UNIVERSITY OF CALIFORNIA
Los Angeles

Learning Hidden Structure:
Derived Environment Effects
and the Richness of the Base

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

by

Adeline R Tan

2024

ABSTRACT OF THE DISSERTATION

Learning Hidden Structure:
Derived Environment Effects
and the Richness of the Base

by

Adeline R Tan
Doctor of Philosophy in Linguistics
University of California, Los Angeles, 2024
Professor Kie Ross Zuraw, Chair

Hidden structure refers to the units of organization that a child cannot directly observe when they are learning language (*e.g.* phonemes, morpheme boundaries, URs, phrases). In this dissertation, I propose a novel computational model that learns hidden structures in-tandem with the grammar. My model consists of two Maximum Entropy sub-models that are chained via the product rule. Since I treat the hidden structure as a latent variable, the learner is free to match the observed surface pattern via different intermediary URs. Latent variable models have no guarantee of concavity, so I develop a novel sampling technique to simulate a population of language learners.

When presented with the same surface information, different human learners may arrive at different analyses (*i.e.* inter-speaker variation in analyses). In a production task with the *-ity* suffix and nonce stems, Pierrehumbert (2006) found that 2 in 10 participants never applied velar softening. This suggests that approximately 20% of learners may not learn the grammar for velar softening, but may instead memorize full underlying forms (*e.g.* /ɪlɛktɪsɪti/) for existing words. For velar softening, my model not only correctly predicts that there are multiple solutions for one surface pattern, it also correctly predicts the proportion of human speakers that will pick each solution.

In the Rich Base problem, there are two grammars that satisfy one surface pattern. However, humans only acquire one grammar – the Rich Base Grammar (as evidenced by loan word adaptation). In my simulated population of language learners, I find an overwhelming preference for the Rich Base Grammar to be learned. This preference emerges from my model's ability to leverage the superior utility of the Rich Base Grammar over its non-Rich Base counterpart (without needing to build in any extra mechanisms or biases).

English CiV lengthening appears at first blush to be a derived environment effect, whose triggering condition – the derived environment – cannot be directly observed. My experiments confirm the productivity of CiV lengthening. I reanalyze CiV lengthening as the emergence of the unmarked Stress-to-Weight Principle, thus simplifying CiV lengthening from a complex hidden structure problem to a surface true phenomenon.

The dissertation of Adeline R Tan is approved.

Timothy Hunter

Donka Minkova Stockwell

Claire Moore-Cantwell

Kie Ross Zuraw, Committee Chair

University of California, Los Angeles

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Words can scarcely portray the gratitude and admiration I hold for Kie Zuraw. Your support, wisdom, guidance and care have been constant through my journey in graduate school. From the undergraduate class in phonology I took with you to the meetings we have had while I'm dissertating, I have learned so much from you on how to analyze and think critically about phonological data and frameworks. Your creativity and ability to always see the applications of my work have helped me greatly! I am also fortunate to have served as your Teaching Assistant and Research Assistant for the maxent.ot software in R. Thank you for supporting my drive to propose, code and train a novel model from scratch!

I am immensely grateful to Tim Hunter for his incredible patience and interest in the different model architectures that I experimented with. I thank you for your steadfast belief that I would discover a model that works! Tim has always been generous with his time, reading multiple drafts of my SCiL abstract and paper with a short turnover period. Thank you for asking difficult questions, which have improved the rigor of my ambitious claims and brought clarity to my arguments!

Donka Minkova's warmth, kindness and encouragement is matched only by her impressive depth of knowledge of English phonology. Благодаря ви много!

Claire Moore-Cantwell's insightful questions have made me think hard about the deeper phonological issues, and consequently improved my writing.

I thank Ed Keenan for first opening my eyes to the mathematical structure of language as an undergraduate. Thank you for helping me to understand the difference between the reals and the naturals and the extent of the claims that I could make in Ch 6. I also thank you for your caring advice.

This dissertation has also been enriched by the following individuals: Thank you Robert Daland for introducing me to Tesar's (2003) surgery paper! Ch 5 and Ch 6 would not exist without you. During the writing process, I found myself referring back to Bruce Hayes' incredibly thorough class notes. I thank Ed Stabler for teaching me to think about learnability through a computational lens. From Megha Sundara, I have learned much on experimental design and analysis.

I thank ZL Zhou for making such high-quality recordings of the stimuli when we were all remote due to the pandemic.

When writing a dissertation in LaTeX, there's always something to fix. I'm grateful to Connor Mayer, who is always quick to share tableaux and formatting tips, and also to Kevin Yin for sharing his formatting tips to meet UCLA's dissertation requirements.

I also thank the anonymous reviewers and audiences at SCiL (2022), GLOW (2023) and WCCFL (2023) for helpful feedback and comments, which have improved the work in this dissertation.

This journey in graduate school was possible because of the potential that Ed Keenan, Yosuke Sato, Dominique Sportiche and Kie Zuraw saw in me when I was still an undergraduate. I thank each of you for the encouragement and the advising meetings we had!

It is always a pleasure to bump into Sun-Ah around the department. I have enjoyed the syntax classes I took with Hilda Koopman, Anoop Mahajan, Dominique Sportiche and Tim Stowell. Thank you for the inspiring and stimulating classes!

With fondness, I remember the late Russ Schuh and his lovely wife Maxine. I am honored to have worked closely with him.

In the final months of writing this dissertation, I have been refreshed by the dances, meals, friendship and laughter I have enjoyed with my friends at DCU.

Robin Roberts has been a pillar of support. Thank you for seeing my strengths and believing in me!

Above all, I am exceedingly grateful to my parents for their unconditional love and unceasing support!

VITA

2013 B.A. (English Language), National University of Singapore.

PUBLICATIONS

Mayer, Connor; Adeline Tan; and Kie Zuraw. accepted. Introducing maxent.ot: an R package for Maximum Entropy constraint grammars. *Phonological Data and Analysis.*

Tan, Adeline. in press. CiV Lengthening: Productivity and the Emergence of the Unmarked. *Proceedings of the 41st West Coast Conference on Formal Linguistics.*

Tan, Adeline. 2022. Concurrent hidden structure & grammar learning. *Proceedings of the Society for Computation in Linguistics (SCiL), 5*(5), 55-64.

CHAPTER 1

Introduction

When a child encounters a human language for the first time, they are faced with a continuous sound stream. One of the tasks involved in learning the language is segmenting the sound stream into smaller units – sentences, phrases, words, phonemes, *etc.* These smaller units form a hierarchy of representations along which language is organized. In order to produce language, the child needs to learn the contents of these representations as well as the mappings between each level of representation within the representational hierarchy. Let's imagine that a child would like to communicate the plural of goat. The task is to get from the WORD GOATS to the surface representation (SR) [goʊts]. I will treat the surface representation as a proxy for the pronounced form. In order to go from the WORD GOATS to the sequence of sounds that communicates the idea GOATS, the child would need to learn the following:

(1) a. The underlying representation (UR) is /goʊt+z/, which is obtained by concatenating the URs of the morpheme for GOAT and the plural morpheme in the correct order.

b. The surface representation (SR) is [goʊts].

c. The mapping from the WORD to the UR.

d. The mapping from the UR to the SR.

