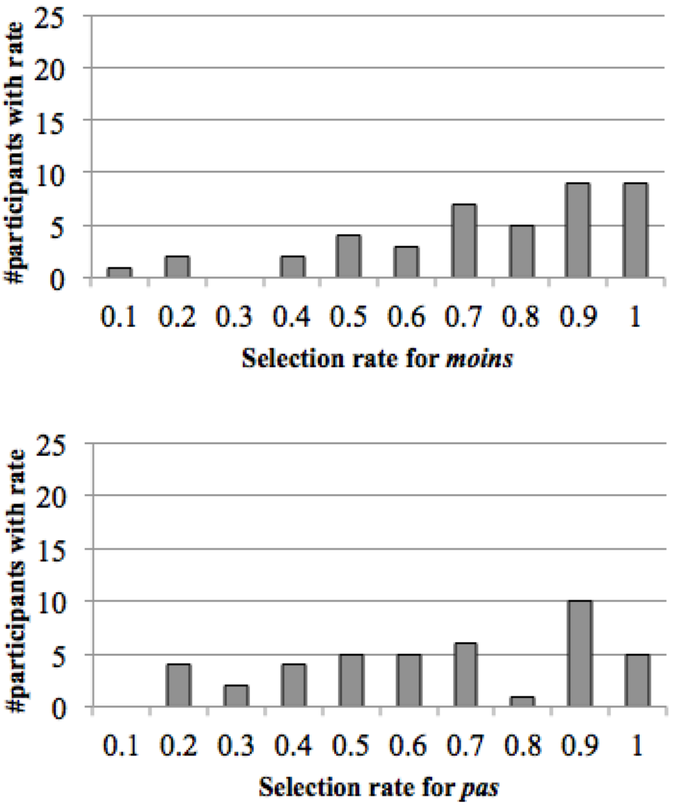


Figure 20: *histograms plotting rate against number of participants*

These observations suggest that the word-specific gradience observed in the overall experimental rates are not exclusively the result of categorical preferences—i.e., more participants categorically choosing liaison for one word over another. Rather, they display structured variation in the *degree* to which they prefer liaison, tending to prefer it more after *très* and *plus*, less so after *bien*, even less so after *moins*, and still even less so after *pas*.

5.3 Summary and implications

To review, we find that the results mirror the distinctions obtained in de Jong's, Mallet's, and my corpus investigations: liaison was preferred more with *très* and *plus* than with *bien*, more with *bien* than *moins*, and more with *moins* than *pas*. Overall these results suggest that learners do not merely internalize fixed pronunciations together with overall rates in the lexicon; rather, they also track rates in words with vacillating pronunciations. The results support theories in which the triggering or undergoing status of individual morphemes is implemented on an entire spectrum (Moore-Cantwell & Pater 2016, Smolensky & Goldrick 2016, Zuraw & Hayes 2017, Tanaka 2017). Ultimately, the theory of the grammar and lexicon must elucidate how lexical propensities are represented and learned.

The results also have implications for lexical indexation. Prior research suggests that speakers of any given language memorize a great number of word pairs and access the pairs wholesale in online production (Zuraw 2000, 2010; Bybee 2001, 2002). But the nonce probe task results show that rate tracking can occur on the level of a *single* morpheme (here, word)—it cannot be that all the variation acquired by speakers can be chalked up to memorizing a large set of word pairs. Recent literature on liaison has referred to frequent Word1-Word2 pairs in their account of variation in French liaison (*cf*. Bybee 2001, 2002), and while factors for word pairs may improve fit to the data, my experiment results suggest that speakers do indeed acquire propensities associated with *individual* words. Thus morphophonological theory must be able to refer to propensities of individual morphemes. If it were not capable of doing so, then we would not be able to predict that speakers replicate corpus distinctions for *très*, *plus*, *bien*, *moins* and *pas* when they come before novel adjectives.

Chapter 6:

Modeling the learning of a frequency-matching grammar together with lexical propensities

This section investigates how to model the learning and representation of lexical variation. How can we model the language learner who frequency matches to trends across the lexicon while acquiring lexical idiosyncrasy—lexical propensities in particular? As discussed in Section 2.2.1, Zuraw (2000, 2010) adopts OT with stochastic ranking and the Gradual Learning Algorithm to model the learning of lexical trends together with idiosyncrasies, using general constraints to capture trends, and constraints enforcing listedness together with high-ranking faithfulness constraints to capture idiosyncrasies.[7] But recent research (Zuraw & Hayes 2017, Smith & Pater 2017) has challenged stochastic OT as a framework for capturing variation, obtaining that probabilistic Harmonic Grammar is capable of handling a broader range of paradigms.

A recent popular approach aims to capture trends with idiosyncrasy in Maximum Entropy Harmonic Grammar, which generates patterns of gradience through constraint weighting rather than ranking. The strategy has been to recruit general, grammatical constraints to model frequency-matching behavior in nonce probe studies, together with lexical(ly indexed) constraints to model lexical idiosyncrasies in the dataset (Moore-Cantwell & Pater 2016, Zuraw & Hayes 2017, Tanaka 2017). Though this approach is capable of representing lexical propensities, this section shows with a series of learning simulations that it encounters a

[7] See also Nazarov (2018) for implementation and some early results of a new model of lexical variation couched in stochastic OT.

GRAMMAR-LEXICON BALANCING PROBLEM: lexical constraints are so powerful in explaining the dataset that that the learner comes to acquire the behavior of each form using only these constraints, at which point the general constraint is rendered ineffective.

In particular, as learning proceeds, the grammatical constraint explains an infinitesimal portion of the dataset; at this point, its weight begins to vacillate, as a wide range of values for its weight fit the dataset well. Once the dataset is learned perfectly using lexical constraints, the weight of the grammatical constraint plummets to zero. I claim that the choice to embed both grammatical and lexical constraints in Maximum Entropy Harmonic Grammar in particular is what leads to the imbalance. In MaxEnt, the grammatical constraint and lexical constraints are treated as *equally viable* explanatory devices for learning the dataset and its patterns, with the learner favoring neither variety of constraint in particular during the learning process. The weights of the lexical constraints therefore rise rapidly in magnitude to explain the dataset, with the grammatical constraint coming to explain less and less of the data, eventually leading to the convergence problem.

I attribute these results, as well the findings that real language learners are nonetheless capable of generalizing across idiosyncratic variation, to them possessing a GENERALITY BIAS: they privilege general, grammatical constraints over the more granular lexical constraints when they acquire variable datasets. It is argued that MaxEnt in its current formulation—essentially a canonical logistic regression model—fails to appropriately represent this property, even after taking into consideration a prior penalty term. This section provides a solution to the grammar-lexicon balancing problem by replacing MaxEnt with a hierarchical and similarly logistic model, the mixed-effects logistic regression model—i.e., MIXED-EFFECTS MAXIMUM ENTROPY HARMONIC GRAMMAR. Mixed-Effects MaxEnt is shown to succeed in learning both general and item-specific behavior by encoding general constraints as main effects, and lexical constraints as random effects. Generality bias is rooted in the fixed effect-random effect distinction: though coefficients of the general constraints are fit to the data directly to match overall rates, the coefficients of the levels of the random intercept are determined by a weighted average of the word-specific rate *and* the overall rate in the dataset. The learner treats the grammar and lexicon differently upon positing the distinction between main effect and random effect, such that idiosyncratic effects of the vocabulary are subordinated to broad, grammatical effects in the learning process.

Hierarchical mixed models are used widely across scientific fields, and a growing family of research has employed random intercepts to measure the degree of by-word or by-lexical class idiosyncrasy in datasets displaying morphophonological variation (Fruehwald 2012, Shih & Inkelas 2016, Zuraw & Hayes 2017, Smith & Moore-Cantwell 2017, Shih 2018); Shih & Inkelas (2016) and Shih (2018) even adopt the hierarchical mixed model as a theory of the language learner. Here I present an argument that adopting as a theory of the language learner the *mixed-effects logistic regression model in particular* is a crucial step toward capturing the capabilities of language learners: it can learn and represent lexical trends together with idiosyncrasies, while models couched in simple logistic regression, such as the current formulation of MaxEnt, seemingly cannot.

6.1 Statistical generalizations over idiosyncratic forms and frequency matching by language learners

Any account of variation would have to capture: (i) the idiosyncratic behavior of different morphemes; (ii) statistical generalizations over these morphemes. As it pertains to the latter, we seek to predict the frequency matching behavior of language learners using the real language data to which they are exposed (Eddington 1996, 1998, 2004; Frisch, Broe, & Pierrehumbert 1996; Coleman & Pierrehumbert 1997; Berkley 2000; Zuraw 2000, 2010; Bailey & Hahn 2001; Frisch & Zawaydeh 2001; Pierrehumbert 2002; Albright 2002; Albright & Hayes 2003; Ernestus & Baayen 2003; Hayes & Londe 2006; *et seq*).

Recent research adopts a model for learning a frequency matching grammar together with lexical idiosyncrasy using Maximum Entropy Harmonic Grammar (Moore-Cantwell & Pater 2016, Zuraw & Hayes 2017, Tanaka 2017). Taking a toy example from Moore-Cantwell & Pater (2016), if we suppose that ALIGN-R and NONFIN have weights of 4 and 1 respectively, the mathematics behind MaxEnt would yield penultimate stress around 95% of the time:

/bætækæ/	*p*	*H*	ALIGN-R 4	NONFIN 1
bə(ˈtækə)	95%	-1		-1
(ˈbætə)kə	5%	-4	-1	

Table 30a: *A grammar choosing penultimate stress 95% of the time (Pater & Moore-Cantwell, p. 56)*

Moreover, setting the weights of ALIGN-R and NONFIN to be equal—e.g., at 2 and 2—yields 50-50 variation between penultimate and antepenultimate stress:

/bætækæ/	*P*	*H*	ALIGN-R *w* = 2	NONFIN *w* = 2
bə(ˈtækə)	50%	-2		-1
(ˈbætə)kə	50%	-2	-1	

Table 30b: *A grammar choosing penultimate stress 50% of the time (Pater & Moore-Cantwell, p. 56)*

But as Moore-Cantwell & Pater (2016) note, the tableaux above would only be appropriate if the two competing forms occurred in free variation. As we have already seen (see Chapters 2, 3, and 4), statistical generalizations are found only across the lexicon as a whole, with individual words primarily displaying fixed pronunciations. This, for example, holds true for French liaison and Slovenian palatalization—the cases surveyed in this dissertation—which display a U-shaped distribution. If we consider all words in the data set, most words are either clear liaisers/palatalizers or non-liaisers/non-palatalizers, and only a minority of words vary in their pronunciation. This is visible in the tables below.

Slovenian palatalization (*Dict. Standard Slovenian* & *Gigafida*)		
	Number of words	% whole data set
Words with rate greater than 95%:	3735	77%
Words with rate between 5% and 95%:	405	8%
Words with rate less than 5%:	701	15%

Table 31a: *Extreme word-level palatalization rates in Slovenian*

French liaison (*PFC*)		
	Number of words	% whole data set
Bigrams with rate greater than 95%:	2803	25%
Bigrams with rate between 5% and 95%:	325	3%
Bigrams with rate less than 5%:	8272	72%

Table 31b: *Extreme bigram-level liaison rates in French*

Of course, the counts in the table above include many words with frequency count 1, exaggerating the end points. But even if we consider Slovenian words and French bigrams occurring only ten or more times in the corpus, we still find that most of them have fixed pronunciations:

Slovenian palatalization (*Dict. Standard Slovenian* & *Gigafida*)		
	Number of words	% whole data set
Words with rate greater than 95%:	1960	81%
Words with rate between 5% and 95%:	162	7%
Words with rate less than 5%:	284	12%

Table 32a: *Extreme word-level palatalization rates in Slovenian, word frequency* ≥ 10

French liaison (*PFC*)		
	Number of words	% whole data set
Bigrams with rate greater than 95%:	171	34%
Bigrams with rate between 5% and 95%:	127	25%
Bigrams with rate less than 5%:	209	41%

Table 32b: *Extreme bigram-level liaison rates in French, word frequency* ≥ 10

To illustrate the problem, consider the Slovenian palatalization data in Table 31a. We find that approximately 75% of words regularly palatalize, while 25% fail to regularly palatalize. Classifying variable palatalizers as non-palatalizers (for simplicity purposes), we might conclude that the grammar palatalizes across roughly 75% of the lexicon. Thus if a Slovenian speaker were to serve as participant in a nonce probe study which takes the form of a two-alternative forced choice task and which tests knowledge of variable palatalization, we would expect her to select palatalized nonce forms roughly 75% of the time, per the law of frequency matching. The analyst might conclude that the participant has set the weight a constraint driving palatalization—PAL in the tableau below (following Jurgec 2016)—to 2, and the counteracting faithfulness constraint to 1, yielding the trend.

Nonce form with subsequence /...k+i .../	*p*	*H*	PAL 2	IDENT 1
ki	25%	-2	-1	
ci	75%	-1		-1

Table 33: *Nonce probe data*

As for attested words with fixed pronunciation—e.g., (/nag+ʦ/, [nag-əʦ], 'naked'-DIM) and (/dvonog+ʦ/, [dvonoʒ-əʦ], 'biped'-DIM) —the account makes incorrect predictions, at least without further intervention. It fails to predict the idiosyncratic, categorical behavior of either word—that is to say, the fact that each of their pronunciations are fixed: the model selects palatalization for each of these forms merely 70% of the time, rather than 100% and 0% of the time respectively, resulting in severe analytical error at the level of word:

/nag+ʦ/	Observed rate	*p*	*H*	PAL 2	IDENT 1
nag-əʦ	100%	25%	-2	-1	
naʒ-əʦ	0%	75%	-1		-1

Table 34a: *Failure to predict categorical non-palatalization in* [nag-əʦ]

/dvonog+ʦ/	Observed rate	*p*	*H*	PAL 2	IDENT 1
dvonog-əʦ	0%	25%	-2	-1	
dvonoʒ-əʦ	100%	75%	-1		-1

Table 34b: *Failure to predict categorical palatalization in* [peʃ-əʦ]

Moore-Cantwell & Pater (2016) respond to this problem by proposing that the grammar contains general constraints that regulate whole sets of forms, as well as form-specific constraints—called *lexically indexed constraints*—that regulate the idiosyncratic behavior of some attested form in particular (also Kraska-Szlenk 1995, Pater 2000, *et seq*; and relatedly, Pater, Staubs, Smith & Jesney 2012, Smith 2015, Zuraw & Hayes 2017; *inter alia*). For example, on top of the general constraints PAL and IDENT, we might posit highly weighted constraints PAL-/dvonog+ʦ/, and IDENT-/nag+ʦ/ to derive the idiosyncratic behavior of (/nag+ʦ/, [nag-əʦ]) and (/dvonog+ʦ/, [dvonoʒ-əʦ]) in particular, much as Moore-Cantwell & Pater (2016) posit lexically-specific

ALIGN-R and NONFIN constraints to get idiosyncratic stress placement (see Section 2.3.1). Their activity is shown in the tableaux below:

	Observed	*p*	*H*	PAL	PAL-/dvonog+ʦ/	IDENT	IDENT-/nag+ʦ/
/nag+ʦ/	rate			2	5	1	6
nag-əʦ	100%	100%	-2	-1			
naʒ-əʦ	0%	0%	-7			-1	-1

Table 35a: *Prediction of categorical non-palatalization in* [nag-əʦ] *using lexical constraints*

	Observed	*p*	*H*	PAL	PAL-/dvonog+ʦ/	IDENT	IDENT-/nag+ʦ/
/dvonog+ʦ/	rate			2	4	1	6
dvonog-əʦ	0%	0%	-6	-1	-1		
dvonoʒ-əʦ	100%	100%	-1			-1	

Table 35b: *Prediction of categorical palatalization in* [peʃ-əʦ] *using lexical constraints*

Under this approach, it would appear that we are able to capture both the frequency-matching behavior of speakers in nonce probe studies—for example, by using the general constraint PAL to match to the overall palatalization rate—and idiosyncrasies of the lexicon, using lexically specific constraints. Yet questions remain: what weights does the model obtain for general constraints and lexically specific constraints when we leave it to learn a dataset displaying statistical generalizations together with lexical propensities? Are both the grammar and the lexical knowledge sustained throughout model learning?

6.2 MaxEnt fails to learn statistical generalizations together with idiosyncrasy: the grammar-lexicon balancing problem

We have reviewed a recent popular MaxEnt-based model for learning statistical generalizations together with lexical propensities: use general constraints to mimic language learners' behavior to frequency match to statistical generalizations, and lexical constraints to capture idiosyncrasy. The following investigation scrutinizes this approach to assess whether it in fact succeeds in learning generalizations together with idiosyncrasies. In a footnote, Moore-Cantwell & Pater (2016) mention that "the parameter settings can affect the outcome, especially the setting of the regularization term. The need to tune the parameters to match the experimental data is a potential weakness of this approach" (p. 62). Tanaka (2017) further surmises that their approach may lead to overfitting of lexical constraints to the idiosyncratic data—and to underfitting of grammatical constraints to lexical trends—if learners are not properly biased to favor grammatical constraints over lexical constraints. We find in what follows is that the concern is indeed well-founded: a series of learning simulations reveal that the model fails to learn a frequency matching grammar in the face of lexical idiosyncrasy. Under the MaxEnt-based approach, lexical constraints are so powerful that they come to explain the entire dataset, to the point where general constraints are rendered ineffective.