The WORD represents the concept(s) to be communicated while the surface representation represents actual produced speech. Hence, both of these levels of representation are known. In contrast, the underlying representation is unobserved, and thus can be called hidden structure. One of the goals of this dissertation is to model how the unseen hidden structure is learned. To simplify the learning problem, I will assume in this dissertation that the word boundaries have already been learned. That is to say, the child has already segmented the speech stream into words, so the

learning problem is to find the best mapping(s) from a WORD to its surface representation via appropriate underlying representations (*e.g.* GOATS → ??? → [goʊts]). Here, the characters ??? symbolize the hidden structure(s).

It is perhaps easier to illustrate what goes into an underlying representation when considering what a child needs to learn in order to learn the underlying representation for a WORD.

(2) a. How to build the UR of a WORD?

   i. *e.g.* Rregarding the WORD GOATS, should the child use the unanalyzed UR /goʊts/, which is perhaps stored as a memorized form? Or should she instead compose the UR from the UR of GOAT /goʊt/ and the UR of the plural morpheme /-z/, which when concatenated together in a particular order creates the UR /gout+z/?

   b. What are the segments of the underlying representation of a morpheme?

   i. *e.g.* Is the plural morpheme /-s/ or /-z/?

   c. What are the suprasegmentals such as stress, length or tone (as relevant) for each syllable of a morpheme.

   i. *e.g.* Is it /ˈælɪˌgeɪtɚ/ or /ˌælɪˈgeɪtɚ/ for the word ALLIGATOR?

   d. Decisions regarding organizational units as moras and feet also constitute hidden structures (although none of the modeling work in this dissertation will address these types of structures).

1.1 Phonological models: A quick overview

A great proportion of phonological modeling studies assume a known UR and focus the modeling task solely on the mapping between the UR and the SR, with recent work on the UR-SR mapping being dominated by constraint-based models such as Optimality Theory (Prince and Smolensky, 1993/2004) and Harmonic Grammar (Legendre *et al.* (1990), Legendre *et al.* (2006), Potts *et al.* (2010), *a.o.*). While less commonly seen, a good number of papers have modeled the learning of hidden structure within a constraint-based framework. Previous work in this vein include (Akers

(2012), Alderete *et al.* (2005), Boersma and Pater (2016), Jarosz (2006), Jarosz (2017), Jarosz (2015), Merchant (2008), Nazarov and Jarosz (2017), Nelson (2019), Tesar and Smolensky (2000), Tesar *et al.* (2003), Tesar (2004), Tesar and Prince (2007)).

1.1.1 The MaxEnt Grammar

Maximum Entropy Grammar (MaxEnt; Smolensky (1986), Goldwater and Johnson (2003)) belongs to the family of Harmonic Grammars (Pater, 2009). Of the variants of Harmonic Grammar, MaxEnt has recently gained ground as a popular model for the UR-SR mapping. While the connection between MaxEnt and prior OT models had been explored in earlier works (Eisner (2000), Johnson (2002)), the seminal work describing the MaxEnt model and its advantages over other competing models is Goldwater and Johnson (2003). It is Goldwater and Johnson (2003)'s version of MaxEnt that I present in the equations below.

$$Pr(y|x) = \frac{1}{Z(x)} exp(\sum_{i=1}^{m} w_i f_i(y, x)) \tag{1.1}$$

In Equation 1.1, the symbol x represent a single UR, and the symbol y represents a single SR. The symbol $\mathcal{Y}(x)$ represents the set of SR candidates that arise of the UR x. Features are used to parameterize the mapping between the UR and the SR. For example, the feature, f_{voice} could be used to track whether the voicing of any sounds change between the UR and its associated SR. The UR-SR pair (/gout+z/, [gouts]) would violate f_{voice} one time because the final voiced consonant in the UR /gout+z/ becomes voiceless in the SR [gouts]. The UR-SR pair (/gout+z/, [kouts]) would violate f_{voice} twice because the initial and final voiced consonants in the UR /gout+z/ become voiceless in the SR [kouts]. Thus, the expression f_{voice}([gʊʊts], /gʊʊt+z/) has a value of 1 while the expression f_{voice}([kʊʊts], /gʊʊt+z/) has a value of 2. In Eq 1.1, there are m features that parameterize the UR-SR mapping, and w_i represents the weight of the i^{th} feature, f_i. In phonological parlance, features that parameterize the UR-SR mapping are called phonological constraints. Hence, the expression $\sum_{i=1}^{m} w_i f_i(y, x)$ represents the weighted sum of the UR-SR pair (x, y)'s constraint violations. Within phonology, the weighted sum of constraint violations of a UR-SR pair is commonly called its harmony score while the exponential of the UR-SR pair's

harmony score is often called its MaxEnt score of MaxEnt value. The term $Z(x)$ serves as a normalizing constant:

$$Z(x) = \sum_{y \in \mathcal{Y}(x)} exp(\sum_{i=1}^{m} w_i f_i(y, x)) \tag{1.2}$$

$Z(x)$ is a function of the UR x. It is calculated by summing the MaxEnt values of all SRs, $\mathcal{Y}(x)$, arising out of the UR x. Hence, a MaxEnt model produces a conditional distribution, $P(y|x)$, over the set of SRs, $\mathcal{Y}(x)$, for a given UR x.

The MaxEnt model stands out amongst competing models of Harmonic Grammar because the MaxEnt model is essentially a log-linear model. Another way to put it, the log probability of the SR y given the UR x, (*i.e.* $ln(Pr(y|x))$) is proportional to the weighted sum of phonological constraint violations, $\sum_{i=1}^{m} w_i f_i(y, x)$. Being a log-linear model brings several advantages. First, the log-linear model is a general statistical model, so its mathematical properties have been well-studied and understood. Second, multiple optimization algorithms that have proofs of convergence are available for log-linear models (Goldwater and Johnson, 2003). Such algorithms include the Conjugate Gradient algorithm (Press *et al.*, 1992) and the iterative scaling algorithms (Berger *et al.*, 1996). These algorithms are used to find the parameter weights (*i.e.* phonological constraint weights) that maximize the likelihood. In practice, the log likelihood is usually maximized in place of the likelihood. Since the logarithm function is monotonic, maximizing the log likelihood is functionally equivalent to maximizing the likelihood, while having the benefit of delaying (or altogether avoiding) the numerical underflow problems of the latter. In phonology, the MaxEnt model produces a distribution over SRs for each UR. In other words, the probabilities that phonologists are usually concerned with are *conditional* probabilities (*e.g.* $Pr(SR|UR)$). Hence, the function to be optimized is the log conditional likelihood (rather than the likelihood). Importantly, this function has been shown to be concave (Berger *et al.*, 1996). Thus, a phonological MaxEnt model has the property that all standard optimization algorithms will converge to the same trained model (*i.e.* the best weights that are found will be the same no matter the algorithm used for optimization) since there is only one global maximum[1]. MaxEnt also has the nice property that the best model is the one that makes no further assumptions about unknown data (*i.e.* remaining maximally noncommittal about unknown data) while matching the known data as closely as possible. In other

words, a trained MaxEnt model produces the least biased distribution possible – in the absence of information, no assumptions are to be made about the unknown data.

A quick aside, to clear up some terminology and to be explicit about what the 'known data' in phonology is: In phonology, the known data is the observed rates of constraint violation profiles. The term 'constraint' in phonology is basically equivalent to the term 'feature' in machine learning. The term 'constraint violation profile' is equivalent to a 'feature vector'.

1.1.2 MaxEnt-ized models for hidden structure

As previously mentioned, multiple models have been proposed for the learning of hidden structures. Several of these proposed models incorporate a MaxEnt sub-model for the UR-SR mapping (Eisenstat (2009), Pater *et al.* (2012), Staubs and Pater (2016), Nazarov and Pater (2017) and O'Hara (2017)). Of these models, the vast majority have an overarching architecture in which multiple MaxEnt sub-models are chained together (Eisenstat (2009), Pater *et al.* (2012), Staubs and Pater (2016) and Nazarov and Pater (2017)).

In Eisenstat (2009) and Pater *et al.* (2012), one of the MaxEnt sub-models is used to model the WORD-UR mapping. This WORD-UR MaxEnt sub-model is parameterized by UR constraints (Zuraw, 2000; Boersma, 2001), and produces a conditional probability distribution over URs for each WORD. The model that I propose in this dissertation (Chapter 3) follows in this tradition by employing a WORD-UR MaxEnt sub-model that is similarly parameterized by UR constraints. The WORD-UR sub-model that my overarching model employs departs from these predecessor models because it produces a joint probability distribution over WORD-UR pairs. Thus, my overarching model produces a joint probability distribution over WORD-UR-SR triples, which enables it to model differences in WORD frequencies (in addition to UR frequencies and SR frequencies). In contrast, prior models produce conditional probability distributions over UR-SR pairs for each WORD, which allow them to model only the differences in SR frequencies and UR frequencies (but not WORD frequencies). O'Hara's (2017) model differs from all previous models by directly

[1]This assumes a sharp peak. In the case where the peak is flat, for each constraint, a range of weights (rather than a singular weight) will all produce models with equally good "best" fits to the data.

adjusting the frequencies of the unobserved URs rather than by parameterizing the WORD-UR mapping with UR constraints.

In the models discussed so far, the nature of the hidden structure mostly falls out from the WORD-UR mapping (*e.g.* the ways in which a WORD is broken down into morpheme(s), the UR(s) of a morpheme). Nevertheless, hidden structure has also been posited to exist between the UR and the SR levels of representation. Phonological frameworks which posit such hidden levels of representation include Harmonic Serialism (McCarthy (2000), McCarthy (2007)) and Stratal OT (Kiparsky (2000)). MaxEnt-ized models of Harmonic Serialism and Stratal OT have been proposed by Staubs and Pater (2016) and Nazarov and Pater (2017). These models essentially consist of chains of "UR-SR" sub-models because the only relevant parameters are those that characterize the UR-SR mapping.

1.2 Phonological phenomena

The combined learning of hidden structure and grammar is especially important for phonological phenomena in which the identity[2] of the hidden structure plays a crucial role in determining the surface forms. Such phonological phenomena, whose in tandem learning of hidden URs and grammars have been modeled, include the lexically-conditioned variation observed in French vowel deletion (Pater *et al.*, 2012) and abstract URs in Klamath (O'Hara, 2017). Nazarov and Pater (2017) have also shown that such models are useful in modeling opaque interactions (*e.g.* French tense-lax vowel alternations, the interaction between diphthong raising and flapping in Canadian English), in which "hidden" intermediary levels of representation between the UR and the SR have been posited by analysts.

Whilst the modeling work on the concurrent learning of hidden structure and grammars has covered several types of phonological phenomena, two families of phonological phenomena have not yet been modeled: Derived Environment Effects and the Richness of the Base. In the following two sections (§1.2.1 and §1.2.2) I introduce these two phenomena respectively, and explain

[2]Or more generally, the distribution over hidden structures.

why they are especially pertinent to the concurrent learning of hidden structure and the mapping between the hidden structure and the SR.

1.2.1 Derived Environment Effects

Derived environment effects present an especially challenging learning problem because the application of a phonological process is dependent on the derived status of the hidden structure. For example, English velar softening (/k/ → [s]) has been described as a derived environment effect because it only applies in a very specific morphologically derived context: underlying "hidden" /k/ softens to "observed" surface [s] in the morphologically complex derived word ELECTRIC+ITY, but not in the non-derived word KITTY.

In constraint-based grammars, a phonological process is motivated by a (hard or soft) ban against certain disfavored surface strings. For English velar softening, it is the surface [kɪ] string that is disfavored. If we take derived environment effects at face value, then in order to prevent underlying derived /kɪ/ strings from surfacing, a phonological process called velar softening must be assumed to apply to underlying derived /kɪ/ strings, turning them into surface [sɪ] strings. However, this process should not apply to underlying simple /kɪ/. This presents a challenge for a child learning the language.

In order to master a language with a derived environment effect, a child would need to learn the derived status of words as well as the phonological process. Taking English as an example, if the child were born with full knowledge of the derived status of words in English, it would be easy for her to deduce the Velar Softening (*i.e.* /k/ → [s]) process. However, when children first encounter language, they are unaware of the derived status of words, which essentially constitute a type of hidden structure. The child is thus confronted with a confusing puzzle where the /k/ → [s] phonological process occurs in some words but not in others. In the absence of information about the derived status of words, the best that a child can do observe that on the surface, the [kɪ] sequence occurs less than expected, and posit a Velar Softening process that applies at a frequency less than 100% (since surface[kɪ] sequences are not outright banned in English as in the word KITTY). Over time, the child's vocabulary would increase and we'd expect that her larger

vocabulary would enable her to see the relationships between morphologically related words like ELECTRIC∼ELECTRICITY, which would in turn help her to learn Velar Softening. Even here, however, the relationship between ELECTRIC and ELECTRICITY is itself obscured by Velar Softening. For example, the relationship between *den*[s]*e* and *den*[s]*ity*, where the stem-final consonant remains the same, is clearer than the relationship between *electri*[k] and *electri*[s]*ity*. In other words, knowledge of morphology (*i.e.* derived status) would aid in the learning of the derived environment phonological process while knowledge of the derived environment phonological process would aid in the learning of morphology. To put it another way, derived environment effects present a chicken-and-egg problem because prior knowledge of the hidden morphological structure would aid in the learning of the grammar while prior knowledge of the grammar would aid in the learning of the hidden morphological structure.

One question we can ask is whether the above characterization of derived environment effects is accurate. Many cases of *morphological* derived environment effects (e.g. Turkish *k*-deletion, Javanese *h*-deletion, Estonian spirantization) are reportedly triggered only by a *subset of exceptional suffixes that do not form a phonological natural class*. This means that it is not enough to learn just the derived status of a word. In order to master the derived environment effect, the child must also learn the specific affixes present in words as well as the special set of affixes that trigger the derived environment effect.

One approach to derived environment effects is to reduce as many cases of morphological derived environment effects as possible to phonological ones. Lubowicz's (2002) theory of conjoined constraints was the most extreme version of this approach – affixation always resulted in a phonological change, and it was the phonological change itself (rather than any morphological factor) that triggered the derived environment effect phonological process. Chong (2019) took a more nuanced approach, treating only the cases where all the suffixes triggered the derived environment effect as phonological derived environment effects. He further hypothesized that phonological derived environment effects received general phonotactic support in the lexicon. Following this line of reasoning, it is the morpheme-internal sequences that lacked phonotactic support and were thus exceptional. Chong argued that such cases were better understood as cases of exceptional non-undergoers, where a language-wide general markedness constraint banned the marked sequence

while the morpheme-internal marked sequences were protected by specific indexed faithfulness constraints. Chong's study of derived environment effects illustrated the utility of careful case studies that elucidated the differences between different types of derived environment effect. Importantly, lines of evidence from experimental and corpus data could be used to show that derived environment effects were not a singular monolithic concept.