6.2.1 MaxEnt fails to learn statistical generalizations with strict exceptionality

Suppose we wanted to model a variable phonological system in which 5,000 forms behave regularly with respect to the grammar, and 100 forms behave irregularly. This sort of dataset resembles English plural or past-tense formation, in which the majority of forms behave regularly, but a small set of forms are exceptional, undergoing a different rule (e.g., *beep*, *beeps*;

jeep, *jeeps*; *creep*, *creeps*; but *sheep*, *sheep*). This dataset also corresponds quite well with the so-called *hi:d* stem paradigm in Hungarian (Hayes & Londe 2006): although stems containing a single front unrounded vowel usually take the harmonic *–nɛk* suffix in dative constructions, roughly 8% of them in a corpus take the disharmonic form of the suffix, *–nɔk* (so-called *hi:d* stems, as in [hi:d-nɔk], 'bridge'-DAT; Hayes & Londe 2006; p. 63, 66); moreover, speakers closely match the 8% irregularity rate in Hayes & Londe's nonce probe study, accepting *–nɔk* 7% of the time overall (p. 72). The goals are as follows: we want the model to predict accurately the fixed pronunciation of each word; and we want the model to mimic frequency-matching behavior in nonce probe studies—that is to say, the model should select an irregular nonce form 100/(100+5000) ≈ 2% of the time.

We adopt Maximum Entropy Harmonic Grammar as the analytical framework. For illustrative purposes, we use three schematic constraints to satisfy our goals— BEREGULAR, and BELEXICAL(regulars) and BELEXICAL(irregulars) —with the definitions given below. Note that the usage of BELEXICAL below requires us to assume that SRs are listed in the lexicon together with URs.

(22a) BEREGULAR: assess 1 violation to any irregular form in the language.

(22b) BELEXICAL(regulars): for each member (x_i, y_i) in a set of *n* attested UR-SR pairs labeled as regular in the language, assess 1 violation to any form z_i such that $z_i \neq y_i$.

(22c) BELEXICAL(irregulars): for each member (x_i, y_i) in a set of *n* attested UR-SR pairs labeled as irregular in the language, assess 1 violation to any form z_i such that $z_i \neq y_i$.

BEREGULAR is a general, grammatical constraint that regulates the overall rate of regularity across words. The BELEXICAL constraints govern the behavior of individual UR-SR pairs. The violation profiles are as in the table below. The weights of the three constraints are initialized to be 0.

		BeRegular 0	BeLex(reg) 0	BeLex(irreg) 0
Regular forms	Correct: 5000			
	Incorrect: 0	-1	-1	
Irregular forms	Correct: 100	-1		
	Incorrect: 0			-1

Table 36: *MaxEnt strict exceptionality toy input*

One might expect the learner to eventually arrive at constraint weights that yield the following results: the model, when presented with nonce forms that must be classified as regular or irregular, would favor the irregular form roughly 2% of the time; and the model would get actual words pronunciations correct 100% of the time. The learner ideally would arrive at a high weight for the BeLexical constraints, so that she gets the right the fixed pronunciation for each word. Moreover, the learner ideally would arrive at a weight for BeRegular that frequency matches the nonce irregularity rate of the lexicon—roughly 2%—when confronted with a set of nonce forms. For example, if the learner converges at the set of weights *w*BeRegular = 4 and *w*BeLexical = 10 (for both regulars and irregulars), then it achieves the roughly desired rates: regulars are acquired as regular 99% of the time, and irregulars are acquired as irregular 99% of the time; moreover, the weight of BeRegular would result in the learner selecting a nonce irregular at around a 1.8% rate.

Let us look at a learning simulation. Recall that learning in MaxEnt proceeds as follows. Pick the constraint weights that maximize the log-probability of the data set, minus a penalty term to avoid overfitting. Suppose a dataset $\{(x_i, y_i)\}_{i=1}^{n}$ of *n* UR-SR pairs (x_i, y_i). The formula is given below in Figure 1. The smaller the σ, the less w will move away from its expected value (typically set to 0) to fit to input rates.

(23)

$$\underbrace{\sum_{i=1}^{m} \log\big(P(y_i|x_i)\big)}_{\text{log probability of the dataset}} \quad - \quad \underbrace{\sum_{j=1}^{n} \frac{\left(w_j - \mu_j\right)^2}{2\sigma_j^2}}_{\text{penalty term}}$$

Microsoft Excel's Solver (Fylstra et al. 1998; *cf.* Walsh & Diamond 1995, Harris 1998), which recruits Newton's method (Tay, Kek & Abdul-Kahar 2009) to find optimal parameter values for non-linear models (i.e., parameter values that occur at the maximum of the equation in (23)). The Solver was used simulate MaxEnt's learning of frequencies proportionate to those given in Table 36. We assess if the model is able to acquire both the 2% nonce irregularity rate, governed by of BEREGULAR, together with the behaviors of individual items, governed by the BELEXICAL constraints. We track learning by varying the frequencies of the various forms in the dataset, keeping constant the ratio of irregular forms to regular forms —that is, 1 irregular form for every 50 regular forms. In other words, we vary number of times the learner is re-presented with the dataset. In an earlier trial, for example, we multiply the frequencies in the dataset by 1, resulting in a dataset with 50 regulars and 1 irregular form; Solver then finds optimal weights for the constraints given that dataset, starting from weights initialized at 0. In one of the later trials, we would multiply the frequencies in the dataset by 100, resulting in a dataset with 5000 regulars and 100 irregular forms; Solver then finds optimal weights for the constraints given that dataset, starting from weights initialized at 0. In an even later trial, we multiply the frequencies in the dataset by 10,000, resulting in a dataset with 500,000 regulars and 10,000 irregular forms; Solver again finds optimal weights for the constraints given that dataset, starting from weights initialized at 0. (As will be shown below, running a learning simulation in this way yields the same learning outcomes as holding frequencies constant but varying σ in the penalty term.)

We give a learning simulation with the MaxEnt penalty setting μ = 0, σ = 1,000 for each constraint. We begin with very small frequency multipliers in the childhood phase—the results of which are given in the table below. She learns regulars quickly—getting 98% of them correct after the first trial of learning—but she learns the irregulars less rapidly, only getting 53% of them right. By the time she reaches the 0.01 multiplier, she already acquired most of the attested lexicon, classifying the vast majority of regulars and irregulars correctly—this is not surprising, considering that the lexical constraints are completely undominated, preferring only winners. Such is not the case for the general constraint, and as such the learner selects nonce irregulars at a rate of roughly 19% around this point in learning rather than the desired 2%.

Freq. multiplier	Be Reg	BeLex (regs)	BeLex (irregs)	Regular correct	Irreg. correct	Nonce irreg. rate
0	0	0	0	0.5000	0.5000	0.5000
0.00001	1.64	2.79	1.76	0.9883	0.5319	0.1622
0.0001	1.43	4.75	3.71	0.9979	0.9071	0.1917
0.001	1.39	6.73	5.66	0.9997	0.9861	0.1986
0.01	1.35	8.80	7.64	0.9999	0.9981	0.2049

Table 37a: *learning simulation in MaxEnt: "childhood" phase*

At adolescence, we begin observe to observe overfitting, with the learner vacillating in the weight of BEREGULAR and the nonce irregularity rate. She learned the lexicon nearly perfectly at this point, before BEREGULAR could be weighted high enough to result in a nonce irregularity rate of around 2%—with no period of frequency matching to the lexical trend. BELEXICAL soars in weight and takes over all of the explanatory opportunity in the dataset, while BEREGULAR vacillates, taking on a spectrum of weight between 0 and 3.5 throughout these learning trials.

Freq. multiplier	Be Reg	BeLex (regs)	BeLex (irregs)	Regular correct	Irreg. correct	**Nonce irreg. rate**
0.1	3.32	8.83	11.55	0.9999	0.9997	**0.0349**
1	0.73	13.41	11.64	0.9999	0.9999	**0.3256**
10	3.45	11.09	15.87	0.9999	0.9999	**0.0306**
100	2.35	16.24	15.15	0.9999	0.9999	**0.0868**
1000	0.00	22.31	17.16	1	0.9999	**0.5000**
10000	1.89	20.57	67.91	1	1	**0.1311**

Table 37b: *learning simulation in MaxEnt: "adolescent" phase*

At adulthood, the learner acquires the attested regulars and irregulars *with essentially perfect accuracy*, at which point the weight of BEREGULAR plummets to and remains at zero, such that she begins selecting nonce irregulars at a fifty-fifty rate. BELEXICAL constraints continue to soar into the firmament, while the grammatical constraint BEREGULAR remains lifeless at 0, in perpetuity. This is visible in the table below.

Freq. multiplier	Be Reg	BeLex (regs)	BeLex (irregs)	Regular correct	Irreg. correct	**Nonce irreg. rate**
100000	0.00	34.96	21.72	1	1	**0.5000**
1000000	0.00	27.95	26.83	1	1	**0.5000**
10000000	0.00	27.98	59.57	1	1	**0.5000**

Table 37c: *learning simulation in MaxEnt: "adulthood" phase*

Even before the learner comes to vacillate and eventually "forget" her grammar, the 19% rate that the learner stalls at in the childhood phases is not particularly close to the desired 2% nonce irregularity rate. The BELEXICAL constraints exert too powerful an effect, being undominated in the input, and come to explain the entire data set before BEREGULAR arrives at a weight that results in the learner frequency matching to the lexical trend. If we take the weights of the BELEXICAL constraints found when we take a frequency multiplier of 100,000—that is to say, 34.96 and 21.72, respectively—and set the weight of BEREGULAR to 0, we still get a near perfect match to the observed data. BEREGULAR is therefore explaining *none* of the dataset by this point in the development of learning, and is entirely redundant.

The complete learning simulation—showing the learner's success in learning the entire dataset via BELEXICAL constraints, and its failure to frequency match to the lexical trend as a whole by BEREGULAR—is summarized in the table and figure below.

Developmental phase	Freq. multiplier	Be Reg	BeLex (regs)	BeLex (irregs)	Regular correct	Irreg. correct	**Nonce irreg. rate**
"Childhood"	0	0	0	0	0.5000	0.5000	0.5000
	0.00001	1.64	2.79	1.76	0.9883	0.5319	0.1622
	0.0001	1.43	4.75	3.71	0.9979	0.9071	0.1917
	0.001	1.39	6.73	5.66	0.9997	0.9861	0.1986
	0.01	1.35	8.80	7.64	0.9999	0.9981	0.2049
"Adolescence"	0.1	3.32	8.83	11.55	0.9999	0.9997	**0.0349**
	1	0.73	13.41	11.64	0.9999	0.9999	**0.3256**
	10	3.45	11.09	15.87	0.9999	0.9999	**0.0306**
	100	2.35	16.24	15.15	0.9999	0.9999	**0.0868**
	1000	0.00	22.31	17.16	1	0.9999	**0.5000**
	10000	1.89	20.57	67.91	1	1	**0.1311**
"Adulthood"	100000	0.00	34.96	21.72	1	1	**0.5000**
	1000000	0.00	27.95	26.83	1	1	**0.5000**
	10000000	0.00	27.98	59.57	1	1	**0.5000**

Table 37d: *full learning simulation in MaxEnt*

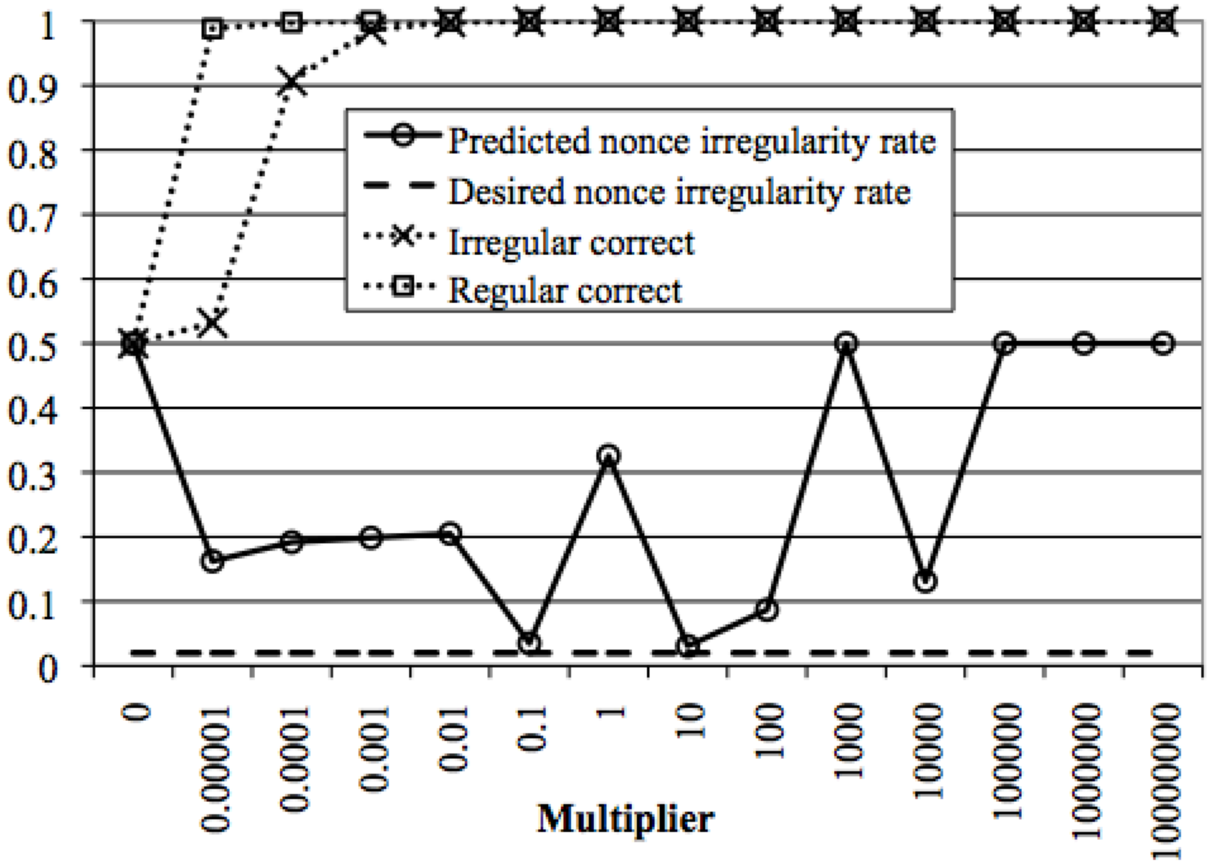


Figure 21: *ineffectiveness of the broader grammar to frequency match to statistical generalization, due to overfitting of lexical constraints*

Given below is a graph of the weights of BEREGULAR and BELEXICAL constraints across learning trials. The weight of the BELEXICAL constraints soar, while BEREGULAR reaches around 1.4 before it begins to vacillate. Once the weights of the BELEXICAL constraints reach a high enough value, the weight of BEREGULAR drops to and remains at zero.

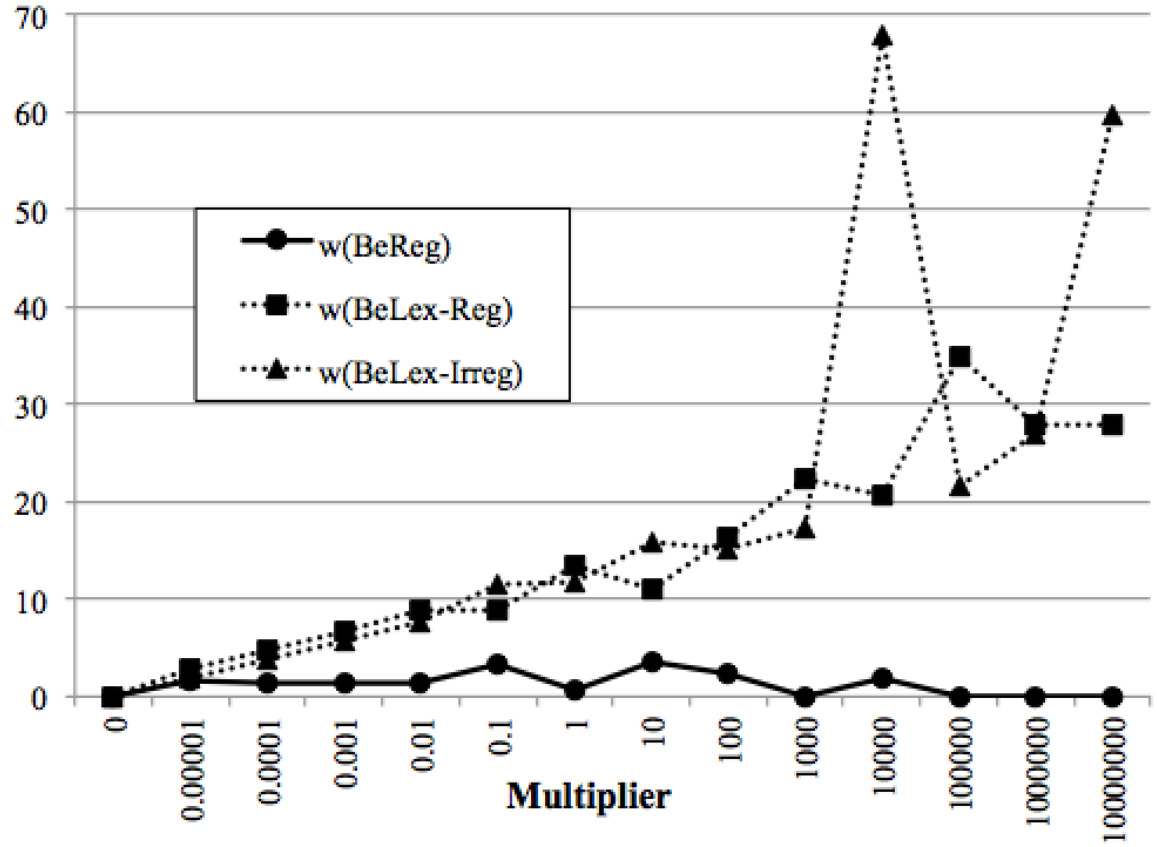


Figure 22: *soaring weights for* BELEXICAL *constraints, stalling and zeroing out of weights for* BEREGULAR

In fact, both the weights of BEREGULAR and the two BELEXICAL constraints come to vacillate throughout the learning process given above. This is because the BELEXICAL constraints are undominated—coming to fit to the data with near-perfect accuracy—and so their ideal weight is infinite. As such, many values for their weight work well to provide the near-perfect fit.