In my dissertation, I investigate two cases of derived environment effects: Velar Softening and CiV Lengthening. Velar Softening appears to be a morphological derived environment effect because when looked at in greater detail, the phonological /k/ → [s] process turns on the identity of the suffix. For example, the *-ity* /-ɪti/ suffix triggers Velar Softening, but the *-ish* /-ɪʃ/ suffix does not. Since both suffixes produce identical phonological environments as they both begin with the high front lax vowel /ɪ/, the triggering condition cannot be reduced to a phonological factor, and the identity of the suffix must be taken into account. In contrast, the morphological character of CiV Lengthening is in question. At first blush, CiV Lengthening appears to be a derived environment effect because (1) a lax vowel cannot appear in the __CiV context when a morpheme boundary intervenes, so a tensing process occurs (*e.g.* the lax [ə] in *Can*[ə]*da* corresponds to the tense [eɪ] in *Can*[eɪ]*dian*); but (2) a lax vowel can surface when there is no morpheme boundary (*e.g.* the lax [æ] in *c*[æ]*meo*). However, recent work by Steriade (2019) challenges the derived environment status of CiV Lengthening, and reframes CiV Lengthening as driven by mainly phonological rather than morphological factors.

To recap, in derived environment effects we have an interesting case of hidden structure (here: the morphological structure of words) informing the phonological process and the phonological process likewise informing hidden structure. This reciprocal chicken-and-egg relationship appears especially suited to a model that does not assume hidden structure in advance but instead allows the learning of the hidden structure to proceed alongside the learning of the phonological process.

Before modeling any phonological phenomena, it is useful to have data against which to evaluate the modeling results. Since I am modeling the concurrent acquisition of hidden structure and phonological processes, the data against which I will evaluate my modeling results would ideally be results from humans. More specifically, I will compare the generalization properties of my trained models against how native speakers of English generalize the phonological processes

in question to novel test items. To my knowledge, no experimental results were available for CiV Lengthening, so I designed and performed an experiment to test the generalizability of CiV Lengthening (Chapter 2). These experimental results indicated that CiV Lengthening was driven by phonological factors. I then propose a novel model of CiV Lengthening in which it is analyzed much more simply as a case of the emergence of the unmarked. I thus show that CiV Lengthening is driven entirely by phonological factors, which reduces the complexity of the parameters needed to model CiV Lengthening.

For Velar Softening, human generalization results were available from a production study conducted by Pierrehumbert (2006). I designed and trained 200 randomly initialized models for Velar Softening (Chapter 4). 27 of these trained models produced equally good "best" fits to the training data. These 27 models were then subject to a generalization task, and found to generalize in a way that mimicked human behavior.

1.2.2 Richness of the Base

One of the differences between constraint-based theories versus previous rule-based ones is that Morpheme Structure Constraints are no longer present in constraint-based theories (Kager (1999)). Morpheme Structure Constraints operated solely on the UR, and were used to prohibit certain structures from the UR (Chomsky and Halle (1968)). For example, a Morpheme Structure Constraint might be used to prohibit front rounded vowels from the URs of English (these vowels are not part of the phonemic inventory of English).

One of the main ideas in constraint-based theories is that all predictable patterns that are productive must be enforced by the grammar (Prince and Smolensky (1993/2004)). The grammar refers to the mapping between the UR (hidden) and the SR (observed). In Optimality Theory, the grammar consists of ranked constraints while in probabilistic constraint-based models such as MaxEnt, the grammar consists of a set of constraint weights. If English lacks front rounded vowels, then according to constraint-based theories, this cannot come about from the URs of morphemes; instead, the prohibition against front rounded vowels must arise from mapping between the UR and the SR. In other words, the constraint weights (or constraint ranking in the case of OT) must

be able to convert any illegal underlying front rounded vowels in the UR to some legal sound (or sound sequence) in the SR.

The removal of Morpheme Structure Constraints meant that the Duplication Problem was no longer an issue. With the inclusion of Morpheme Structure Constraints, the prohibition of illegal structures (such as the aforementioned front rounded vowels in English) would be dealt with both at an earlier stage (UR) and at the mapping between the UR and the SR. If a particular ban against illegal structures is known to be productive (*e.g.* front rounded vowels are disallowed even in English loanwords), then we know that the illegal structures are dealt with at the level of the grammar. Thus, the removal of Morpheme Structure Constraints from constraint-based theories meant that the work of prohibiting illegal structures was now placed entirely on a single level – the grammar, which we know is needed to deal with illegal structures in loans. This in turn eliminated the Duplication Problem.

One consequence of the removal of Morpheme Structure Constraints was that the grammar was now entirely responsible for the shape of SRs. Another way to put it is that the grammar should be able to turn any UR into a legal sequence of sounds in the SR. This principle is called the Richness of the Base, where the term "rich" refers to an unrestricted UR. That is to say, there should be no restrictions on the UR.

One of the curiosities regarding the Richness of the Base is how the grammar is learned. In the absence of conflicting information, the UR is generally treated as identical to the SR. Let's take the English front rounded vowels as an example. Since there are no existing English words with front rounded vowels, these vowels do not exist in the URs of any extant English morphemes. Thus, such vowels are entirely absent from URs in English. Consequently, the grammar is under no pressure to learn any constraint weights that would modify illegal front rounded vowels into legal sounds since it never encounters front rounded vowels. Yet, such a grammar must have been acquired somehow because the illegal front rounded vowels are repaired when loanwords featuring such vowels are borrowed into English.

The general strategy used by some analysts is to imagine a hypothetical input that has the illegal sound/sound sequence. This hypothetical input would then force the grammar to learn a constraint

weight setting that would repair the illegal sound/sound sequence into something that is legal at the SR. Yet, the status of this hypothetical input in unclear – it is not an actual UR, so do language learners actually consider it?

My proposed model and the learning algorithm I used in this dissertation give equal priority to learning the UR at the same time as learning the grammar. The learner is able to continually adjust the parameter (constraint) weights for both the UR and the grammar until it comes to a satisfying solution. This presents an opportunity to observe whether a Rich Base grammar, which is able to deal with illegal sound/sound sequences, can be acquired when no "helpful" illegal URs are stipulated in advance.

1.3 Roadmap

The rest of this dissertation is laid out as follows:

While CiV Lengthening has been described in multiple studies (Chomsky and Halle (1968), Burzio (2005), Baković (2013) Steriade (2019)), no experimental study has yet been conducted to ascertain its productivity. Chapter 2 presents an experiment that I performed to determine the productivity of CiV Lengthening (§2.2). It also includes a follow-up experiment that explores the triggering environment for CiV Lengthening (§2.3). The chapter closes with a re-analysis of CiV Lengthening that recasts the phenomenon as the emergence of the unmarked Stress-to-Weight Principle. In doing so, I show that CiV Lengthening can be reduced to a simple interaction of phonological constraints, and does not need to be modeled with the complex constraint machinery (Lubowicz (2002), McCarthy (2003)) needed for derived environment effects.

In Chapter 3, I introduce my hidden structure model. My model consists of one MaxEnt WORD-UR sub-model and one MaxEnt UR-SR sub-model that are chained together via the product rule. I employ the Expectation-Maximization algorithm (Dempster *et al.*, 1977) to find the parameter weights that produce the best fit to the data. Two English Voicing Assimilation data sets serve as the running examples in this chapter. My overarching model, which produces a joint probability distribution over WORD-UR-SR triples, differs from previous models in the literature because my model is generative while previous models are discriminative. The second data set is

specifically introduced to show how the model generalizes.

In Chapter 4, I model English Velar Softening. I filter the resultant trained models to collect the set of best trained models (as indicated by the models' log-likelihoods). This set of best trained models functions as a random sample of the global maxima in the solution space. I treat this random sample as a simulated population of language learners. The generalization properties of these models as a group are compared against human generalization behavior on novel test items, and are observed to show a close match.

In Chapter 5, I explore the learning of Rich Base grammars when URs are not stipulated in advance. The task of the learner is to learn two schematic stress languages. For each of these languages, it is possible to match the surface training pattern via a Rich Base grammar or a non-Rich Base one. It turns out that there is an overwhelming preference for the Rich Base grammar to be learned. This preference is maintained even for trained models that eventually conclude their learning journey around the solution space by landing on a set of parameter weights that correspond to a non-helpful UR. The generalization properties of the acquired Rich Base grammars were confirmed by testing the ability of the trained models to generalize predictable stress to test items that were designed to resemble loan words.

In Chapter 6, I show that the preference towards acquiring a Rich Base Grammar is present at every single point of my model's solution space. Rich Base Grammars are more useful than non-Rich Base ones because the former can handle a larger set of URs. My over-arching model, which jointly learns the hidden structure and the grammar, naturally leverages the greater utility of Rich Base Grammars to cause the gradients at every point of the solution space to point towards learning a Rich Base Grammar without the need to build in any extra mechanisms.

In the latter portion of Chapter 6, I discuss two scenarios that commonly arise in phonological modeling. The first scenario involves the inclusion of unobserved SR candidates. The second senario is the over-parameterization that arises form the conversion of one OT constraint ranking into two MaxEnt constraints whose weights can be freely adjusted. I create two toy cases to illustrate these scenarios. The first case features over-parameterization, and produces a solution space with an infinite number of global maxima (*i.e.* the ridge). The second case features both

over-parameterization and unobserved SR candidates, and produces a solution curve without any global maxima (*i.e.* the asymptote). I explain why an iterative method of optimization is compatible with the solution curve that lacks a global maximum. I then refine Goldwater and Johnson's generalization to make room for these two solution curve shapes.

Chapter 7 concludes.

CHAPTER 2

CiV Lengthening

2.1 Introduction

In English CiV Lengthening, non-high lax vowels are reportedly not allowed in the _CiV environment when a morpheme boundary intervenes (Chomsky and Halle, 1968; Baković, 2013). For example, *Caˈn[æ]d-ian with the lax [æ] is ill-formed while Caˈn[eɪ]d-ian with the tense [eɪ] is grammatical. This restriction against lax vowels does not apply to non-derived words. For instance, monomorphemic ˈc[æ]meo is grammatical because no morpheme boundary occurs within the [æ]CiV sequence.

At first blush, CiV Lengthening appears to be a derived environment effect because it is blocked from applying in non-derived environments like *cameo*. However, CiV Lengthening is subject to an additional blocking effect. Burzio (2005) observes that CiV Lengthening is similarly blocked in derived environments when the target vowel is not newly stressed. Consider the stem *Orwell* (ˈOrˌw[ɛ]ll), which retains the lax [ɛ] upon *-ian* affixation: ˌOrˈw[ɛ]ll-ian, *ˌOrˈw[i]ll-ian. This contrasts with the stem *Canada* (ˈCan[ə]da), which disallows the lax [æ] in the very same derived environment: Caˈn[eɪ]d-ian, *Caˈn[æ]d-ian.

The two reported generalizations regarding the applicability of CiV lengthening in derived environments are summarized in (3):

(3) a. CiV Lengthening applies to a vowel upon affixation with a CiV Lengthening suffix.

b. except that CiV Lengthening is blocked from applying to a non-newly stressed vowel upon affixation with the same CiV Lengthening suffix.

It should be noted that *-ian* is not the only CiV Lengthening suffix. Other CiV Lengthening suffixes

include *-ious* (*e.g.* mel[ə]dy: mel[oʊ]d-ious, *mel[ɑ]d-ious) and *-ial* (*e.g.* cust[ə]dy: cust[oʊ]d-ial, *cust[ɑ]d-ial). The relevant tense-lax pairs for CiV Lengthening are the very same ones observed for Trisyllabic Shortening (TSS). Take for instance the TSS pair [oʊ]∼[ɑ][1]. CiV Lengthening has applied to mel[oʊ]d-ious. Yet, under different phonological conditions, the same stem shows up with the lax [ɑ]: mel[ɑ]dic.

The rest of this chapter is organized as follows. In §2.2, I report experimental results that show that the two major generalizations regarding CiV Lengthening in (3) are productive (Experiment 1). In §2.3, I report experimental results that indicate tentative support for having just one C in the _CiV environment (Experiment 2). A summary of experimental findings is presented in §2.4.

Since the two major CiV Lengthening generalizations that were set out in (3) are productive, I refine the generalization in (3b) from one of blocking to that of a retention effect. In the case of CiV Lengthening, the retention effect creates a challenge to the traditional derived environment effect story. Since the retention effect is synchronically active, I propose an alternative analysis of CiV Lengthening, in which the application of CiV Lengthening in (3a) is treated not as a case of a derived environment effect, but rather as a case of the emergence of an unmarked preference for stressed vowels to be heavy. In short, when affixation with a CiV Lengthening suffix moves stress onto a [ə], the vowel must change, and it prefers to become tense in order to satisfy the Weight-to-Stress Principle. In contrast, full vowels by virtue of being able to bear stress in English, can remain unchanged. This analysis is presented in §2.5.1. The chapter closes with a discussion of alternative analyses of CiV Lengthening in §2.5.2.

2.2 Experiment 1

The two generalizations to be tested are shown in (3). In order to generate the relevant experimental conditions, it is necessary to take a closer look at each of these generalizations.

Let us first turn our attention to the generalization in (3a), which states that "CiV Lengthening applies to a newly-stressed vowel upon affixation with a CiV Lengthening suffix". In order for

[1] *e.g.* prov[oʊ]k, prov[ɑ]cative.

the target vowel to become newly-stressed in the affixed form, it must have been unstressed in the unaffixed form. In English, this means that the vowel must have been a reduced vowel in the unaffixed form. Since CiV Lengthening has been reported to be restricted to the **non-high** vowels, I will restrict the set of reduced vowels to [ə]. In other words, all unaffixed forms relevant to the generalization in (3a) can only bear [ə] as their target vowel. Examples of such unaffixed test stems include [ˈkudəb] and [ˈziləv][2]. This thus forms the first of the experimental conditions: the ə-stems. In order for the generalization in (3a) to be found productive, the ə-stems must show a preference for a tense vowel upon affixation with a CiV Lengthening suffix.