The model fails to learn an adequate, frequency-matching weight for BEREGULAR for the very reason that the BELEXICAL constraints alone are enough to fit to the data perfectly. BEREGULAR is imperfect for the purposes of replicating the training data, as it is violated by the irregulars in the dataset, whereas the two BELEXICAL constraints are enough to replicate the data nearly perfectly. BEREGULAR is thus treated as a superfluous constraint for learning the dataset.

The overfitting result is general across values of σ. According to (23), multiplying σ by a factor of k (e.g., $k = 10$) has the same effect on learning as dividing the frequencies of the dataset by a factor of k^2 ($k^2 = 100$). Likewise, dividing σ by k ($k = 10$) has the same effect on learning as multiplying the frequencies by k^2 ($k^2 = 100$). This is evident in the table below, which presents the results of a series of learning simulations of the dataset from above (but only fitting the weight of BeRegular to it). For example, setting σ = 1 and frequency multiplier $m = 100$ yields the same learning outcome as setting σ = 10 and $m = 1$.

	σ = 1		**σ = 10**		**σ = 100**	
	irreg. rate	weight	irreg. rate	weight	irreg. rate	weight
***m* = 0.01**	0.4748	0.1008	0.1127	2.0629	0.0213	3.8258
***m* = 1**	0.1127	2.0629	0.0213	3.8258		
***m* = 100**	0.0213	3.8258				

Table 38: *identical learning outcomes across different values of m and* σ

Hence decreasing σ across constraints merely extends the time at which the learner begins to stall and forget her grammar—that is to say, decreasing σ merely has the effect of delaying the stages of learner overfitting.[8]

One might imagine that setting a high value of σ for the general constraint and a low value of σ for the lexical constraints would solve the overfitting problem. My investigation into this approach has yielded negative results: setting σ = 1,000 for BeRegular and σ = 10 for the lexical constraints, for example, still yields poor frequency-matching predictions for the nonce irregularity rate, especially around the period where the learner achieves a near-perfect fit to the lexicon:

[8] Manipulating values of μ in the penalty also has no effect in surmounting the overfitting problem: in trials with positive frequency multiplier, the learner simply adjusts the constraint weights to the values found above, regardless of whether they start at 0.

Freq. multiplier	Be Reg	BeLex (reg)	BeLex (irreg)	Regular correct	Irreg. correct	Nonce irreg. rate
1	2.00	4.23	4.23	0.9980	0.9025	**0.118543281**
10	1.96	6.19	6.19	0.9997	0.9857	**0.123141569**
100	1.95	8.22	8.22	0.9999	0.9981	**0.123982297**
1000	1.90	10.38	10.26	**0.9999**	**0.9999**	**0.353377945**
10000	0.38	13.87	10.85	**0.9999**	**0.9999**	**0.404501717**
100000	0.27	16.13	12.76	**0.9999**	**0.9999**	**0.431732639**
1000000	0.41	18.58	15.23	**0.9999**	**0.9999**	**0.397693575**

Table 39: *learning simulation in MaxEnt,* σ(BeRegular) = 1,000*,* σ(lexical constraints) = 10

Even with this σ-based bias towards general constraints, the weights of the two lexical constraints still soar, while grammatical weight vacillates and remains too low across the learning simulation to be effective.

6.2.2 MaxEnt fails to learn statistical generalizations with lexical propensities

One might ask whether the model would fare better with a different dataset—for example, one with lexical propensities. Suppose we have twelve words in the dataset with equal numbers of tokens, each with a different propensity across tokens to undergo some phonological process. Such a dataset is reminiscent of those observed in variable phonology in which words undergo a variable process at different rates, including cases discussed in prior sections, and in other works (Hayes & Londe 2006; Zuraw 2009, 2016; Smith & Moore-Cantwell 2017; Tanaka 2017). In the aforementioned MaxEnt-based approach to lexical variation, the general constraint should enforce language learners' ability to frequency match to lexical trends in nonce probe studies, and so its weight should eventually reach a value that results in frequency matching to the overall average rate across the twelve words. The weights of the lexically specific constraints should, in tandem with the weight of the general constraint, govern the individual behaviors of each word in the dataset, matching their rates.

Consider the dataset below consisting of twelve words, with propensities distributing across a spectrum. We pick these rates in particular so that the overall average rate across all forms, roughly **61%**, is not close to any one of the word-specific rates—hence there would be reason to allocate positive weight to a lexical constraint for each word, rather than adjusting the general constraint weight to fit to any one of the words with a propensity that happens to match to the overall rate.

Word	Rate	Word	Rate	Word	Rate
1	0.00	5	0.30	9	1.00
2	0.00	6	0.80	10	1.00
3	0.10	7	0.90	11	1.00
4	0.20	8	1.00	12	1.00

Average over all rates: 0.61

Table 40: *dataset consisting of twelve words with differing propensities*

For any Word*i*, the predicted probability of the candidate Word*i* surfacing is one minus the probability that Word*i*-alt surfaces. Hence we get the following input to be fed into Solver after assigning each word's frequency their respective probability in Table 40 and resetting the weights to 0 (and thus the predicted probabilities back to 0.5). We use the general constraint APPLY, whose weight should result in frequency matching to the 60% rate of application across the dataset, and FAITH$_1$, FAITH$_2$, ..., FAITH$_5$, APPLY$_6$, APPLY$_7$, ..., APPLY$_{12}$ as the twelve lexical constraints, which in tandem with APPLY should result in frequency matching to word-specific rates. Note that we use FAITH$_1$, FAITH$_2$, ..., FAITH$_5$ to enforce lower-than-chance rates of application in the first five words, and APPLY$_6$, APPLY$_7$, ..., APPLY$_{12}$ to enforce higher-than-chance rates in the last seven words.

UR	SR	Freq.	Pred. P	H	APPLY 0	FAITH$_1$ 0	...	FAITH$_5$ 0	APPLY$_6$ 0	...	APPLY$_{12}$ 0
Word1	Word1-alt'n	0.00	0.5	0		1					
	Word1-faith	1.00	0.5	0	1						
...	...	...	...	...	...	...	...		...	...	
Word5	Word5-alt'n	0.30	0.5	0				1			
	Word5-faith	0.70	0.5	0	1						
Word6	Word6-alt'n	0.80	0.5	0							
	Word6-faith	0.20	0.5	0	1				1		
...	...	...	...	...	...	...	...		...	...	
Word12	Word12-alt'n	1.00	0.5	0							
	Word12-faith	0.00	0.5	0	1						1
Nonce	Nonce-alt'n	0.00	0.5	0							
	Nonce-faith	0.00	0.5	0	1						

Table 41: *propensity toy dataset input to Excel Solver*

Notice that the only candidate pair assigned zero frequency are the nonce candidates. Since Nonce is a nonce form, its candidates receive zero frequencies; rather, its frequencies are predicted by the weight of APPLY alone—which should frequency match to the overall average rate of 61%—after the learner has found optimal weights for all thirteen constraints. We can say that this learning model succeeds if it learns weights for the general constraint and twelve lexically indexed constraints such that: 1) the predicted probabilities for the alternated forms of each attested word match the word-specific alternation rates; and 2) the predicted probability of the alternated nonce form—the nonce regularity rate—frequency matches to the average over all of the rates of alternated words in the data set—the desired, overall regularity rate—i.e., 0.61. The following set of weights results in frequency matching to both word-specific rates and the lexical trend overall:

Constraint weight should match:	Constraint	Weight	Observed rate	Predicted rate
Overall rate	**APPLY**	**0.40**	**0.61**	**0.61**
Word1 rate	FAITH$_1$	10.00	0.00	0.00
Word2 rate	FAITH$_2$	10.00	0.00	0.00
Word3 rate	FAITH$_3$	2.65	0.10	0.10
Word4 rate	FAITH$_4$	1.85	0.20	0.20
Word5 rate	FAITH$_5$	1.28	0.30	0.30
Word6 rate	APPLY$_6$	0.95	0.80	0.80
Word7 rate	APPLY$_7$	1.75	0.90	0.90
Word8 rate	APPLY$_8$	10.00	1.00	1.00
Word9 rate	APPLY$_9$	10.00	1.00	1.00
Word10 rate	APPLY$_{10}$	10.00	1.00	1.00
Word11 rate	APPLY$_{11}$	10.00	1.00	1.00
Word12 rate	APPLY$_{12}$	10.00	1.00	1.00

Table 42: *set of successful weights for propensity dataset*

We set forth with learning simulations. We begin with a learning simulation under which the learner takes the settings μ = 0 and σ = 1,000 for the penalty term. As we did before, we multiply each of the word frequencies by a small factor, learn the weights, and record the results; then, we start over with weights set to 0, multiply each of the word frequencies by a larger factor, re-learn the weights, and record the results; and so on.

The results of the full learning simulation are given in the table and graph below, which include information on the predicted nonce regularity rate along with the predicted rates for Word5, Word6, and Word12. Though the predicted rates for the twelve words accurately match the desired rates, the learner never experiences a sustained period of frequency matching to the overall rate. The learner's predicted nonce regularity rate vacillates primarily between 0.50 and 0.80—in the latter case because it occasionally recruits APPLY to fit to the rate of Word6, whose rate is closest to the overall average rate relative to the other words in the dataset. Only for one setting of the frequency multiplier—100—does the learner achieve a roughly frequency matching rate. And past a multiplier of 10,000, the weight of APPLY drops to and remains at zero. Once again, the entirely undominated lexical constraints are so powerful that they come to rapidly explain the entire dataset, such that the general constraint APPLY is ineffective as a device for frequency matching nonce rate to the overall rate in the dataset.

Freq. multiplier	APPLY	FAITH$_5$	APPLY$_6$	APPLY$_{12}$	**Pred. nonce rate**	Pred. Word5 rate	Pred. Word6 rate	Pred. Word12 rate
0	0.00	0.00	0.00	0.00	**0.50**	0.50	0.50	0.50
0.001	0.00	0.84	1.39	11.19	**0.50**	0.30	0.80	1.00
0.01	0.00	0.85	1.39	10.64	**0.50**	0.30	0.80	1.00
0.1	1.39	2.23	0.00	8.77	**0.80**	0.30	0.80	1.00
1	1.40	2.25	0.00	11.85	**0.80**	0.30	0.80	1.00
10	1.37	2.22	0.02	16.17	**0.80**	0.30	0.80	1.00
100	0.50	1.34	0.89	8.25	**0.62**	0.30	0.80	1.00
1000	1.41	2.27	0.00	6.16	**0.80**	0.30	0.80	1.00
10000	0.00	0.85	1.39	12.60	**0.50**	0.30	0.80	1.00
100000	0.00	0.85	1.39	11.84	**0.50**	0.30	0.80	1.00
1000000	0.00	0.85	1.39	11.84	**0.50**	0.30	0.80	1.00

Table 43: *propensity learning simulation in MaxEnt*

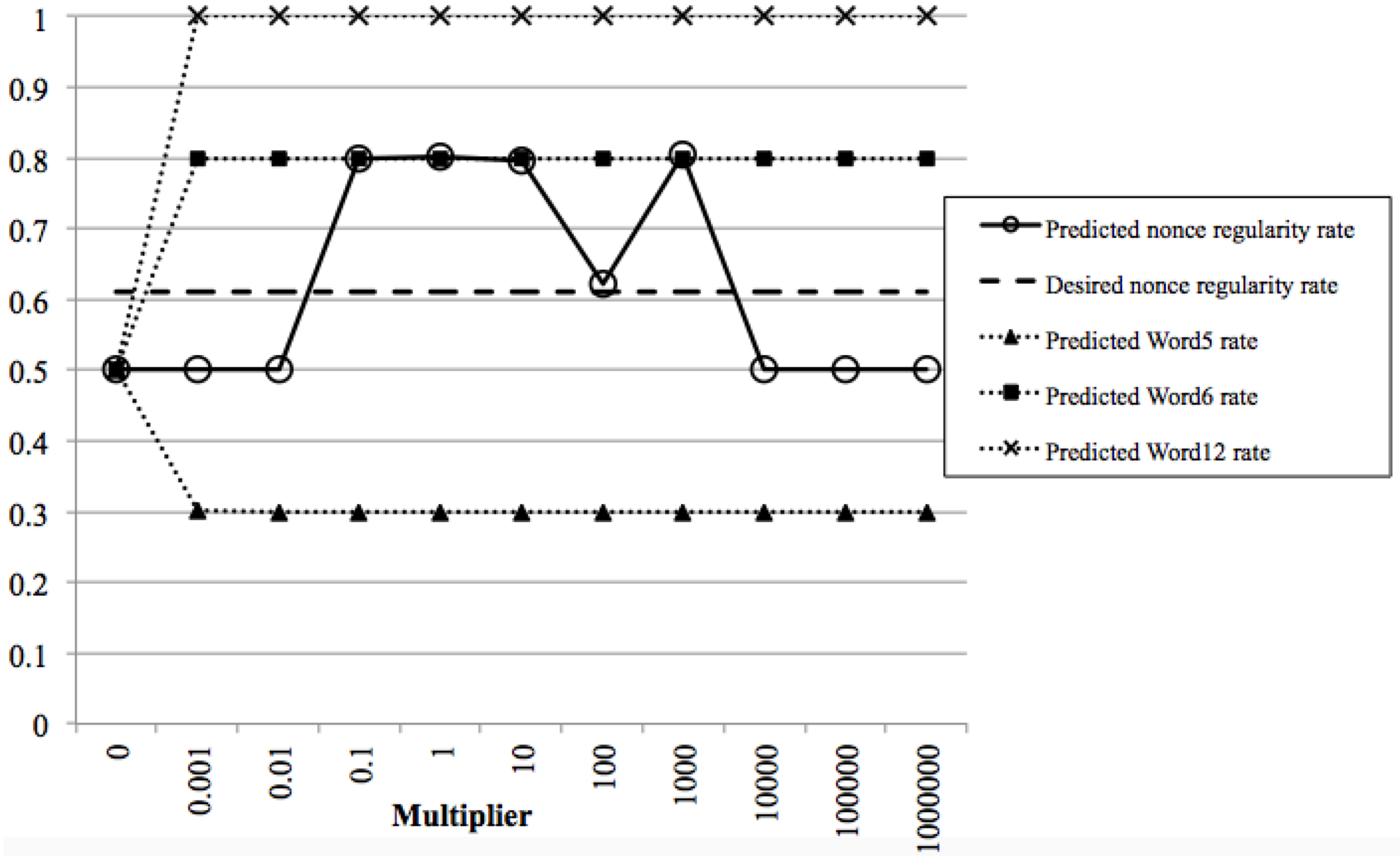


Figure 23: *correct learning of word-specific rates, incorrect learning of overall application rate*

The model fails. Very reasonable assumptions have led to a GRAMMAR-LEXICON BALANCING PROBLEM: in MaxEnt Harmonic Grammar—essentially a canonical logistic regression model—given that the learner makes use of both general and lexically indexed constraints, the learner merely acquires the lexicon, but not the grammar. Under our simulations with the first dataset, the weight of BEREGULAR:

- begins to stall and vacillate around the incorrect nonce irregularity rate once the weights of the BELEXICAL constraints grow high enough such that the learner acquires near-perfect knowledge of the dataset;

- eventually zeroes out, when the weights of the BELEXICAL constraints grow high enough such that the learner acquires perfect knowledge of the dataset.

In the second dataset with word-specific propensities, the lexical constraints quickly frequency match to the desired word-specific rates, while the general constraint is again rendered ineffective throughout the learning process. We want statistical models of variable phonology to frequency match rates across the lexicon. Moreover, we want the learner's knowledge of the grammar that she acquired from the distribution of the input to *remain* in the adult system, even after mastering the lexicon. Our learning theory should not predict that at some age, the learner would suddenly come to fail nonce probe studies into general statistical knowledge of a variable phonological system. Rather, it is natural and desirable to assume that the learned adult will pass such a test, no matter their age (see Shademan 2007 for results suggesting that healthy elderly speakers, though they display greater analogical effects relative to younger speakers, nevertheless still display grammatical generalizations in experiments). We must respect our elders, and thus we seek to identify a solution to the grammar-lexicon balancing problem.

6.3 Mixed-effects logistic regression as a model of the grammar and lexicon

Let us return to the toy dataset on word propensities that we were working with in the prior section.

Word	Rate	Word	Rate	Word	Rate
1	0.00	5	0.30	9	1.00
2	0.00	6	0.80	10	1.00
3	0.10	7	0.90	11	1.00
4	0.20	8	1.00	12	1.00

Average over all rates: 0.61

Table 44: *dataset consisting of twelve words with differing propensities*

Assuming there are two candidates for /Word-*i*/ in this dataset—the faithful candidate and the alternated candidate—the MaxEnt model is essentially a simple BINOMIAL LOGISTIC REGRESSION MODEL. In this model, the general constraint APPLY and the set of lexical constraints FAITH_1, FAITH_2, ..., FAITH_5, APPLY_6, ..., APPLY_{12} are treated as *equally viable* explanatory variables for learning the dataset and its patterns. As such, the general constraint and the set of lexical constraints compete to explain the dataset, with lexical constraints—which are completely undominated—coming to explain its entirety, such that the general constraint fails to mimic the frequency matching behavior of language learners in nonce probe experiments.

We therefore seek a model that possesses a strong GENERALITY BIAS: one in which general, grammatical constraints are privileged to lexical constraints. Such a model would adopt lexical constraints to the extent that they can explain lexical idiosyncrasies in the data, but in such a way that they would not constitute the only explanation of the dataset. That is to say, the model would be posed such that the dataset would be *primarily explained by the general constraint*, with deviations from the trend obtained by the general constraint being explained by *subsidiary, lexical constraints*.