The generalization in (3b) states that "CiV Lengthening is blocked from applying to a non-newly stressed vowel upon affixation with a CiV Lengthening suffix". In this case, the target vowel must have borne some level of stress (primary or secondary) in the unaffixed form. In other words, the target vowel must have been a "full" vowel (*i.e.* not a reduced vowel) even in the unaffixed form. We have already seen that full lax vowels retain their laxness upon affixation (*e.g.* ˈOr͵w[ɛ]ll: ͵Orˈw[ɛ]ll-ian, *͵Orˈw[i]ll-ian). While generally not discussed for CiV Lengthening, it can be seen that full tense vowels retain their tenseness upon affixation too (*e.g.* ͵Uˈkr[eɪ]n: ͵Uˈkr[eɪ]n-ian, *͵Uˈkr[ɛ]n-ian). Thus, an alternative way to characterize the blocking effect is to regard it as a retention effect. That is to say, there is pressure for the tenseness value of full vowels in the unaffixed form to be retained upon affixation with a CiV Lengthening suffix. To test for a retention effect, two additional conditions are necessary: the lax-stem condition & the tense-stem condition. Examples of unaffixed lax-stems include [səˈdæl] and [ˈtɪ͵dɑn]. Examples of unaffixed tense-stems include [pəˈboʊk] and [ˈzu͵seɪm]. For the retention effect to be productive, the tense-stems should exhibit a greater preference for a tense vowel upon affixation than their lax-stem counterparts do.

The generalizations in (3) have thus been operationalized to the following hypotheses (4):

(4) a. ə-stems prefer a tense vowel upon affixation with a CiV Lengthening suffix.

[2]Upon affixation with the CiV Lengthening suffix, the target vowel [ə] would have to become a non-reduced vowel. Looking ahead, the vowel choices available in this experimental task will be [eɪ, æ, oʊ, ɑ], as these are the trisyllabic shortening pairs that do not include any high vowels. (CiV Lengthening has been reported to not apply to high vowels.)

b. Upon affixation with a CiV Lengthening suffix, tense-stems prefer a tense vowel at a higher rate than their lax-stem counterparts.

2.2.1 Method

I employed the two-alternative forced choice (2AFC) method to determine the tenseness preference of *-ian* affixed forms for the three tenseness conditions: ə-stem, tense-stem, and lax-stem.

2.2.2 Stimuli

2.2.2.1 Conditions

Stems were of the shape $C_1V_1C_2V_2C_3$ (*i.e.* all stems were disyllabic). The target vowel was V_2, V_2 being the vowel whose tenseness might change upon affixation. All experimental conditions were reflected only on V_2. All other segments were irrelevant to experimental manipulation.

For test stems, the target vowel, V_2, was subject to the following manipulations. There were three tenseness conditions: ə-stems, tense-stems, and lax-stems. Tenseness was crossed with backness such that within each of the tenseness conditions, half of the stems had a back target vowel and half had a front target vowel. For ə-stems, the backness condition was visible on the affixed options (but not on the unaffixed stem).

To illustrate, while the ə-stems [ˈkudəb] and [ˈziləv] belonged to the front and back conditions respectively (Table 2.1), their backness condition was indistinguishable in the unaffixed form. Their membership in their respective front and back conditions only became clear upon encountering the affixed options. For example, [ˈkudəb], belonging to the front condition, had the affixed options [ˌkuˈdeɪbiən] and [ˌkuˈdæbiən], both of which had target vowels that were front [eɪ, æ]. [ˈziləv] belonged to the back condition, so it had the affixed options [ˌziˈloʊviən] and [ˌziˈlɑviən], which both featured the back vowels [oʊ, ɑ].

When unaffixed, the V_2's of ə-stems were necessarily unstressed. In contrast, the V_2's of tense-stems and lax-stems had to bear stress. There were two stress-levels such that half of the V_2's of tense-stems had primary stress and the other half had secondary stress. Stress-level was similarly

crossed with the lax-stem condition. This resulted in a total of 96 test stems. Table 2.1 summarizes the distribution of stems[3] for each experimental condition.

		Backness	
Tenseness	***Stress***	*Front*	*Back*
ə-stem	*None*	16 [ˈkudəb]	16 [ˈziləv]
Tense-stem	*Primary*	8 [jəˈteɪk]	8 [pəˈboʊk]
	Secondary	8 [ˈzuˌseɪm]	8 [ˈtæˌsoʊp]
Lax-stem	*Primary*	8 [səˈdæl]	8 [ɹəˈlɑn]
	Secondary	8 [ˈnɪˌmæb]	8 [ˈtɪˌdɑn]

Table 2.1: *Distribution of test stems across conditions (Expt1).*

2.2.2.2 Materials

CiV Lengthening was reported to be restricted to the non-high vowels. Hence, I decided to restrict the vowel pairs to the non-high vowels. The vowel pairs for the front and back conditions were [eɪ]∼[æ] and [oʊ]∼[ɑ] respectively[4].

I wrote a script to randomly generate the rest of the test stems *(i.e.* C_1, V_1, C_2, and C_3). The set of onsets (C_1 & C_2) were [p, t, k, b, d, ɡ, m, n, f, θ, s, ʃ, h, v, ð, z, ʒ, t͡ʃ, d͡ʒ, ɹ, l, j, w]. The set of codas (C_3) were [p, k, b, ɡ, m, n, f, θ, ʃ, v, ð, ʒ, l][5]. The set of vowels were [i, ɪ, eɪ, ɛ, æ, ɑ, ʌ, oʊ, ʊ, u, aɪ, aʊ, ɔɪ]. The frequency at which the script picked a sound depended on the sound's unigram

[3]Other than the target vowel, all other segments were randomly generated.

[4]The three remaining trisyllabic shortening pairs are [i]∼[ɛ] (*e.g. serene∼serenity*), [aɪ]∼[ɪ] (*e.g. divine∼divinity*), and [aʊ]∼[ʌ] (*e.g. profound∼profundity*). The [aɪ]∼[ɪ] and [aʊ]∼[ʌ] pairs were excluded because CiV Lengthening does not apply when its application would result in the [aɪ] and [aʊ] diphthongs. The [i]∼[ɛ] pair was excluded in order to have a balanced data set whose front target vowels had height features that mirrored their back target vowel counterparts exactly.

frequency in the CMU Pronouncing Dictionary. Stems and affixed options that sounded like real words were rejected.

The 2AFC affixed options were generated by modifying the test stems in the following ways (5):

(5) a. The *-ian* [ɪən] suffix was added.
    i. All affixed options carried primary stress on the target vowel.
  b. Target vowels were modified such that for each 2AFC pair, one option would carry a tense target vowel while the alternative option would carry its lax counterpart.
    i. For ə-stems, half of the stems showed the front pair in the affixed forms while the other half showed the back pair in their affixed forms.

For example, the ə-stem [ˈkudəb] would have these two affixed options: [ˌkuˈdeɪbiən], [ˌkuˈdæbiən]. Since the target vowels in the affixed forms are the front vowels [eɪ] and [æ], this particular ə-stem is classified as belonging to the front group. The lax-stem [ˈtɪˌdɑn], which has the two affixed choices: [ˌtɪˈdoʊniən], [ˌtɪˈdɑniən], is similarly classified as belonging to the back group. The tense-stem [jəˈteɪk], which clearly belongs to the front group since its target vowel in the unaffixed form is already the front vowel [eɪ], would have the affixed options: [jəˈteɪkiən], [jəˈtækiən].

Let us now turn our attention from the test stems to the filler stems. There were 48 filler stems. The target vowels for the unaffixed filler stems were restricted to [ə] as well as the [aɪ]∼[ɪ] and [aʊ]∼[ʌ] pairs (*i.e.* trisyllabic shortening pairs that were not used as the target vowels in the experimental task). The distribution of the filler target vowels was identical to that of the test stems (albeit halved). The distribution of the filler stems amongst the three Tenseness conditions is shown in Table 2.2 (vertical axes).

[5]The sounds [t, d, s, z, t͡ʃ, d͡ʒ] were removed from the set of codas for test stems. This was done to sidestep potential confounding factors that might arise from *yod*-coalescence. The sounds [h, j, w] were removed too because they couldn't serve as codas. [ɹ] was excluded because tenseness is neutralized before an [ɹ] coda. Since C_3 becomes an onset upon *-ian* affixation, [ŋ] was excluded because it could not serve as an onset.

In the filler task, which served as a distractor, participants had to choose among palatalized and unpalatalized affixed options. For this reason, the unaffixed stem-final consonant was restricted to [t, d, s, z, t͡ʃ, d͡ʒ]. The 2AFC filler options were created by modifying the filler stems in the following ways (6):

(6) a. The *-ian* [ɪən] suffix was added.

   i. All affixed options carried primary stress on the target vowel.

b. For each pair of affixed options, one option would coalesce[6] the stem-final vowel with [ɪ] (palatalized option), while the other remained unchanged (unpalatalized option).

c. Regarding target vowels that bore stress in the unaffixed form. For half of these stems, their target vowels for <u>both</u> affixed options were changed to their tense∼lax counterpart upon affixation. The other half retained their original tenseness value upon affixation.

d. For each pair of affixed options, both affixed options would carry one of [aɪ, ɪ, aʊ, ʌ]. The distribution of [aɪ, ɪ, aʊ, ʌ] was split evenly among ə-stem filler stems.

For example, the filler stem [ˈko͡ʊsəs] would have these two affixed options: [ˌko͡ʊˈsɪsiən], [ˌko͡ʊˈsɪʃən] (6a-b).

In addition to the consonant changes above, modifications to the target vowel of the affixed options were also necessary. In particular, stressed stem vowels were hypothesized to retain their tenseness value even if this would result in a dispreferred lax vowel in the __CiV context. For the test stems, the task was to choose between retaining or alternating the target vowel to its tense counterpart. Hence, it wouldn't do for all the filler items to only show retention. To avoid inadvertently biasing participants towards vowel retention, filler stems with stressed target vowels were modified following (6c). For example, half of the filler stems with stressed target vowels

[6]Coalescence pairs: [tɪ] → [t͡ʃ], [dɪ] → [d͡ʒ], [sɪ] → [ʃ], [zɪ] → [ʒ]. The last two pairs were, strictly speaking, deletion: [t͡ʃɪ] → [t͡ʃ], [d͡ʒɪ] → [d͡ʒ].

would have this pattern: [ˈɹɛˌlɪs] ↔ {[ˌɹɛˈlaɪsiən], [ˌɹɛˈlaɪʃən]}, where the target vowel [ɪ] in the unaffixed stem alternates to [aɪ] in both affixed options. The other half would have this pattern: [təˈlɪs] ↔ {[təˈlɪsiən], [təˈlɪʃən]}, where the target vowel in the unaffixed stem [ɪ] is retained in both affixed options. The distribution of retained/alternated target vowels is shown with blue and yellow highlights respectively in Table 2.2.

Stem		***Affix Tenseness***	
Tenseness	***Stress***	*Tense*	*Lax*
ə	*None*	4	4
Tense	*Primary*	2	2
	Secondary	2	2
Lax	*Primary*	2	2
	Secondary	2	2

Tenseness: retained, alternated

(a) Front

Stem		***Affix Tenseness***	
Tenseness	***Stress***	*Tense*	*Lax*
ə	*None*	4	4
Tense	*Primary*	2	2
	Secondary	2	2
Lax	*Primary*	2	2
	Secondary	2	2

(b) Back

Table 2.2: *Distribution of filler stems across conditions (Expt 1)*

In Table 2.2, the vertical axis indicates the tenseness of the target vowel in the unaffixed stem while the horizontal axis indicates the tenseness of the target vowel in the affixed options. When tenseness is retained, tenseness values for both the vertical and horizontal axes match. When the target vowel alternates, tenseness values for the vertical and horizontal axes differ.

Finally, for each filler stem with a [ə] target vowel, the affixed options were modified following (6d).

The other segments of the filler stem (*i.e.* C_1, V_1, C_2) were irrelevant to the research question. Just as for the test stems, they were randomly generated according to their unigram frequencies.

The stimuli were recorded by a phonetically-trained male native speaker of American English who was naïve to the research questions. Word-final stops were released. Otherwise, the pronunciation was as in American English.

2.2.2.3 Procedure

The Experigen software (Becker and Levine, 2014) was used to run the experiment online. Participants were instructed to find a quiet space and to put on headphones before beginning the experiment. They were then presented with the audio of two practice stems (*Portugal* and *Vietnam*), and given feedback on whether they correctly identified these words. Participants could play these practice stems multiple times to help adjust their headsets to a comfortable and clear volume.

The experiment began with two practice trials. In each trial, they saw two sentences, which each had a missing word. The first sentence had a missing noun while the second sentence had a missing adjective. Participants were then presented with a pair of audio files. Each audio file contained a noun∼adjective sequence. For example, one audio file would play *Portugal∼Portuguese* while the other would play *Portugal∼Portulese*. Participants then picked the audio option that they preferred. An example trial screen is shown in Figure 2.1.

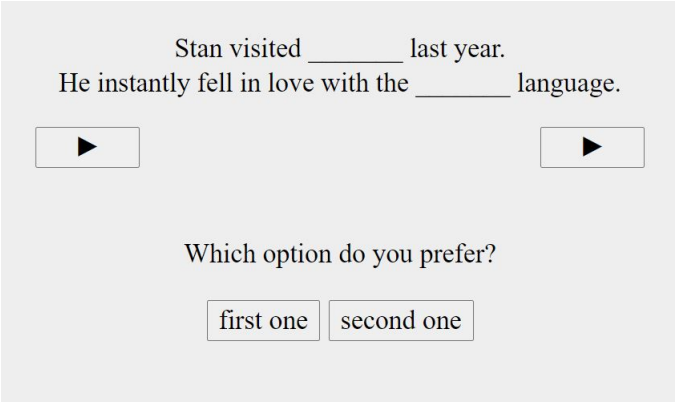

Figure 2.1: *Trial screen*

The two practice trials were followed by the test and filler trials. Participants were informed that they might now encounter words that they were not familiar with, and that repeating the words out loud might aid them in the task. Half the trials played the tense affixed option played first while the other half played the lax affixed option first. The trials were shuffled, so that the order of the presentation of the audio files (tense first or lax first), the order of stems, and the sentence-frame-to-audio-file pairings were different for each participant.

2.2.2.4 Participants

Participants were recruited from the UCLA Psychology Subject Pool, which means that they were undergraduate UCLA students enrolled in a course in psychology or a related field. I excluded participants who were not native English speakers (2 participants), as well as those who had taken more than three linguistics classes (2 participants). After exclusions, a total of 24 participants remained.

2.2.3 Predictions

Before presenting the experimental results, let us recall the two research questions that I sought to answer by conducting Experiment 1 (7):

(7) a. In the absence of vowel information, was there a tense preference in the __CiV context?

b. When the unaffixed stem's target vowel was known, was there a preference to retain its original vowel quality even if that would result in a dispreferred lax vowel in the __CiV context?

The ə-stem condition was relevant to the first question. To support a tense preference in the __CiV context, participants should pick the tense affixed option at a rate greater than chance (4a).