A similar model to the logistic regression model is the MIXED-EFFECTS LOGISTIC REGRESSION MODEL. We might call an analytical framework based on this model MIXED-EFFECTS MAXENT HARMONIC GRAMMAR. For our purposes, the relevant difference between a simple logistic regression model and a mixed-effects logistic regression model is that the latter model posits a distinction between FIXED EFFECTS and RANDOM EFFECTS. Constraints that constitute fixed effects are those that concern the population of words as a whole—that is to say, we would not have to change our constraints if we substituted out the dataset at hand for another one; i.e., general, grammatical constraints. Constraints constituting a random effect are those that capture the idiosyncrasies in the dataset itself, such as lexically indexed constraints, without threatening the general constraint to capture the broader trend. Mixed effects logistic regression models, as a subset of mixed models (linear, logistic, Poisson, etc.) are used widely across scientific fields, with fixed effects being used to capture broad trends across a dataset, and with random effects used to account for idiosyncrasies particular to the dataset at hand. Moreover, the mixed model has been argued to carry various advantages relative to other models in capturing various insights within the field of linguistics (Baayen 2004, 2008; Baayen et al. 2008; Jaeger 2008; Quené & van den Bergh 2008; Levy 2010; Fruehwald 2012; Shih & Inkelas 2016; Zuraw & Hayes 2017; Smith & Moore-Cantwell 2017; Shih 2018). A growing family of research has recruited the random intercept to measure the degree of by-word or by-lexical class idiosyncrasy in datasets displaying morphophonological variation (Fruehwald 2012, Shih & Inkelas 2016, Zuraw & Hayes 2017, Smith & Moore-Cantwell 2017, Shih 2018); Shih & Inkelas (2016) and Shih (2018) even adopts the hierarchical mixed model as a theory of the language learner.

MaxEnt is equivalent to the simple logistic regression model (Smolensky 1986, Goldwater & Johnson 2003, Smolensky & Legendre 2006) when the candidate set is restricted to two outcomes. Here we restrict attention to the binomial version of MaxEnt—those which model the conditioning of the relative probabilities of two candidates: a candidate *x* in which some phonological process has applied, and a faithful candidate *y*. The equation in (24a) is *x*'s *harmony* (Hayes & Wilson 2008), and the probability that *x* surfaces, $P(x)$, is as in (24b) below. Here we denote any constraint *k* over the set of *K* constraints as C_k, and its weight as w_k.

(24a)

$$H(x) = \sum_{k=1}^{K} w_k * C_k(x)$$

(24b)

$$P(x) = \frac{e^{-h(x)}}{e^{-h(x)} + e^{-h(y)}}$$

$P(x)$ is the probability of candidate x, and, since this is a binomial regression, $P(y) = 1 - P(x)$. The objective function to be maximized by finding optimal weights is the probability of the observed dataset, i.e., the product of all probabilities $P(x)$ of every candidate x in the input.

In the lexical indexation approach to lexical variation, a lexically indexed constraint is treated as any other constraint in CON, its weight estimated in the same way that any general, grammatical constraint's weight is estimated. In mixed effects logistic regression models, which feature the distinction between fixed effects and random effects, we code general, grammatical constraints as fixed effects—constraints that target populations (in our case, any morpheme that contains a particular phonological configuration)—and the weights of lexical constraints as coefficients of levels of a random intercept—constraints that target particular data in the sample dataset at hand; i.e., constraints over particular morphemes. Here we regard CON as the union between two kinds of constraints: *K* general, grammatical constraints which constitute main effects, and *L* lexically indexed constraints which constitute a random intercept. We denote any fixed effect constraint *k* as C_k and its weight as w_k, and we denote any constraint that is part of a random intercept, *l*, as U_l, and its weight as v_l. The harmony of candidate *x* surfacing is as in (25).

(25)

$$H(x) = \sum_{k=1}^{K} w_k * C_k(x) + \sum_{l=1}^{L} v_l * U_l\,(x)$$

The constraints U_l might be lexically indexed constraints (Pater 2000, Moore-Cantwell & Pater 2016, *inter alia*), UR-constraints (Pater, Staubs, Jesney & Smith 2012, Smith 2015), or USE constraints (Zuraw & Hayes 2017), etc. Notice that the formula in (25) above is nearly the same as that in (24a), except now we are distinguishing between constraints regarded as fixed effects and those regarded as part of a random intercept.

The mixed effects logistic regression model is hierarchical, in that random effects are "depreciated" relative to fixed effects in explaining the dataset. What does this mean? In the dataset in Table 44, we have twelve Words—twelve *groups of tokens*. The central distinction between constraints that constitute fixed effects and constraints that constitute random effects is that their weights are determined differently. If we were to posit a factor targeting a particular group in simple logistic regression—for example, a constraint targeting a particular word, APPLY–Word-i—its coefficient (i.e., weight) would be estimated by the group mean, or the application rate across Word2 tokens—call this μ_{Word-i}. Without lexical constraints, a general constraint such as APPLY would apply to the entire population of words, and thus its coefficient would be estimated by the population mean—call this $\mu_{all\,words}$. But in mixed models, the coefficient of a level of a random intercept—the weight of APPLY–Word-i, for example—is determined by combining both group information and population information. Whereas in a simple logistic regression the coefficient is a free parameter (see below), here the coefficient is not a free parameter, but rather is determined by a weighted average of the mean of Word-i (rate across Word-i tokens) and mean over all words (Snijders & Bosker 2012, p. 62-3):

(26a) $\lambda_{Word-i} * \mu_{Word-i} + (1 - \lambda_{Word-i}) * \mu_{all\,words}$

The expression in (26a) is called the empirical Bayes estimate, which produces a so-called posterior mean (see Efron & Morris 1975, Gelman 2004, Snijders & Bosker 2012). λ_{Word-i}, called the reliability of the mean of Word2, is determined as follows (Snijders & Bosker 2012, p. 62-3, 291):

(26b)

$$\frac{\tau_0^2}{\tau_0^2 + \frac{\sigma^2}{n_{Word-i}}}$$

In the above, τ_0^2 is the variance of the random effect for Word, and σ^2 is the residual variance. The reliability takes on a value between 0 and 1, and depends on the number of observations in a particular group—e.g., the number of tokens of Word2, n_{Word-i}—with more observations resulting in values closer to 1, and fewer observations resulting in values closer to 0. That the reliability is sensitive to the size of the group is of particular importance. Since we deal with groups of tokens that are finite, n_{Word-i} in (26b) is necessarily finite, and so λ_{Word-i} is less than 1. Hence the population mean at least in part determines the posterior mean for Word-*i*, and as a result, the Word-*i* mean is always slightly *shrunken* toward the population mean (see Snijders & Bosker 2012, Raudenbush & Bryk 2012 for further discussion of shrinkage). In our dataset, Word-*i* has a different application rate than the overall average; if Word-*i* were to have a small number of tokens, then the weight of APPLY–Word-*i* would be more shrunken towards the weight of APPLY than if Word-*i* were to have lots of tokens. The property that a lexical weight departs from general weight more when it targets a larger group of data, and less when it targets a smaller group of data, predicts a pattern of variation observed in Morgan & Levy (2016) and Smith & Moore-Cantwell (2017) whereby more idiosyncrasy is observed within frequent forms, but more grammatical behavior within infrequent forms.

The parameters of a mixed model are the coefficients of the fixed effects—the weights of our non-lexical constraints—and the variances of the random effects and residual variance, as in τ_0^2 and σ^2 (Snijders & Bosker 2012, Raudenbush & Bryk 2012). The objective function of a mixed logistic regression model—the probability of the observed dataset given the fixed effect coefficients and the random effect variances (leaving aside a regularization term)—does not possess a closed form, and so these parameters are estimated by maximizing a Laplace approximation to this likelihood. The approximation is maximized by applying the penalized iterative reweighted least squares algorithm, which performs batch gradient descent first on the fixed effect coefficients and then on the random variances—iteratively—to determine optimal values, until the relative change in predictors has fallen below a threshold value, at which point the iterates are said to have converged. Explanations of the details of the approximation and the learning algorithm lie outside the scope of this dissertation, but see Bates (2009), p. 28-31 for more details. R uses the Laplace approximation and penalized iterative reweighted least squares algorithm to determine the values of the coefficients for the main and random effects. See Bates (2009), Snijders & Bosker (2012), Raudenbush & Bryk (2012) (*inter alia*) for further discussion of the mathematical details behind how mixed-effects logistic regression models are implemented, and how simple logistic regression models and mixed-effects logistic regression models differ.

We seek to predict the following: 1) with the general constraint APPLY, the behavior of speakers in nonce probe studies to frequency match to the overall average rate of application across all Words in the dataset; 2) and with the set of lexical constraints, the attested rates for each Word. We run a model of the dataset in Table 44 using the `glmer` function of the *lme4* package R. Each Word had 10,000 tokens, assigned proportionally to their rate—for example, Word3 had 1,000 tokens of an alternated output, and 9,000 tokens of the faithful output. I inputted into R the following command in (27). We set `family = “binomial”` to indicate that the model is within the logistic family. Below, `application` is the dependent variable, `1` is the general intercept—that is, its coefficient corresponds to *w*APPLY—and `(1|word)` is the random intercept for word, with different levels of the intercept corresponding to the different weights of the lexical constraints.

(27)

```
model =   glmer(application ~ 1 + (1|word),
          data = toyPropensityData, family = “binomial”)
```

Note that while in the MaxEnt formulation we restricted coefficients to be positive (through using a combination of lexically indexed APPLY and FAITH constraints), here the coefficients of the level of the random intercept can also be negative.

From this, we obtain the modeling results in Table 45. The coefficients obtained for APPLY and the levels of the random intercept result in a good fit to the data. In particular, while the overall rate of application across the twelve words is 0.608, we find that the model predicts a 0.667 application rate to nonce words. We see that the weight of APPLY contributes to fitting accurate propensities, as it factors into how each of the lexical weights are fitted. On the other hand, we found that the weight of the general constraint contributed nothing to MaxEnt’s ability to fit to lexical rates in the strict exceptionality toy dataset. Here, with *w*APPLY set to zero, we observe a drastic decrease in the model’s accuracy in predicting word-specific rates, especially at medial rates. This is because the coefficients of the intercept are predominantly negative (except for Words 8 through 12)—we see that the positive coefficient for APPLY is being recruited in conjunction with the coefficients of the intercept to fit to these medial rates effectively.

Word	*Weight*	*Actual rate*	*Predicted rate*	*Predicted rate, w*APPLY = 0	*Accuracy difference*
Word1	-16.56	0.000	0.000	0.000	0.038
Word2	-16.56	0.000	0.000	0.000	0.096
Word3	-7.32	0.100	0.100	0.001	0.099
Word4	-6.51	0.200	0.200	0.001	0.199
Word5	-5.97	0.300	0.300	0.003	0.297
Word6	-3.74	0.800	0.800	0.023	0.777
Word7	-2.93	0.900	0.900	0.051	0.849
Word8	7.14	1.000	0.999	0.999	0.000
Word9	7.14	1.000	0.999	0.999	0.000
Word10	7.14	1.000	0.999	0.999	0.000
Word11	7.14	1.000	0.999	0.999	0.000
Word12	7.14	1.000	0.999	0.999	0.000

***w*APPLY = 5.130**

OVERALL AVERAGE APPLICATION RATE: 0.608
PREDICTED APPLICATION RATE TO NONCE WORDS: 0.667

Table 45: *output of the mixed-effects logistic regression model for the propensity dataset*

Note that the predicted nonce application rate is obtained differently in mixed models than it is in simple logistic regression. In MaxEnt, for example, we would simply take the inverse logit of *w*APPLY to be the predicted nonce application rate. But for the mixed-effects logistic regression model, we cannot do the same by simply "zeroing out" the weights of the lexical constraints in order to determine the nonce application rate, as it would result in nonce rate predictions that are non-frequency matching and highly exaggerated, tending towards the poles 0 and 1 (Pavlou et al. 2015).[9] The exact predicted nonce rate produced by the mixed-effects logistic regression model cannot be calculated analytically, and involves a complex integral over the random effect

[9] I note here that even Moore-Cantwell & Pater (2016) obtained exaggerated nonce rate predictions using the inverse logit of the weight of the general constraints, when they tested their MaxEnt model on Ernestus & Baayen (2013)'s dataset on Dutch voicing alternations. Though this is speculation, it could be that Pavlou et al. (2015)'s findings explain the exaggeration effect produced by their model too.

(Skrondal, Rabe-Hesketh 2009; Pavlou et al. 2015). To obtain an estimate of the model-predicted application rate to nonce words, I calculated Zeger et al. (1998)'s approximation to the model-predicted nonce rate, which roughly equates to taking the inverse logit of the value obtained by dividing the model-obtained value of *w*APPLY by the standard deviation of the random intercept, τ (see Pavlou et al. 2015 for justification for using this approximation to these ends). The actual equation for this approximation is given below—*c* is a constant equal to $\frac{16\sqrt{3}}{15\pi}$.

(28)

$$\frac{\exp\left(\frac{w\text{APPLY}}{\sqrt{c^2\tau^2+1}}\right)}{1+\exp\left(\frac{w\text{APPLY}}{\sqrt{c^2\tau^2+1}}\right)}$$

Thus inputting *w*APPLY = 5.130 by the standard deviation of the random intercept, τ = 12.38, yields the 0.667 rate. This rate closely matches the overall application rate across attested words, and thus the model mimics the frequency-matching behavior of language learners during nonce probe studies.

As we see in Table 45, this model: 1) predicts every word-specific rate, thereby learning the lexical idiosyncrasy displayed by the data; 2) predicts the nonce application rate that closely matches the overall application rate across attested data—mimicking participants in nonce probe studies, without going haywire at high levels of input. The grammar is sustained in learning without the weights of the general constraint beginning to vacillate together with general, grammatical rates once the learner achieves a high degree of fit to word-specific rates. The mixed-model outcome stands in stark contrast with that of the MaxEnt analysis given in the prior sections, whereby in no period of learning does the learner achieve a particularly close fit to the overall average via the weight of the grammatical constraint *w*APPLY, and in which the period of learning during which the learner achieves a close fit to the word-specific idiosyncrasies is characterized by vacillating grammatical rates, and—in the most extreme case where the learner acquires the lexicon perfectly—the zeroing out of grammatical weights such that the learner begins to select palatalization in eligible nonce probes at chance. Modeling the learning of a frequency-matching grammar together with lexical propensities was made possible by: switching from the simple logistic regression-based MaxEnt to the mixed-effects logistic regression model—Mixed-Effects MaxEnt; and realizing that we can obtain accurate, frequency-matching nonce regularity rates from the latter by using Zeger et al. (1998)'s approximation to Skrondal et al. (2009)'s method.

What about real data? We observed in Chapters 3, for example, that the model does indeed match to palatalization rates across the Slovenian lexicon. The learning model based in mixed-effects logistic regression acquires idiosyncratic application rates depending on suffix identity and stem identity, and yet is able to detect and track accurately trends across the lexicon. For example, the learner closely tracks that palatalization applies more readily to stem-final *k* than to *g*; moreover, it applies categorically where the faithful candidate would produce a geminate. Note that the predicted rates below were obtained using Zeger et al. (1998)'s approximation, as in (28).

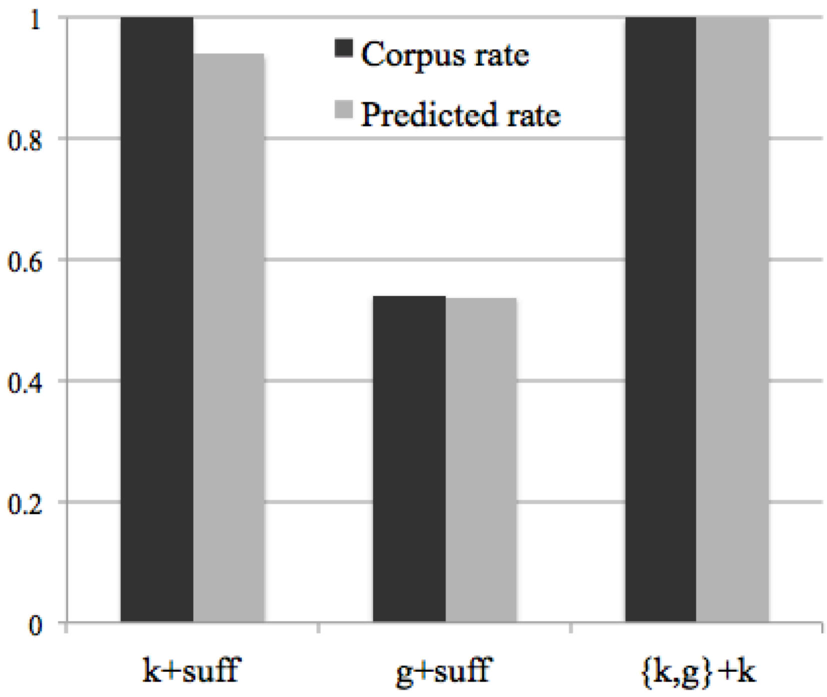


Figure 24: *model succeeds in predicting phonological trends in Slovenian*

I therefore submit **mixed-effects logistic regression** as a viable future approach to the modeling of lexical variation—to modeling the learning and representation of the lexicon and the grammar. That is to say, I propose Mixed-Effects MaxEnt Harmonic Grammar as a theory of the language learner. I note here that Mixed-Effects MaxEnt is not the only regression-based model currently on the market for adequately explaining speaker behavior when they learn datasets displaying lexical variation. Smith & Moore-Cantwell (2017) posit a MaxEnt-based model of lexical that uses UR constraints (Pater, Staubs, Jesney & Smith 2012, Smith 2015) to represent lexical idiosyncrasy. UR constraints are situated within an online error-driven learner in which learning data are sampled according to lexical frequency, and UR constraints are induced only when needed, and decay when they are not used—see Smith & Moore-Cantwell (2017) for further details. That these constraints are set to decay may be crucial in ensuring that a frequency-matching grammar is sustained throughout the learning process. Further research should be conducted to compare Smith & Moore-Cantwell (2017)'s model with the mixed-effects logistic regression model. But, thus far, the mixed model succeeds for my data, and has proven to be effective in modeling other variable datasets, both in morphophonology and other fields of linguistics (Baayen 2004, 2008; Baayen et al. 2008; Jaeger 2008; Quené & van den Bergh 2008; Levy 2010; Fruehwald 2012; Shih & Inkelas 2016; Zuraw & Hayes 2017; Smith & Moore-Cantwell 2017; Shih 2018).

Further research should be undertaken to assess more fully the capabilities of mixed-effects logistic regression models in capturing the behavior of language learners. The dissertation has introduced the binomial mixed-effects logistic regression model as a way to capture the learning of datasets displaying variation. Future work should flesh out the multinomial mixed-effects logistic regression model, to cover datasets in which underlying forms re associated with three or more surface forms. In addition, more work should be conducted to determine exactly what constraints must constitute fixed effects, versus a random effect. Invoking the hierarchical fixed-random distinction may be necessary for a variety of phenomena that require multiple levels of generalization—for example, cases where we need to invoke a general constraint together with constraints governing the behavior of different lexical classes (Shih & Inkelas 2016). Even further research should be conducted to assess how we can model generalizations at three or more levels: e.g., to cover a case where different lexical classes behave differently, and within each class, different morphemes behave differently. Lastly, more research should be conducted to assess the use of random slopes in capturing morphophonological variation (though some work showing their promise has already been put forth—see Shih & Inkelas 2016).

6.4 Summary

This section investigates how we can model the language learner who frequency matches to trends across the lexicon while also acquiring lexical propensities. A popular approach aims to capture trends with idiosyncrasy in Maximum Entropy Harmonic Grammar, using general, grammatical constraints to model frequency-matching behavior in nonce probe studies, together with lexical constraints to model lexical idiosyncrasies in the data. Though this approach is capable of representing lexical propensities, I have shown with a series of learning simulations that it encounters a GRAMMAR-LEXICON BALANCING PROBLEM: lexical constraints are so powerful in explaining the dataset that that the learner comes to acquire the behavior of each form using only these constraints, at which point the general constraint is rendered ineffective. I have argued that the choice to embed both grammatical and lexical constraints in Maximum Entropy Harmonic Grammar in particular, in which the grammatical constraint and lexical constraints are treated as equally viable explanatory devices for learning the dataset and its patterns, is what leads to the overfitting problem. The negative results obtained in the learning simulations, as well the findings that real language learners are nonetheless capable of generalizing across idiosyncratic variation, suggests that learners possess a generality bias, privileging general, grammatical constraints over the more granular lexical constraints when they acquire variable datasets. It is argued that MaxEnt in its current formulation—essentially a canonical logistic regression model—fails to appropriately represent this property, even after taking into consideration the MaxEnt penalty term. Privilege of general constraints to granular, lexical constraints can be represented in a hierarchical mixed-effects logistic regression model—Mixed-Effects MaxEnt—by encoding general constraints as fixed effects and lexical constraints as a random effect. The learner treats the grammar and lexicon differently upon positing the distinction between main effect and random effect, such that idiosyncratic effects of the vocabulary are subordinated to broad, grammatical effects in the learning process. The mixed model has been shown to succeed in learning both a frequency-matching grammar together with lexical propensities.

Chapter 7:

On morphosyntactic generality and specificity in phonology and phonological learning

A number of investigators have uncovered evidence for phonological learning biases: biases inherent in learners that favor certain language phonologies over others (Wilson 2006; Martin 2007, 2011; Finley 2012; Hayes & White 2013; White 2014; McMullin & Hansson 2014; Myers & Padgett 2014; Chong 2016, 2017; amongst many others; *cf*. Moreton & Pater 2012a, b). How strong, and how pervasive, are these biases? And to what extent can a learning bias be defied in language? Ultimately, in what form must these biases take in theories of language learning? These questions bear directly on the theory of phonological learning, as they address the limits of learner capability. The previous chapter found that models allowing general constraints to be pitted directly against lexically specific constraints overfits the latter to the dataset. It was proposed that learners must therefore be endowed with a generality bias, such that general constraints are privileged to lexically specific constraints for purposes of learning variable phonology—as in mixed-effects logistic regression, rather than canonical logistic regression. The model must privilege general principles to the extent that learners can extract broad generalizations from noisy data, but it also must be able to represent the numerous idiosyncrasies of the lexicon. The generality bias proposal, as well as the balance between representing generality and specificity in grammar, raises a number of questions. Beyond lexical idiosyncrasy, how does generality bias extend to how learners generalize out of whole morphosyntactic domains? Is generalization bias defiable, such that speakers learn a process specific to a single domain or affix, with no evidence of the process's working in other domains (e.g., phonotactics)? This chapter reviews prior findings relevant to generality and specificity at the phonology-morphosyntax interface, and raises a series of questions concerning how it should be modeled in the future.

7.1 The learner's tendency to extend phonological generalizations across morphosyntactic domains

A growing family of findings suggests that learners tend to favor phonological constraints that are morphosyntactically general—i.e., are obeyed by at least several morphemes, or in multiple or all grammatical contexts. That phonological alternations are typically corroborated by the phonotactic constraints of a given language was observed as early as Chomsky & Halle (1968) (Kenstowicz & Kisseberth 1977; McCarthy 2002; *et seq*), but the generalizing tendency just mentioned has also been observed in a number of recent corpus studies.

Martin's (2007) and (2011) studies, for example, find cases of grammatical "leaking", in which strong phonotactic restrictions tend to manifest as weaker statistical generalizations across compound boundaries. Martin focuses on two cases of grammatical leaking: Navajo sibilant harmony and English geminate avoidance. In Navajo, sibilants in a root must agree for anteriority. For example, the roots [t**ʃ**'o**ʒ**], 'worm' and [t**s**'o**z**i], 'slender' are attested in the language, but forms like *[**s**o**ʃ**] are forbidden. The harmony restriction can be observed in the cross-boundary domain as well. Prefixes in the Navajo undergo sibilant harmony as an alternation in order to match the anteriority of a sibilant in the root (e.g., /ji+**s**+lee**ʒ**/ → [ji-**ʃ**-lee**ʒ**]; Fountain 1998, Martin 2011). But across a compound boundary, sibilant harmony is not obligatory. Though many compounds abide by the restriction, a number of compounds do not:

(28) *Disagreeing sibilants across a compound boundary* (Young & Morgan 1987, Martin 2011)

[t**ʃ**ei#t**s**'i:n], heart#bone = 'ribcage'
[t**s**^{h}e#t**ʃ**e:ʔ], stone#resin = 'amber'

Martin investigates in a corpus whether compounds with disagreeing sibilants across a boundary are underattested in the lexicon. To determine whether the number of compounds disobeying the restriction is significantly above chance, Martin conducts a Monte Carlo test for significance (Kessler 2001). He indeed finds that the number of compounds that disobey the restriction is well below what chance alone would predict. Compounds in Navajo therefore obey the sibilant harmony restriction, though gradiently—the phonotactic restriction "leaks" into the domain of compounding. Martin shows that a similar leaking phenomenon arises in English compounds: though English bans geminates within roots, compounds such as *bookcase* and *bus stop*—wherein a geminate is formed at the boundary—are underattested. Zuraw (2015) and Shih & Zuraw (2018) further observe cases of grammatical "leaking", in which strong phonotactic restrictions tend to manifest across word boundaries, or affect the choice between grammatical constructions.

Martin attributes grammatical leaking to a generality bias, and in particular, to the penalty term adopted in the objective function of MaxEnt. Under small settings of σ, the following result is obtained: when the learner encounters the need for a structure-sensitive constraint (e.g., *s...ʃ/*morpheme-internal* in Navajo), it also posits a structure-sensitive constraint (*s...ʃ); since σ is low, the learner prefers small weights, and so it "spreads" some of the weight that would be obtained by *s...ʃ/*morpheme-internal* to the weight of *s...ʃ, thereby deriving the leaking effect (see Martin 2007, 2011 more in depth discussion). That this approach to generality bias suffices to model grammatical leaking but fails to model frequency matching to general lexical trends when lexically specific constraints are included in the model (see Section 6) raises an interesting question: what should the shape of generality bias take within the learning theory of the phonology-morphosyntax interface, so as to account for all of the generalization effects observed thus far? Mixed logits also contain a regularization term, much as MaxEnt does; but would encoding the general constraint as a fixed effect and a set of domain-specific constraints as levels of a random effect suffice?

Chong's (2016) and (2017) corpus investigation obtains a finding closely related to Martin's: a set of phonological phenomena previously claimed to be derived environment effects—morphophonological alternations that lack a corresponding phonotactic generalization in the lexicon—are merely apparent. In Korean palatalization, for example, /t, t^h/ palatalize—mapping to [c, c^h]—before suffixes beginning with high front vocoids, yet some roots fail to display the palatalization requirement morpheme-internally (Kiparsky 1973, 1993; Iverson & Wheeler 1988; T. Cho 2001). The Korean pattern was taken, based on these data, as evidence for a derived environment effect, and for the Derived Environment Condition—that is to say, morphological derivedness as a condition for a process to apply. However, Chong's corpus investigation reveals that roots that fail to display palatalization before high front vocoids are underattested within the Korean lexicon. In other words, Korean palatalization is merely an *apparent* derived environment effect, in fact displaying a strong trend towards palatalized sequences in roots, in addition to displaying palatalization as an alternation.

Generalization effects were also borne out in a set of artificial language learning experiments. Myers & Padgett (2014) found that participants generalize an utterance-final devoicing pattern to the word-final domain without exposure to unambiguous evidence.

The results are relevant here in that they suggest that language learners prefer general phonological principles—insensitive to morphosyntactic domain—showing a tendency to generalize a phonological principle from one domain into another. Chong (2017) features another experiment with results that support learner generalization tendency. In a series of artificial language learning studies consisting of a blick test (Scholes 1966 *et seq*) followed by a wug test (Berko 1958), he found that participants more readily learned an artificial suffixal harmony alternation in the wug test when they were exposed to higher rates of root harmony in the blick test. The experimental results support the proposal that phonotactic generalizations assist in acquiring alternations (Tesar & Prince 2003, Hayes 2004, Jarosz 2006, *a.o.*; *cf.* Chomsky & Halle 1968; Kenstowicz & Kisseberth 1977, 1979; McCarthy 2002; *et seq*). Overall, the results of the Myers & Padgett (2014)'s artificial language learning experiment—wherein learners generalize an utterance-final phonotactic constraint to a word-final constraint—and Chong (2017) artificial language learning experiment—which establishes a direct correlation between the degree to which a phonotactic constraint is expressed in stems and the learner's ability acquire an alternation driven by that constraint—suggest that learners favor phonological principles that are general across morphosyntactic domains.

On the other hand, the bias toward morphosyntactically general phonologies must constitute a soft learning bias (Goldsmith 1990; Beckman 1998; Lombardi 1998; de Lacy 2002, 2006; Wilson 2006; Moreton & Pater 2012a, b; Staubs 2014; *a.o.*), or else we would not observe structure-sensitive phonology whatsoever in language. Moreover, there is some reason to suspect that the generality preference can be overridden. Vaux (1998) and Paster (2013), for example, provide the case of Marash Armenian, wherein adjacent root vowels must agree in backness and roundness, but affixed words can be disharmonic. Moreover, Archangeli & Pulleyblank (2007) discuss the case of Ngbaka, wherein root vowels must agree for [ATR], but affix vowels do not alternate. Returning to Chong (2017), the dissertation features a second corpus investigation into another apparent derived environment effect—Turkish velar deletion (Lewis 1967; Sezer 1981; Inkelas 2000, 2011, 2014; *inter alia*), whereby velar obstruents delete intervocalically when the velar neighbors a boundary (/bebek+A/ → [bebe-e], 'baby'-DAT)—showing that $VC_{[+\mathrm{DORS}]}V$ sequences are in fact not underattested in roots. The alternation itself is not entirely productive, it would seem (Inkelas 2011): though deletion applies when the stem is velar-final and the affix is vowel initial ($VC_{[+\mathrm{DORS}]}+V$), it does not apply if the velar is contained in the affix ($V+C_{[+\mathrm{DORS}]}V$); in addition, deletion does not apply if the stem is verbal; and finally, the process as a whole has lexical exceptions. Yet the alternation nevertheless applies in *some* derived environments, albeit not all of them, suggesting a moderate degree of productivity. Finally, Chong (2017) investigates assibilation in Finnish, a proposed derived environment effect whereby /t+i/ maps to [s-i], but [ti] is permitted morpheme internally (e.g., /tilat+i/ → [tilas-i], *[silas-i], *[tilat-i], 'order'-PAST) (Kiparsky 1973, 1993; Karlsson 1983; Anttila 2006). Chong shows that, as in the Turkish case, there is no evidence for underattestation of [ti] in the Finnish lexicon. Though assibilation does not apply in all derived environments, three suffixes regularly undergo assibilation, and one suffix optionally assibilates, to avoid [t+i] (Karlsson 1983; Anttila 2006). The Finnish system suggests that, once again, that derived environment effects can persist in language, at least to a degree. All this being said, it is an open question whether there exists a language with a complete derived environment effect (Inkelas 2011)—for example, a language Finnish′ whereby all eligible suffixes undergo assibilation, with phonotactics showing no dispreference toward [ti] sequences.

In the following section, I present a case of extreme morphosyntactic specificity—not quite Finnish′—but still one that deserves a place in the discussion of morphosyntactic generality and specificity in phonological learning: the case of Malagasy backness dissimilation, whereby dissimilation applies to a single suffix in the language—the only suffix even eligible to undergo dissimilation—even though phonotactics displays no dissimilatory preference whatsoever, but in fact a weak but highly significant backness harmony preference.

7.2 A case of extreme specificity: Malagasy backness dissimilation targeting a single affix, with no accompanying phonotactic tendency

It would seem that the above research into morphosyntactic generality bias is pointing towards the following conclusion: within any language, given that a morpheme or a set of morphemes in a domain undergoes a phonological alternation, we should find accompanying evidence for the alternation-driving constraint elsewhere in the language too (e.g., phonotactics). The grammar of Malagasy (Austronesian; Madagascar), as I will argue, challenges this conclusion, and complicates our current understanding of learners' tendency to posit morphosyntactically general constraints. Malagasy displays backness dissimilation, an alternation that has persisted across multiple generations that sends a back vowel to front in the presence of a nearby back vowel. The process applies very consistently to the passive imperative suffix, *–u*, and displays blocking behavior typical of dissimilation, suggesting the working of an OCP constraint. But *–u* is the only affix in the language that undergoes dissimilation, and is the only *suffix* even eligible to undergo it. Moreover, stems in the lexicon show no preference for dissimilation whatsoever; in fact, they display a modest but highly significant opposing preference for harmony. This suggests that Malagasy learners induce a morphologically specific OCP constraint—specific either to *–u* alone or to the suffix domain as a whole—without the need for a corroborating phonotactic trend. These findings suggest that *no* degree of morphosyntactic generality is a *necessary* condition for learning. Though learners might be biased towards acquiring grammatically general constraints, the Malagasy system suggests that they are capable of overriding this bias completely. I present this system below, and discuss the problems it poses for a theory in which learners favor grammatically general constraints.

7.2.1 Backness dissimilation applying to the passive imperative suffix

Unless otherwise specified, the data below come from the Malagasy Dictionary and Encyclopedia of Madagascar (hereafter MDEM; malagasyword.org; de la Beaujardière 2004), an annotated online corpus containing ~92,000 Malagasy words. The Malagasy vowel inventory is composed of [i e a u] (Parker 1883, de la Beaujardière 2004). There are four suffixes: the passive suffixes *–ina* and *–ana*, the active imperative suffix *–a*, and the passive imperative suffix *–u* (Parker 1883, Richardson 1885).

The passive imperative suffix conditionally undergoes backness dissimilation (Parker 1883, Zymet 2015): underlying *–u* (29a-b) surfaces as *–i* after stems containing *u* (30a-d) unless a front vowel intervenes (31a-b). The alternation conforms to patterns driven by the Obligatory Contour Principle (Leben 1973, Goldsmith 1976, *et seq*).

Underlying –u	(29a)	/bata+u/	[bata-u]	lift-PASS.IMP
	(29b)	/sava+u/	[sava-u]	inspect-PASS.IMP
Backness dissimilation	(30a)	/bab**u**+**u**/	[bab**u**-**i**]	plunder-PASS.IMP
	(30b)	/t**u**v+**u**/	[t**u**v-**i**]	fulfill-PASS.IMP
	(30c)	/s**u**av+**u**/	[s**u**av-**i**]	bless-PASS.IMP
	(30d)	/**u**^{n}dan+**u**/	[**u**^{n}dan-**i**]	bolster-PASS.IMP
Blocking by front vowels	(31a)	/t**uri**+**u**/	[t**uri**-**u**]	preach-PASS.IMP
	(31b)	/f**ules**+**u**/	[f**ules**-**u**]	thread-PASS.IMP

3,675 words in MDEM with the passive imperative suffix were extracted. The counts in Table 46 show that dissimilation is triggered by the presence of stem-internal *u*, applies regularly when the trigger is local and semi-regularly across *a*, and is regularly blocked by front vowels.

Context (ignoring consonants)	***–u***	***–i***	**Dissim. rate**	**Example**
No trigger	1877	7	0.0%	bata-u
Adjacent trigger	4	989	99.6%	bab**u-i**
Intervening *a*	196	201	50.9%	t**u**da-**i**
Intervening front vowel	399	2	0.4%	**turi-u**

Table 46: *Counts for Malagasy backness dissimilation*

Multiple lines of evidence suggest that Malagasy speakers acquire this alternation. Dissimilation is observed across at least two generations: it was reported as early as Parker (1883), and evidence for it appears in dictionaries since then (e.g., Abinal & Malzac 1888, Rajemisa 1985, de la Beaujardière 2004). Dissimilation and its blocking can be observed even when *–u* comes after loaned stems, as in (4a-d) below. The stems given below can be found in the World Loanword Database (wold.clld.org; Adelaar 2009), except /matsu/, which is marked as a loan in MDEM.

Dissimilation	(32a)	/mats**u**+**u**/	[mats**u-i**]	march-PASS.IMP	English loan
	(32b)	/kirar**u**+**u**/	[kirar**u-i**]	shoe-PASS.IMP	Bantu loan
	(32c)	/kuhukuh**u**+**u**/	[kuhukuh**u-i**]	cluck-PASS.IMP	Bantu loan
Blocking	(32d)	/bur**usi**+**u**/	[bur**usi-u**]	brush-PASS.IMP	French loan

Remarkably, the passive imperative suffix is the *only* affix to undergo dissimilation, and, assuming the process sends back vowels to front but not vice versa, is the only suffix even eligible to undergo it (being the only one to contain *u*). Even if we assume that dissimilation sends back vowels to front *and* vice versa, it is still not displayed by any other affix, according to an MDEM search. Given below are all affixes in MDEM occurring with at least 20 stems and that can place a front/back vowel tier-adjacent to a front/back root vowel. None of them alternate based on the root vowel (see http://malagasyword.org/bins/derivLists?form#longScroll), except for *–in–*/*–un–*, which displays some evidence of a harmony alternation (see Section 7.2.2 below).

Pref.	**# forms w/ pref.**	**Circumf.**	**# forms w/ circumf.**	**Inf.**	**# forms w/ inf.**	**Suff.**	**# forms w/ suff.**
fi- ‘manner of doing X’	2618	fi-…-ana ‘instance of X’	2144	-in-/-un- -PASS-	288+14	-ina/-na[10] -PASS	1700+32
ki-[11] ‘act of doing/ state of being X’	78	i-…-ana renders X into relative verb	1991				
mi- ACTIV-	4312	a^{m}pi-…-ina renders X into passive verb	31				
^{m}pi- ‘one who provides X’	1975						
t͡si- ‘instance of X’	46						
ku- ‘that which is X’	44						
fa^{m}pi- PASS-	53						
ma^{m}pi- ACTIV-	693						
^{m}pa^{m}pi- ‘one who provides X’	20						

Table 47: *different frequently occurring affixes and their counts*

[10] The counts of the *–na* allomorph might be inaccurate, as it also serves as the allomorph to another passive suffix, *–ana* (Richardson 1885). Regardless *–na* surfaces as a result of hiatus repair in the language (*cf.* Albro 2005, Lin 2005, O'Neill 2015, *a.o.*).

[11] *ki-*/*ku-* could be allomorphs of the same morpheme—but even if this is were true, their distributions do not appear to be conditioned by neighboring vowels (*ki-*: http://malagasyword.org/bins/derivLists?form=ki~#longScroll; *ku-*: http://malagasyword.org/bins/derivLists?form=ko~#longScroll).

If there were other evidence for a dissimilatory tendency in the grammar, we would expect to find it in phonotactics. We now turn to a corpus study of roots to assess whether this is the case.

7.2.2 A backness harmony tendency in Malagasy stem phonotactics

Surprisingly, roots display a modest but highly significant tendency toward backness *harmony*. MDEM gives numerous harmonic roots:

(33)	k**iri**	‘small hole’	sar**ut**^{r}**u**	‘cape’	**u**z**u**na	‘curse’
	l**u**f**u**	‘persistence’	**te**v**i**ka	‘spasm’	t͡s**i**ndr**i**	‘compression’
	g**egi**	‘indiscreet’	v**ulu**	‘color’	d**uku**	‘identity’

Counts of tier-adjacent pairs involving only front or back vowels (*i*, *e*, and *u*) were enumerated across 4,514 roots that were extracted from MDEM. The counts reveal no preference for disharmonic sequences in roots, as Table 2 reveals below. Note that the majority of roots in the corpus are classified as nouns (2,737), adjectives (729), or adverbs (733); verbs are derived through affixation (*cf.* Keenan & Polinsky 1998).[12]

[12] Some words displaying reduplication (*cf.* Lin 2005) were classified as roots in the corpus; in these cases, only the root involved in reduplication contributed to the counts, rather than the reduplicated stem as a whole. A conference reviewer points out that there could exist productive pseudoreduplication, with the first syllable being a copy of the second, potentially inflating the harmony rate. The corpus revealed that only 115 of the 4,514 roots have matching first and second syllables, with only 64 beginning with a front or back vowel ([didit^{r}a] = ‘twisting’, [vuvuka] = ‘dust’). It is not at all obvious that the language possesses pseudoreduplication, considering how low the count is here.

	# harmonic **V**C_0**V** *seq.s*	*# disharmonic* **V**C_0**V** *seq.s*	*# harmonic* **V**C_0aC_0**V** *seq.s*	*# disharmonic* **V**C_0aC_0**V** *seq.s*
Within noun roots	786	602	35	44
Within adj. roots	185	183	27	11
Within adv. roots	312	188	109	49
Within interj., conj., prep. roots	96	41	24	14
TOTAL	**1379**	**1014**	**205**	**118**

Table 48: *Raw counts of (dis/)harmonic sequences in roots*

There are around 350 more local harmonic sequences than local disharmonic sequences, and around 100 more nonlocal harmonic sequences than nonlocal disharmonic sequences. This makes backness dissimilation highly morphologically specific: it requires reference to the suffix domain or to the passive imperative suffix in particular, and lacks a counterpart generalization in stem phonotactics.

In fact, as we will see, Malagasy displays a significant tendency toward backness *harmony* in roots—these counts are unlikely to have arisen by chance alone. The observed rates of local and nonlocal harmony are 1379/(1379 + 1014) = 57.3% and 205/(205+ 108) = 63.5%, respectively. We can calculate the expected local harmony rate given the frequencies of front and back vowels by extracting from roots all V_1V_2 sequences in which each vowel belongs to [i e u], and calculating [$p(V_1 = u) \times p(V_2 = u)$] + [$p(V_1 = i$ or $e) \times p(V_2 = i$ or $e)$], where e.g. $p(V_1 = u)$ is the number of instances of *u* in V_1 position divided by the number of instances of *i*, *e*, and *u* in V_1 position. The expected rate of nonlocal harmony is computed analogously over V_1aV_2 sequences. Doing this, we obtain 51.6% and 57.7% as expected rates of local and nonlocal harmony. Comparing the observed and expected rates, we find that observed rates (local: 57.3%; nonlocal: 63.5%) are higher than expected (local: 51.6%; nonlocal: 57.7%). To determine whether harmonic sequences occur *significantly* more than chance would predict, we can run a Monte Carlo simulation (Kessler 2001). To run a simulation for local vowel sequences, we gather pairs of tier-adjacent vowels belonging to [i e u], shuffle the second vowels of each pair and randomly concatenate each of them to a first vowel, calculate the new harmony rate, and then repeat 10,000 times. The simulation for nonlocal sequences (V_1aV_2) can be computed analogously. Figures 25a-b below show histograms of (non/)local harmony rate frequencies after the 10,000 trials.

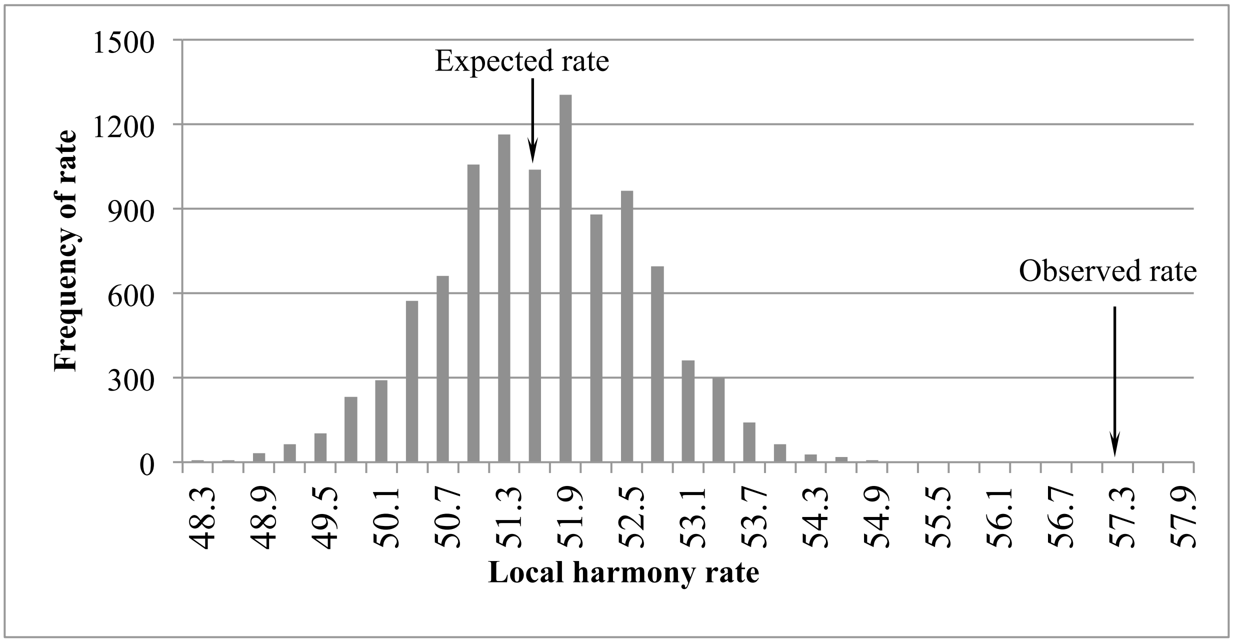


Figure 25a: *distribution of local harmony rates yielded by Monte Carlo trials*

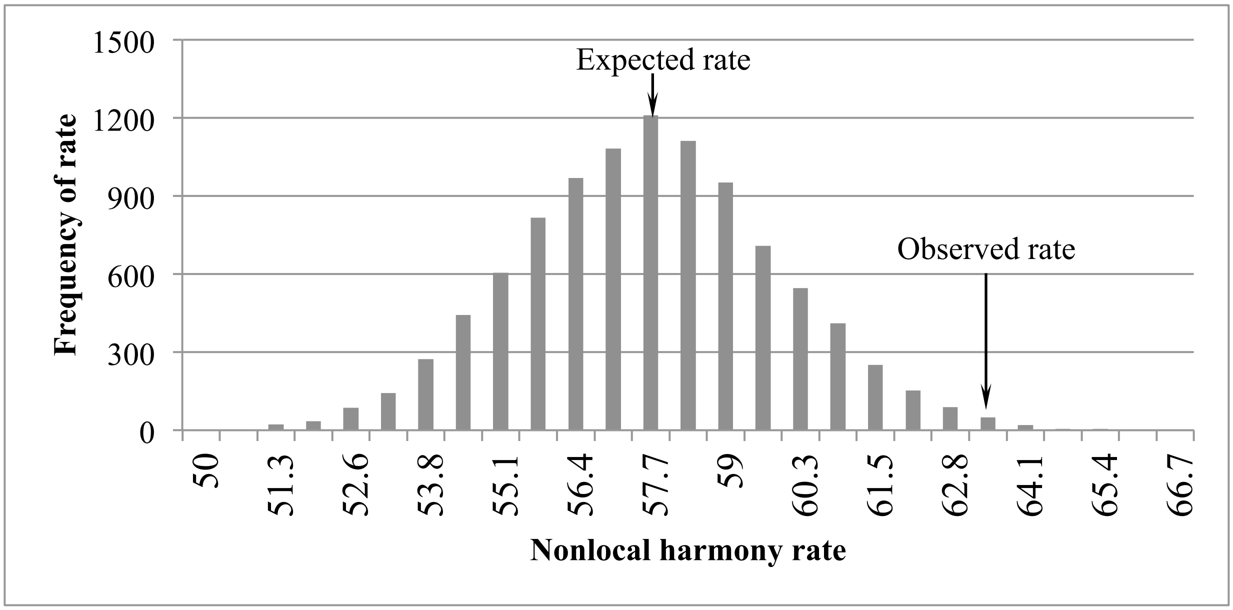


Figure 25b: *distribution of nonlocal harmony rates yielded by Monte Carlo trials*

For local harmony, the observed rate of 57.3% is greater than *any rate* yielded by 10,000 trials, and is thus significantly greater than chance would predict (est. $p < \frac{1}{10000} = 0.0001$). For nonlocal harmony, the observed rate of 63.5% is greater than 9,834 of the trials, and is thus significantly above chance as well (est. $p = \frac{10000-9834}{10000} = 0.008$). The results suggest that overrepresentation is not coincidental, but rather reflects a backness harmony preference in phonotactics. Note that there exists some evidence of a harmony alternation as well. The *–in–*/*–un–*infix is used to create passive verbs ([sava] = clear, [s-in-ava] = clear-PASS; [fidi] = choice, [f-in-idi] = choice-PASS). MDEM gives 288 words with *–in–* and 14 with *–un–*. *–in–* can surface before any vowel, and in particular surfaces before *u* in 56 forms. But in the 14 forms with *–un–*, the following vowel is always *u* ([b**u**ri] = round; [b-**u**n-**u**ri] = round-PASS), suggesting that *–un–* is selected to satisfy a (weak) harmony drive. That an infix can harmonize while another suffix dissimilates is reminiscent of Yucatec Maya, in which two suffixes harmonize for backness and height, but another dissimilates for backness, and yet another for backness and height (Blair

1964; see Krämer 2001 for an account). Altogether, these cases suggest that *contradictory* markedness preferences can target different morphemes or domains.

To summarize, Malagasy backness dissimilation applies consistently to the passive imperative suffix and displays blocking behavior typical of OCP. Roots, however, show no dissimilatory tendency, but rather a modest but highly significant harmony preference.

7.2.3 Discussion and potential analytical directions

The Malagasy system provides evidence that the learner can counteract the tendency to favor morphosyntactically general constraints. This finding patterns with other instances of learning bias defiance, in which systems that have been suggested to be disfavored by learners occasionally arise in the world's languages and persist across generations, providing evidence that they can be apprehended to some extent (Hayes, Zuraw et al. 2009; Hayes & White 2015; Merrill 2015; Beguš & Nazarov 2017).

The Malagasy system complicates the picture of how a morphosyntactic generality bias in phonological learning should be modeled. Martin (2011) finds that phonotactic constraints can "leak" into the cross-boundary domain: in Navajo sibilant harmony and English geminate avoidance, a categorical generalization within roots is mirrored by a statistical tendency across compound boundaries. To account for this, Martin introduces a Gaussian smoothing term into a MaxEnt learning system so that when the learner weights positively a structure-specific constraint (e.g., applying only stem-internally), it gives small positive weight to an analogous domain-general constraint, leading over time to a grammar with the morphosyntactic generality property. A model in which the usage of a structure-specific constraint implies the usage of an analogous structure-insensitive constraint cannot be applied to Malagasy, at least without further elaboration.

Two potential solutions to the problem are entertained. The first is to say that while any particular *affix* is allowed to depart from typical phonological behavior in a language, whole *domains* must overall respect the generality property, at least to a degree. We can say that OCP targets the passive imperative suffix in Malagasy, rather than the entire suffix domain, and so no generalizing tendency should arise. Although this would be a possible approach, we cannot be *sure* that OCP targets *–u* rather than the entire suffix domain: it could be that OCP in Malagasy is triggered only by back vowels and is indexed to the suffix domain, and thus the one suffix with a back vowel, *–u*, undergoes dissimilation. Nonetheless, corpus studies undertaken by Chong (2017) support indexing OCP to *–u* rather than to its domain, as they discount claims of the existence of certain derived environment effects—that is, domain-level mismatches: though prior investigators show that palatalization in Korean (Kiparsky 1973,1993; Iverson & Wheeler 1988) avoid sound sequences that are found in *some* of the languages' roots, Chong shows that such roots are underattested. Korean therefore still displays the morphosyntactic generality property, at least for the most part. That being said, it may be that the generalizing tendency is not universal even for domains, as aforementioned: Finnish shows *no* tendency against [ti] sequences in roots, but three suffixes regularly undergo assibilation, and one suffix optionally assibilates, to avoid [t+i] (Kiparsky 1973, 1993; Karlsson 1983; Anttila 2006; Chong 2017). The Finnish system suggests that even domains can, to some extent, mismatch overall.

Another possible solution is to say that a generalizing bias even applies in the Malagasy case, but that Malagasy learners make use of a harmony constraint that counteracts leaking of the dissimilatory drive into phonotactics. One can imagine that a learner with a generalizing bias, upon encountering the Malagasy system, would invoke a morphologically specific OCP constraint, and then “smooth” over the grammar with a general OCP constraint, so that the dissimilatory drive leaks into stems. This alone could not account for the Malagasy system, since no dissimilatory tendency is observed in phonotactics. Thus, to correct for this, the learner could weight positively a harmony constraint so that the phonotactic dissimilatory tendency is cancelled or overridden (see Zymet 2018 for a MaxEnt model involving this). Some evidence indeed suggests that learners can make use of constraints driving dissimilation in some morphemes or domains but harmony in others: after all, Malagasy displays consistent dissimilation to the passive imperative suffix, but a harmony tendency in phonotactics; in addition, backness dissimilation and harmony constraints seem to condition allomorphy in different suffixes in Yucatec Maya (Blair 1964, Krämer 2001). One might wonder, then, why contradictory-preferences systems are so typologically infrequent. Perhaps they are tied to *backness restriction* in particular. The cases of leaking found in Martin (2011) involve sibilant harmony and geminates; considering that grammars *preferring* disharmonic sibilants or geminates are rare or unattested, we might imagine that learners would not entertain such preferences as hypotheses about different grammatical contexts. As a result, sibilant harmony or geminate avoidance found in one grammatical context would leak into another. But backness harmony *and* dissimilation are observed crosslinguistically (Parker 1883, Esztergár 1971, Campbell 1977, Clements & Sezer 1982, Itô 1984, Harrison 1999, *a.o.*), and so it may be that the learner can entertain constraints driving both backness harmony and dissimilation in hypotheses about these different contexts. Learners might spread the effect of one of these constraints across contexts (e.g., dissimilation), but counteract the effect using the natural opposing constraint (harmony). It could be that generalization effects are only defied in cases where there exists crosslinguistic evidence for the working of two opposing constraints, as in backness dissimilation and harmony. Where there does not, languages requiring restrictions specific to grammatical context may be relatively prone to being generalized.

How might the Malagasy system have arisen if the generalizing bias is true? Here the picture is unclear, but we can speculate: the passive imperative suffix may have been adopted late in the language's development, with dissimilation arising to distinguish the suffix boundary—a drive for recoverability that would directly conflict with the generalizing bias. Or perhaps dissimilation began as a constraint against *u*+*u* sequences, mirroring a ban on pairs of directly adjacent *u*'s in phonotactics, but was somehow generalized to *u*…+*u* sequences. This is a topic here left for further research.

7.2.4 Summary of the Malagasy case

Several findings now suggest that learners tend to favor morphosyntactically general phonological constraints. This section argues that this bias, if it exists, can be overridden. Malagasy backness dissimilation applies very consistently to the passive imperative suffix –*u*, and displays blocking behavior typical of OCP. But –*u* is the only affix in the grammar that undergoes it, and is the only suffix even eligible to undergo it. Stems, on the other hand, display a modest but significant harmony trend. This suggests that Malagasy learners induce a morphologically specific OCP constraint—specific either to –*u* or to the suffix domain as a whole—without the need for a corroborating phonotactic trend. These findings suggest that *no* degree of morphosyntactic generality is a *necessary* condition for learning. Though learners might favor grammatically general constraints, the Malagasy system suggests that they are capable of overriding this bias completely.

7.3 Towards a broad understanding of generality and specificity at the phonology-morphosyntax interface

At present, how *granular* a phonology *can* be—how learner-driven generality effects arise in the phonology-morphosyntax interface, and the extent to which structure specificity in phonological principles can be acquired by learners and persist in grammar—is an open question. To give an overview, the prior studies reviewed in this chapter and the dissertation overall have presented the following evidence for morphosyntactic generality bias in phonological learning: phonological alternations are typically accompanied by a corresponding phonotactic generalization (Chomsky & Halle 1968; Kenstowicz & Kisseberth 1977, 1979; McCarthy 2002; *et seq*); there exist cases in which a strong phonotactic generalization is generalized, "leaking" into other domains such as compound well-formedness (Martin 2007, 2011); Korean palatalization as a derived environment effect is merely apparent—phonotactics in fact displays a strong bias towards the palatalizing restriction (Chong 2016, 2017); in artificial language learning experiments, participants generalize a phonological principle from one morphosyntactic domain to another, without unambiguous evidence (Myers & Padgett 2014); in artificial language learning experiments, participants more readily acquire a harmony alternation when there is accompanying evidence for harmony in phonotactics (Chong 2017); and finally, learners extract broad phonological principles from datasets displaying high degrees of lexical idiosyncrasy, rather than simply internalizing principles on a word-by-word basis (Zuraw 2000, Hayes & Londe 2006, this dissertation; *inter alia*). On the other hand, other studies have presented the following evidence for morphosyntactic specificity in phonological learning: there are a couple—albeit less productive—cases where phonotactics and alternations mismatch, in particular, Turkish velar deletion and Finnish assibilation (Lewis 1967; Kiparsky 1973, 1993; Sezer 1981; Karlsson 1983; Vaux 1998; Anttila 2006; Archangeli & Pulleyblank 2007; Inkelas

2011; Paster 2013; Chong 2017; *inter alia*); Malagasy, furthermore, displays backness dissimilation to a single suffix—the only suffix eligible in the domain to undergo the process—without showing an accompanying tendency in phonotactics (Zymet 2017, this dissertation); though they extract generalizations from datasets displaying lexical variation, speakers nonetheless internalize the fixed pronunciations of attested words (Zuraw 2000, Hayes & Londe 2006, this dissertation; *inter alia*); and finally, French speakers even internalize word-specific lexical propensities in liaison, to some degree of accuracy (this dissertation). How can these facts be reconciled in the learning theory of the phonology-morphosyntax interface? On what grounds does the learner posit a structure-specific constraint, and on what grounds does the learner posit a completely general constraint—insensitive to structure of any kind, whether it be morpheme or morphosyntactic domain? When must the learner posit a set of morpheme-specific constraints, rather than a domain-specific, category-specific, or structure-insensitive constraint? And how do we model the learner's positing of structure-sensitive constraints? Perhaps a clustering algorithm lies in the future of the theory of phonological learning, either bottom-up or top-down: for example, a bottom-up algorithm whereby the learner posits morpheme-specific constraints until she realizes that a broader generalization can be said to govern the data, whether it be domain-specific, category-specific, or structure-insensitive entirely—much like what has been recently proposed in Shih (2018); or one whereby the learner posits a broad, structure-insensitive generalization, until she realizes that smaller clusters must be formed in cases where the evidence is strong enough—at first she might try domain- or category-specific constraints, but if the attempt fails to explain the ambient input sufficiently, then she might try morpheme-specific constraints. Maybe the best way to encode domain specificity or category specificity, furthermore, is as a random effect—much like morphemes—with grammar leaking simply resulting from this distinction. Whatever the answers to these questions may be, they must explain the facts discovered so far—a considerable challenge for investigators to tackle in future studies.

Chapter 8:

Conclusion

Several theories of variation propose that encoded in morphemes is a binary scale ([+/- Rule X]) that determines whether they trigger or undergo a phonological process. Other, more recent theories raise the possibility that encoded in morphemes are gradient parameters ([0.7 Rule X]), predicting that they should display lexical—idiosyncratic, gradient rates at which they trigger or undergo a process. In this investigation, I argue that individual morphemes—both triggering morphemes and undergoing morphemes—can display differing propensities to participate in a variable process, and that learners internalize these propensities. The evidence for these claims comes from a series of corpus investigations into variable Slovenian palatalization and French liaison, and a nonce probe investigation into the intuitions of native French speakers, the results of which suggest that French learners internalize the liaison propensities of different Word1s. These findings favor theories that are capable of encoding a morpheme's status on an entire spectrum, and would suggest that learners are capable of tracking morpheme-specific rates of allomorphy. Moreover, it validates theories that allow encoding on a morpheme-by-morpheme basis, and suggests that variation cannot be explained by referring to idiosyncrasies of stored larger constituents alone, or morphemes grouped arbitrarily or by semantic or lexical class alone. The theory of the grammar and lexicon must elucidate how such idiosyncratic propensities are represented and learned.

Such theories would share the following goals: 1) capturing language learners' behavior to frequency match to statistical generalizations found across the lexicon; 2) capturing the idiosyncratic behavior of individual words or morphemes to abide by or deviate from these trends. The modeling section primarily focused on a recent MaxEnt-based model for learning a frequency-matching grammar together with lexical propensities, which makes use of general constraints that putatively frequency match to general trends across the lexicon, as well as lexically specific constraints that govern the behavior of individual words. In this model, general constraints and lexical constraints are treated as *equally viable* explanatory variables for learning the dataset and its patterns. A series of learning simulations revealed that the approach fails to learn general, grammatical trends for this very reason, as it runs into a grammar-lexicon balancing problem: the lexical constraints are so powerful that the learner acquires the behavior of each word in the dataset well before the general constraints are strong enough to capture the grammatical trends, at which point grammar learning ceases. A generality bias was therefore attributed to learners, such that they privilege general constraints over lexical ones. MaxEnt—essentially an ordinary logistic regression model—fails to represent this property. This dissertation argued that it should replaced with a mixed-effects logistic regression model—Mixed-Effects Maximum Entropy Harmonic Grammar—which was shown to succeed in learning both grammatical *and* item-specific behavior by encoding general constraints as fixed effects and lexical constraints as random effects. The learner treats the grammar and lexicon differently, in that vocabulary effects are subordinated to broad, grammatical effects in the learning process. Mixed-effects logistic regression is used widely in linguistics experiments and across scientific fields. My purpose has been to make the case for why it should be adopted as a model of the language learner.

References

Abinal, Antoine & Victorin Malzac. 1888. *Dictionnaire malgache-français.* Ambozoantany.

Adelaar, Alexander. 2009. Malagasy vocabulary. In: Haspelmath, Martin & Tadmor, Uri (eds.) *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Ågren, John. 1973. *Etude sur quelques liaisons facultatives dans le français de conversation radiophonique: frequence et facteurs*. Uppsala: Acta Universitatis Upsaliensis.

Akaike, Hirotugu. 1973. Information theory and an extension of the maximum likelihood principle. in Petrov, B. N.; Csáki, F., *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, 267–281.

Albright, Adam. 2002. Islands of reliability for regular morphology: Evidence from Italian. *Language 78*: 684-709.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition 90*, 119–161.

Albro, Daniel. 2005. *Studies in Computational Optimality Theory, with special reference to the phonological system of Malagasy.* Ph.D. dissertation, UCLA.

Anttila, Arto. 1997. Deriving variation from grammar. In F. Hinskens, R. van Hout & L. Wetzels (eds.) *Variation, change and phonological theory*. Amsterdam: Benjamins, 35-68.

Anttila, Arto. 2002. Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory 20*: 1-42.

Anttila, Arto. 2006. Variation and opacity. *Natural Language & Linguistic Theory* 24, 893–944.

Archangeli, Diana, Douglas Pulleyblank. 2007. Harmony. In: de Lacy, P. (Ed.), *The Handbook of Phonology*. Cambridge University Press, Cambridge, 353-377.

Armstrong, Nigel. 2001. Social and Stylistic Variation in Spoken French: a comparative approach. Amsterdam: John Benjamins.

Baayen, R. Harald. 2004. Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers 1*, 1-45. University of Edmonton.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. Harald, Dogulas J. Davison, &Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 390-412.

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language 44*, 568–591.

Bajec, Anton et al. 2000. *Slovar slovenskega knjižnega jezika: Electronic edition*. Ljubljana: SAZU and Fran Ramovš Institute for the Slovenian Langauge.

Barreca, Giulia & George Christodoulides. 2017. Analyse fréquentielle de la liaison variable dans un corpus de français parlé. *Journal of French Language Studies 27*: 27-40.

Bates, Douglas. 2009. Linear mixed model implementation in lme4. Ms., University of Wisconson—Madison. Available at http://econ.ucsb.edu/~doug/245a/Papers/Mixed%20Effects%20Implement.pdf.

Bates, Douglas & Martin Maechler. 2011. Package 'lme4'. R.

Baude, Olivier & Céline Dugua. 2011. (Re)faire le corpus d'Orléans quarante ans après: quoi de neuf, linguiste? *Corpus* 10: 99-118.

Becker, Michael. 2009. *Phonological Trends in the Lexicon: The Role of Constraints*. Doctoral Dissertation, University of Massachusetts, Amherst.

Becker, Michael & Maria Gouskova. 2016. Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry 47:3*, 391-425.

Becker, Michael & Jonathan Levine. 2013. Experigen – an online experiment platform. Available at http://becker.phonologist.org/experigen.

Becker, Michael, Andrew Nevins & Nihan Ketrez. 2011. The Surfeit of the Stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language 87:1*, 84–125.

Becker, Michael, Andrew Nevins & Jonathan Levine. 2012. Asymmetries in generalizing alternations to and from initial syllables. *Language 88:2*, pp. 231–268.

Beguš, Gašper & Aleksei Nazarov. 2017. Lexicon against naturalness: Unnatural gradient phonotactic restrictions and their origins. Ms., Harvard University.

Bergounioux G., Baraduc J., Dumont C. 1992. L'Etude socio-linguistique sur Orléans (1966-1991), 25 ans d'histoire d'un corpus. *Langue française* 93: 74-93.

Berkley, Deborah Milam. 2000. Gradient obligatory contour principle effects. Doctoral dissertation, Northwestern University.

Blair, Robert. 1964. *Yucatec Maya noun and verb morpho-syntax.* Ph.D dissertation, Indiana University.

Blanc, Michael & Biggs, Patricia. 1971. L'enquête socio-linguistique sur le français parlé à Orléans. *Le Français dans le Monde* 85: 16–25.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* 21: 43–58. University of Amsterdam.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry 32*: 45-86.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24(3)*: 127–135.

Boula de Mareuil, P., Gendner, V., Adda-Decker, M. 2003. Liaisons in French: a corpus-based study using morphosyntactic information. *ICPhS*, Barcelona, Aug 2003.

Burnham, Kenneth P. & David R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research 33*: 261–304.

Burzio, Luigi. 1997. Cycles, non-derived-environment blocking, and correspondence. Ms., Johns Hopkins University.

Bybee, Joan. 2001. Frequency effects on French liaison. In J. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*, 337–359. Amsterdam: John Benjamins.

Bybee, Joan. 2002. Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition 24*, 215–221.

Campbell, Lyle. 1977. *Quichean Linguistic Prehistory*. University of California Press, Berkeley, CA.

Cedergren, H., & Sankoff, D. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50:333-355.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English.* New York, NY: Harper and Row.

Chong, Adam J. 2016. Gradient phonotactics and derived-environment effects: A comparison of Korean and Turkish. Presentation given at AMP2016.

Chong, Adam J. 2017. On the relation between phonotactic learning and alternation learning. Ph.D. Dissertation, University of California, Los Angeles.

Clements, George & Engin Sezer. 1982. Vowel and consonant disharmony in Turkish. In Harry van der Hulst and Norval Smith (eds.), *Structure of phonological representations (Part II)*, 213-255. Dordrecht: Foris Publications.

Coetzee, Andries & Joe Pater. 2008. The place of variation in phonological theory. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The Handbook of Phonological Theory* (2nd edn.). Malden, MA: Blackwell. Available at: http://roa.rutgers.edu/.

Coetzee, Andries & Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguisic Theory 31*: 47-89.

Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In Third Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop, ed. by John Coleman, 49–56. East Stroudsburg, PA: Association for Computational Linguistics.

Côté, Marie-Hélène. 2011. Understanding cohesion in French liaison. In the *Blackwell Companion to Phonology*, ed.s Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume and Keren Rice.

Côté, Marie-Hélène. 2013. Understanding cohesion in French liaison. *Language Sciences 39*:156-166.

de la Beaujardière, Jean Pierre. 2004. *Malagasy Dictionary and Encyclopedia of Madagascar.*

Delattre, Pierre. 1951. *Principes de phonétique française à l'usage des étudiants anglo-américains*. Middlebury College.

Delattre, Pierre. 1966. *Studies in French and Comparative Phonetics*. The Hague: Mouton.

Dell, François. 1973/1985. *Les règles et les sons*. 2nd edition, 1985. Paris: Hermann.

Durand, Jacques, Bernard Laks & Chantal Lyche. 2002. La phonologie du français contemporain: usages, variétés et structures. In: *Romanistische Korpuslinguistik Korpora und gesprochene sprache/Romance Corpus Linguistics – Corpora and Spoken Language*. Ed. by Claud Pusch and Wolfgang Raible. Tübingen: Gunter Narr. 93–106.

Durand, Jacques & Chantal Lyche. 2008. French liaison in the light of corpus data. *Journal of French Language Studies* 18: 33–66.

Durand, Jacques, Bernard Laks, & Chantal Lyche. 2009. Le projet PFC (phonologie du français contemporain): une source de données primaires structurées. In: *Phonologie, variation et accents du français*. Paris: Hermès, 19–61.

Eddington, David. 1998. Spanish diphthongization as a non-derivational phenomenon, *Rivista di Linguistica 10*: 335-354.

Eddington, David. 2004. *Spanish Phonology and Morphology: Experimental and Quantitative Perspectives.* Amsterdam: John Benjamins.

Efron, Bradley & Morris, Carl. 1975. Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association 70*, 311-319.

Ernestus, Mirjam and R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language 79*, 5–38.

Esztergár, Marianne. 1971. *A generative phonology of nouns and vowel harmony in Hungarian.* Ph.D. dissertation, UC San Diego.

Finley, Sara. 2012. Typological asymmetries in round vowel haramony: Support from artificial grammar learning. *Language and Cognitive Processes* 27: 1550-1562.

Frisch, Stefan A., Janet B. Pierrehumbert & Michael Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:179–228.

Jaeger, T. Florian. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language* 59 (4), 434-446

Frisch, Stefan A. & Zawaydeh, Bushra. 2001. The psychological reality of OCP-Place in Arabic. *Language 77*, 91-106.

Fougeron, Cécile, Jean-Philippe Goldman, Alicia Dart, Laurence Guélat & Clémentine Jeager. 2001a. Influence de facteurs stylistiques et lexicaux sur la réalisation de la liaison en français. *Proceedings of TALN 2001*, 173–182. Available at http://tln.li.univ-tours.fr/Tl.../TALN2001-RECITAL2001/index1.html.

Fougeron, Cécile, Jean-Philippe Goldman & Uli H. Frauenfelder. 2001b. Liaison and schwa deletion in French: An effect of lexical frequency and competition? In Paul Dalsgaard, Børge Lindberg, Henrik Benner & Zheng- hua Tan (eds.) *Proceedings of Eurospeech 2001*, 639–642. Aalborg, Denmark: ISCA Archive.

Fruehwald, Josef T. 2012. Redevelopment of a Morphological Class. *University of Pennsylvania Working Papers in Linguistics 18(1)*. Available at: https://repository.upenn.edu/pwpl/vol18/iss1/10.

Fylstra, D.; Lasdon, L.; Watson, J.; and Waren, A. 1998. Design and use of the Microsoft Excel solver. *Interfaces*, Vol. 28, No. 5, 29-55.

Gelman, Andrew. 2004. Exploratory Data Analysis for Complex Models (with Discussion). *Journal of Computational and Graphical Statistics*, 13(4): 755–779.

Gibson, Edward, Steve Piantadosi, & Kristina Fedorenko. 2011. Using Mechanical Turk to Obtain and Analyze English Acceptability Judgments. *Language and Linguistics Compass 5.8*: 509–524.

Goldsmith, John. 1976. *Autosegmental phonology*. IULC and Garland Press, New York.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In the *Proceedings of the Stockholm workshop on variation within Optimality Theory.*

Gouskova, Maria & Tal Linzen. 2015. Morphological conditioning of phonological regularization. *The Linguistic Review 32:3*, 427-473.

Guy, Gregory. 1980. Variation in the group and the individual: the case of final stop deletion. In W. Labov, ed., *Locating Language in Time and Space*, 1-36. New York: Academic Press.

Guy, Gregory. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change 3*:1-22.

Guy, Gregory. 1997. Violable is variable: Optimality theory and linguistic variation. *Language Variation and Change* 9: 333-348.

Hallanger, Frederik S. 1974. *Diksionera Malagasy-Frantsay*. Trano Printy Loterana.

Hansson, Gunnar Ólafur. 2001. *Theoretical and typological issues in consonant harmony*. Ph.D. Dissertation, University of California, Berkeley.

Harris, Daniel. 1998. Nonlinear Least-Squares Curve Fitting with Microsoft Excel Solver. *Journal of Chemical Education* 75(1).

Harrison, K. David. 1999. Vowel harmony and disharmony in Tuvan and Tofa. *Proceedings of the Nanzan GLOW.*

Hayes, Bruce. 2000. Gradient well-formedness in Optimality Theory. In Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer, eds., O*ptimality Theory: Phonology, Syntax, and Acquisition*, Oxford University Press, 88-120.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. In Kager, Rene, Joe Paterand Wim Zonneveld, (eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press.

Hayes, Bruce & Zsuzsa Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23: 59-104.

Hayes, Bruce & Margaret MacEachern. 1998. Quatrain form in English Folk Verse. *Language 64*: 473-507.

Hayes, Bruce, Kie Zuraw, Peter Siptar & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85: 822-863.

Hayes, Bruce & James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44:45-75.

Hayes, Bruce & James White. 2015. Saltation and the P-map. *Phonology* 32:267-302.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39, 379–440.

Inkelas, Sharon. 2011. Another look at velar deletion in Turkish, with special attention to the derived environment condition. In E. E. Taylan and Bengisu Rona (eds.), *Puzzles of language: Essays in honour of Karl Zimmer*.

Inkelas, Sharon & Cheryl Zoll. 2005. *Reduplication: Doubling in Morphology.* Cambridge: Cambridge University Press.

Irwin, Mark. 2005. Rendaku-Based Lexical Hierarchies in Japanese: The Behaviour of Sino-Japanese Mononoms in Hybrid Noun Compounds. *Journal of East Asian Linguistics 14*: 121-153.

Itô, Junko. 1984. Melodic dissimilation in Ainu. *Linguistic Inquiry* 15, 505-13.

Iverson, Gregory K. & Deirdre W. Wheeler. 1988. Blocking and the Elsewhere condition. In Michael Hammond & Michael Noonan (eds.), *Theoretical morphology: Approaches in modern linguistics*, 325 – 338. San Diego: Academic Press.

Jong, Daan de. 1994. La sociophonologie de la liaison orléanaise. In Chantal, Lyche (ed.) *French generative phonology: Retrospective and perspectives*, 95–130. Salford: Association for French Language Studies.

Kilbourn-Ceron, Oriana. 2017. Speech production planning affects variation in external sandhi. Doctoral dissertation, McGill University.

Klausenberger, Jürgen. 1984. French liaison and linguistic theory. Stuttgart: Franz Steiner Verlag Wiesbaden GMBH.

Jarosz, Gaja. 2006. *Rich lexicons and restrictive grammars: Maximum likelihood learning in Optimality Theory*. Ph.D. Dissertation, Johns Hopkins University.

Jarosz, Gaja. 2017. Defying the Stimulus: Acquisition of Complex Onsets in Polish. *Phonology 34(2)*: 269-298.

Jarosz, Gaja & Amanda Rysling. 2017. Sonority Sequencing in Polish: the Combined Roles of Prior Bias and Experience. *Proceedings of the 2016 Annual Meetings on Phonology, USC.*

Jurgec, Peter. 2016. Velar palatalization in Slovenian: Local and long-distance interactions in a derived environment effect. *Glossa* 1(1): 24.

Karlsson, Fred. 1983. *Suomen kielen äänne- ja muotorakenne* [*The Phonological and Morphological Structure of Finnish*]. Helsinki: Werner Söderström Osakeyhtiö.

Keenan, Edward & Maria Polinsky. 1998. Malagasy Morphology. In A. Zwicky and A. Spencer (eds.), *Handbook of Morphology*, Oxford University Press, Oxford, pp. 563-623.

Kenstowicz, Michael & Charles Kisseberth. 1977. *Topics in phonological theory.* New York: Academic Press.

Kessler, Brett. 2001. *The significance of wordlists*. Cambridge, MA: MIT Press.

Kiparsky, Paul. 1973. Phonological representations. In Osamu Fujimura (ed.), *Three dimensions of linguistic theory*, 1–135. Tokyo: TEC.

Kiparsky, Paul. 1993a. Blocking in nonderived environments. In Ellen Kaisse & Sharon Hargus (eds.), *Phonetics and Phonology 4: Studies in lexical phonology*, 277–313. San Diego: Academic Press.

Kiparsky, Paul. 1993. Variable rules. Handout of paper presented at Rutgers Optimality Workshop 1, Rutgers University.

Kitto, Catherine & Paul de Lacy. 1999. Correspondence and epenthetic quality. *Proceedings of AFLA 4*: 181-200.

Krämer, Martin. 2001. Yucatec Maya Vowel Alternations: Harmony as Syntagmatic Identity. *Zeitschrift für Sprachwissenschaft* 20: 175-217.

Kraska-Szlenk, Iwona. 1995. The phonology of stress in Polish. Ph.D. dissertation, University of Illinois, Urbana-Champaign.

Kullback, Solomon & Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22 (1): 79–86.

Labov, William. 1973. Sociolinguistic Patterns.. University of Pennsylvania Press.

Labov, William. 1989. The child as linguistic historian. *Language Variation and Change 1*, 85–97.

Labov, William. 1994. Principles of Linguistic Change: Internal Factors. Blackwell Publishing.

Labov, William. 2001. Principles of Linguistic Change, Vol. 2: Social Factors. Blackwell Publishers.

Leben, Will. 1973. *Suprasegmental phonology*. Ph.D. dissertation, MIT. Published by Garland Press, New York.

Levy, Roger. 2010. Probabilistic models in the study of language. Ms, University of California, San Diego.

Lewis, Geoffrey. 1967. *Turkish grammar*. Oxford: Oxford University Press.

Lin, Ying. 2005. Two perspectives on Malagasy reduplication: derivational and OT analyses. *UCLA Working Papers in Linguistics 12*.

Linzen, Tal, Sofya Kasyanenko & Maria Gouskova. 2013. Lexical and phonological variation in Russian prepositions. Phonology 30:3, 453-515.

Logar-Berginc, Nataša, Simon Krek, Tomaž Erjavec, Miha Grčar, Peter Halozan & Simon Šuster. 2012. *Giga da corpus*. http://www.giga da.net: Amebis.

Lonergan, J., J. Kay & J. Ross. 1974. Etude sociolinguistique sur Orléans, Catalogue des enregistrements. Colchester: Orléans Archive, University of Essex, Department of Language and Linguistics.

Lubowicz, Ania. 1998. Derived environment effects in Optimality Theory. In: Susan J. Blake, Eun-Sook Kim, and Emary Shahin (eds.), *Proceedings of the 17th West Coast Conference on Formal Linguistics*, Vancouver, British Columbia.

Mallet, Géraldine. 2008. *La liaison en français: Descriptions et analyses dans le corpus PFC*. Ph.D. dissertation, Université Paris Ouest, Nanterre La Défense.

Martin, Andrew. 2007. The evolving lexicon. Ph.D. Dissertation, UCLA.

Martin, Andrew. 2011. Grammars leak: Modeling how phonotactic generalizations interact with in the grammar. *Language* 87(4): 751–770.

McMullin, McMullin and Gunnar Ólafur Hansson. 2014. Locality in long-distance phonotactics: evidence for modular learning. In *Proceedings of the 44th Meeting of the North East Linguistic Society*. GLSA Publications, UMass, Amherst.

Merrill, John. 2015. Nasalization as a repair for voiced obstruent codas in Noon. *LSA Annual Meeting Extended Abstracts* 6:14, 1–5.

Moisset, Christine. 2000. Variable liaison in Parisian French. Ph.D. dissertation, University of Pennsylvania.

Morgan, Emily & Roger Levy. 2016. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition 157*:382–402.

Moore-Cantwell, Claire & Joe Pater. 2016. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. *Catalan Journal of Linguistics 15*, 53-66.

Moreton, Elliott & Joe Pater. 2012. Structure and Substance in Artificial-phonology Learning, Part I: Structure. *Language and Linguistics Compass*, 6(11), 686-701.

Morin, Yves-Charles. 1986. On the morphologization of word-final consonant deletion in French. In: H. Andersen (ed.), *Sandhi Phenomena in the Languages of Europe*. Berlin: Mouton de Gruyter, 167-210.

Morin, Yves-Charles and Kaye, Jonathan 1982. The Syntactic Bases for French Liaison, *Journal of Linguistics 18*: 291-330.

Myers, Scott & Jaye Padgett. 2014. Domain generalization in artificial language learning. *Phonology* (31): 399 433.

Nazarov, Aleksei. 2018. Learning Both Variability and Exceptionality in Probabilistic OT Grammars. *Proceedings of the Society for Computation in Linguistics*: Vol. 1 , Article 36.

New, Boris, Christophe Pallier, Ludovic Ferrand & Rafael Matos. 2001. Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique* 101, 447–462.

O'Neill, Timothy. 2015. *The phonology of Betsimisaraka Malagasy.* Ph.D. dissertation, University of Delaware.

Parker, George Williams. 1883. *A concise grammar of the Malagasy language*. London, Trübner & Co.

Paster, Mary. 2013. Rethinking the 'duplication problem'. *Lingua* 126: 78–91.

Pater, Joe. 2000. Nonuniformity in English stress: the role of ranked and lexically specific constraints. *Phonology* 17: 237--274.

Pater, Joe. 2010. Morpheme-Specific Phonology: Constraint Indexation and Inconsistency Resolution. In Steve Parker, (ed.) *Phonological Argumentation: Essays on Evidence and Motivation. London: Equinox*. 123-154.

Pater, Joe, Robert Staubs, Karen Jesney & Brian Smith. 2012. Learning probabilities over underlying representations. In the *Proceedings of the Twelfth Meeting of the ACL-SIGMORPHON: Computational Research in Phonetics, Phonology, and Morphology*.

Pavlou, Menelaos, Gareth Ambler, Shaun Seaman & Rumana Z. Omar. 2015. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Medical Research Methodology* 15: 59.

Peperkamp, Sharon, Inga Vendelin & Emmanuel Dupoux. 2010. Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics 38*: 422-430.

Prince, Alan & Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Technical Report, Rutgers University and University of Colorado at Boulder, 1993. Revised version published by Blackwell, 2004.

Quené, Hugo & Huub van den Bergh. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language 59*. 413–425.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rajemisa-Raolison, Régis. 1985. *Rakibolana Malagasy*. Ambozoantany.

Raudenbush, Stephen W., & Anthony S. Bryk, 2002. *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Richardson, James. 1885. *A New Malagasy-English Dictionary.* The London Missionary Society, Antananarivo.

Rose, Sharon, & Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language* 80: 475–531.

Rosen, Eric. 2016. Predicting the unpredictable: capturing the apparent semi-regularity of rendaku voicing in Japanese through Gradient Symbolic Computation. In Clem, Emily, Geoff Bacon, Andrew Cheng, Virginia Dawson, Erik Hans Maier, Alice Shen, & Amalia Horan Skilton (eds.), *Proceedings of the 42nd Annual Meeting of the Berkeley Linguistics Society. Berkeley: Berkeley Linguistics Society.*

Rosen, Eric. 2001. *Phonological Processes Interacting with the Lexicon: Variable and Non-Regular Effects in Japanese Phonology*. PhD dissertation, University of British Columbia.

Schane, Sanford. 1968. French Phonology and Morphology. Cambridge, Massachusetts: MIT Press.

Schnoebelen, Tyler & Victor Kuperman. 2010. Using Amazon Mechanical Turk for linguistic research. *Psihologija* 4.

Selkirk, Elisabeth. 1972. The Phrase Phonology of English and French. Ph.D. dissertation, MIT. (1980, New York: Garland).

Scholes, Robert J. 1966. *Phonotactic Grammaticality*. The Hague: Mouton.

Sezer, Engin. 1981. The k/ø alternation in Turkish. In G. N. Clements (Ed.), *Harvard Studies in Phonology*, 354–382. Bloomington: IULC.

Shademan, Shabnam. 2007. Grammar and Analogy in Phonotactic Well-formedness Judgments. Ph.D. dissertation, University of California, Los Angeles.

Shih, Stephanie. 2018. Learning lexical classes from variable phonology. In *Proceedings of AJL2*.

Shih, Stephanie & Sharon Inkelas. 2016. Morphologically-conditioned tonotactics in multilevel Maximum Entropy grammar. In Hansson, Farris-Trimble, McMullin, Pulleyblank (eds). *Proceedings of the 2015 Annual Meeting on Phonology*. Washington, DC: Linguistic Society of America.

Shih, Stephanie & Kie Zuraw. 2018. Phonological conditions on variable adjective-noun word order in Tagalog. *Phonological Data and Analysis*, an online section of *Language* 94: e317-e352.

Skrondal Anders & Rabe-Hesketh, Sophia. 2009. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Statistics in Society, Series A,*172(3): 659–687.

Smith, Brian. 2015. *Phonologically-conditioned allomorphy and UR constraints*. Doctoral dissertation, University of Massachusetts, Amherst.

Smith, Brian W. & Claire Moore-Cantwell. 2017. Emergent idiosyncrasy in English comparatives. In Andrew Lamont and Katie Tetzloff, eds., *NELS 47: Proceedings of the 47th meeting of the North East linguistic Society. Amherst: Graduate Linguistic Student Association*. 127-140.

Smith, Brian W. & Joe Pater. 2017. French schwa and gradient cumulativity. Ms., UC Berkeley & UMass Amherst.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press/Bradford Books. 194–281.

Smolensky, Paul & Matthew Goldrick. 2016. Gradient symbolic representations in grammar: the case of French liaison. ROA 1286.

Smolensky, Paul & Geraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar.* Cambridge, MA: MIT Press.

Snijders, Tom & Roel Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Applied Multilevel Analysis, 2nd edition*. Sage.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavioral Research Methods 43(1)*: 155-67.

Stanton, Juliet & Sam Zukoff. 2016. Prosodic effects in segmental correspondence. *Proceedings of CLS 51.* Chicago Linguistics Society. 501-515.

Tanaka, Yu. 2017. *The sound pattern of Japanese surnames*. Doctoral dissertation, UCLA.

Tay, Kim Gaik, Kek, Sie Long, Abdul-Kahar, Rosmila. 2012. A Spreadsheet Solution of a System of Ordinary Differential Equations Using the Fourth-Order Runge-Kutta Method. *Spreadsheets in Education* (eJSiE) 5(2), 5.

Tesar, Bruce & Alan Prince. 2003. Using phonotactics to learn phonological alternations. *Proceedings of the Thirty-Ninth Conference of the Chicago Linguistics Society, Vol. II: The Panels*.

Tranel, Bernard. 1981. Concreteness in generative phonology. Evidence from French. Berkeley and Los Angeles: University of California Press.

Tranel, Bernard. 1996. French liaison and elision revisited: a unified account within Optimality Theory. In: C. Parodi, C. Quicoli, M. Saltarelli and M.L. Zubizarreta (eds.) *Aspects of Romance Linguistics*. Washington, DC : Georgetown University Press, 433-455.

Toporišič, Jože (ed.). 2001. *Slovenski pravopis*. Ljubljana: SAZU.

Trudgill, Peter. 1974. *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Walker, Douglas. 2001. French Sound Structure. Calgary: University of Calgary Press.

Walsh, S. & D. Diamond. 1995. Non-linear curve fitting using Microsoft Excel solver. *Talanta* 42(4): 561-72.

Walther, Markus & Richard Wiese. 1999. Optimization versus lexical specification. Handout from the workshop Conflicting Rules in Phonology and Syntax, University of Potsdam.

Wilson, Colin & Benjamin George. 2009. Maxent grammar tool. Software. http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/

Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC: Center for Applied Linguistics.

Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* 30.5:945-982.

White, James. 2014. Evidence for a learning bias against saltatory phonological alternations. *Cognition*, 130(1): 96–115.

Zeger, Scott L., Kung-Yee Liang & Paul S. Albert. 1998. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44(4): 1049–1060.

Zonneveld, Wim. 1978. *A Formal Theory of Exceptions in Generative Phonology*. Lisse: The Peter de Ridder Press.

Zuraw, Kie. 2000. *Patterned Exceptions in Phonology*. Doctoral dissertation, University of California, Los Angeles. ROA-788.

Zuraw, Kie. 2009. Frequency influences on rule application within and across words. *Proceedings of CLS 43*.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory 28(2)*: 417-472.

Zuraw, Kie. 2013. Class 1 handout for Linguistics 251A/B: Variation in phonology.

Zuraw, Kie. 2015. Allomorphs of French de in coordination: a reproducible study. *Linguistics Vanguard.*

Zuraw, Kie. 2016. Polarized variation. *Catalan Journal of Linguistics* 15: 145-171.

Zuraw, Kie & Hayes, Bruce. 2017. Intersecting constraint families: An argument for harmonic grammar. *Language* 93: 497-548.

Zymet, Jesse 2015. Distance-based decay in long-distance phonological processes. In U. Steindl et al. (eds.), *Proceedings of the 32nd West Coast Conference on Formal Linguistics*, pp. 72 – 81. Somerville, MA: Cascadilla Press.

Zymet, Jesse. 2018. Contradictory Markedness Preferences across Morphological Domains. In *Proceedings of the 35th West Coast Conference on Formal Linguistics*, ed. Wm. G. Bennett et al., 479-488. Somerville, MA: Cascadilla Proceedings Project.