The tense-stem and lax-stem conditions were relevant to the second question. To support a preference towards the retention of full vowels, we'd expect to see that the tense-stems prefer a tense vowel upon affixation at a higher rate than their lax-stem counterparts (4b).

2.2.4 Results: Hypotheses

The proportion of tense responses was .71 for the ə-stems, .75 for the tense-stems, and .56 for the lax-stems (Figure 2.2). The error bars in all figures in this paper represent the 95% confidence interval. The proportion of tense responses for the three tenseness conditions are shown in Figure 2.2. The error bars represent the 95% confidence interval.

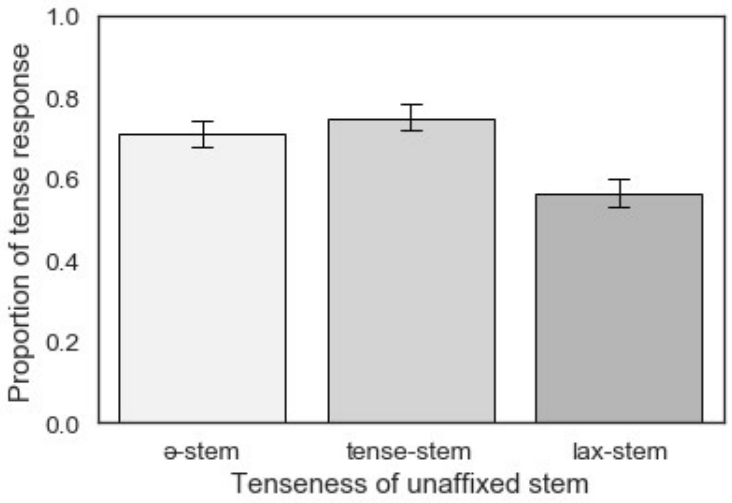

Figure 2.2: *Proportion of tense responses for the three Tenseness conditions.*

The proportion of tense responses was submitted to a mixed-effects logistic regression[7,8] with ə-stem set as the reference level. I included random intercepts for subjects and items. This model's parameters are presented in Table 2.3. For the ə-stem condition, the proportion of tense responses is indeed higher than chance ($\beta_0 = .97,\ p < .001$). This result suggests that a reduced vowel like [ə] prefers to become tense in the __CiV context.

Fixed effects	***Estimate***	***Std. error***	***p***
Intercept	.9702	.1429	<.001***
Tenseness = lax (*vs.* ə)	-.7002	.1350	<.001***
Tenseness = tense (*vs.* ə)	.2052	.1406	.144

Note: *p <.05, **p <.01, ***p <.001

Table 2.3: *Fixed effects (Reference level = ə-stem)*

I reformatted the mixed-effects logistic regression, this time setting tense-stem as the reference level. As before, random intercepts for subjects and items were included. The parameters of this model are presented in Table 2.4. For the tense-stems, the proportion of tense responses is higher than chance ($\beta_0 = 1.18,\ p < .001$). Compared to the tense-stems, lax-stems do in fact show a

[7]I used the `glmer()` function from the `lme4` package (Bates *et al.*, 2015) in *R* (R Core Team, 2018).

[8]These mixed effects logistic regression models only had one independent variable – unaffixed stem tenseness.

reduced porportion of tense responses ($\beta = -.91,\ p < .001$). This result suggests that for non-reduced vowels, the tenseness value in the unaffixed form has an effect on the tenseness value in the affixed form, with tense-stems preferring the tense vowel at a higher rate than their lax-stem counterparts.

Fixed effects	***Estimate***	***Std. error***	***p***
Intercept	1.1754	.1453	<.001***
Tenseness = ə (*vs.* tense)	-.2052	.1406	.144
Tenseness = lax (*vs.* tense)	-.9054	.1374	<.001***

Note: *p <.05, **p <.01, ***p <.001

Table 2.4: *Fixed effects (Reference level = tense-stem)*

2.2.5 Discussion: Hypotheses

Both of the hypotheses set out in §2.2.3 were supported, thus indicating that the generalizations in (3) were productive. ə-stems were found to prefer a tense target vowel upon affixation. This indicated that there was indeed a tense preference in the __CiV context when the original vowel was reduced. Tense-stems were found to have a greater tense preference than their lax-stem counterparts. This pointed towards the presence of a retention effect for non-reduced vowels, wherein a stem's original vowel quality was retained.

The careful reader might have noticed that lax-stems appear to have a tense preference. This was, in fact, marginally so, though failing to meet the criterion for significance when α was set at $.05$. (For the mixed effects logistic regression model with lax-stem set as the reference level, there was a trend for the tense response to be higher than chance ($\beta_0 = .2700,\ p = 0.0518$).) This might seem surprising because if full vowels truly retained their original quality, we'd expect lax-stems to show a lax preference (all else being equal).

Nevertheless, there might be a confounding factor that skewed responses for all conditions towards the tensed affixed form. The experiment was designed to test participants' knowledge about the relation between the affixed form and it unaffixed counterpart. Stimuli were designed with this in mind; in order to nudge participants towards performing an affixation task (*i.e.* a wug task), the

relevant unaffixed form was played immediately before the presentation of an affixed form. Despite this, a phonotactic task was also available in the background. To perform a phonotactic task, one need only compare the grammaticality of the two affixed forms without paying heed to the relation between the affixed form and its unaffixed counterpart. In a phonotactic task, the proportion of tense responses would reflect the goodness of tense vowels in the __CiV context (agnostic of morpheme boundary). The two possible tasks are summarized in (8):

(8) a. Affixation (*i.e.* wug test)
    i. How good is the mapping between the unaffixed stem & its affixed form?
    ii. *e.g.* Given this stem, which affixed option do you prefer?
  b. Phonotactics (*i.e.* blick test)
    i. How good do these words sound on their own?
    ii. *e.g.* Does the first audio file or the second audio file have words that sound better? (In a pure phonotactic task, the mapping between the unaffixed stem and its affixed counterpart is ignored.)

Participants could be thought of as performing a mixture of these two tasks. Different participants might prioritize these two tasks at different rates. For example, one participant might give greater priority to how each of the words sound and give lower priority to the mapping between the unaffixed stem and its affixed form. Another participant might give roughly equal weights to both tasks, while yet another participant might prioritize the affixation task and pay minimal attention to how each of the words sound on their own.

If participants had performed a mixture of the affixation and phonotactic tasks, we'd expect the vowel retention effects to be moderated by the phonotactic task. To perform a phonotactic task, participants would be guided by the frequencies of tense [eɪ, oʊ] versus lax [æ, ɑ (& ɔ)] in the __CiV context. Consulting the Oxford English Dictionary, I found that there were 770 words in the Oxford English Dictionary (OED), that had one of these four vowels in the __CiV context (agnostic of morpheme boundary). Of these 770 words, the proportion of the tense vowels [eɪ, oʊ] out of

the set of vowels [eɪ, æ, ɑ (& ɔ[9]), oʊ][10] in the __CiV context (agnostic of morpheme boundary) was .82. This indicated that any potential effect arising from an interfering phonotactic task should push responses towards the tense affixed form (and thus away from the lax affixed form). Given the potential interaction between the phonotactic and the affixation tasks, the trend towards a tense preference for lax-stems isn't unexpected, and should not constitute evidence against the retention of lax vowels in the __CiV context by lax-stems.

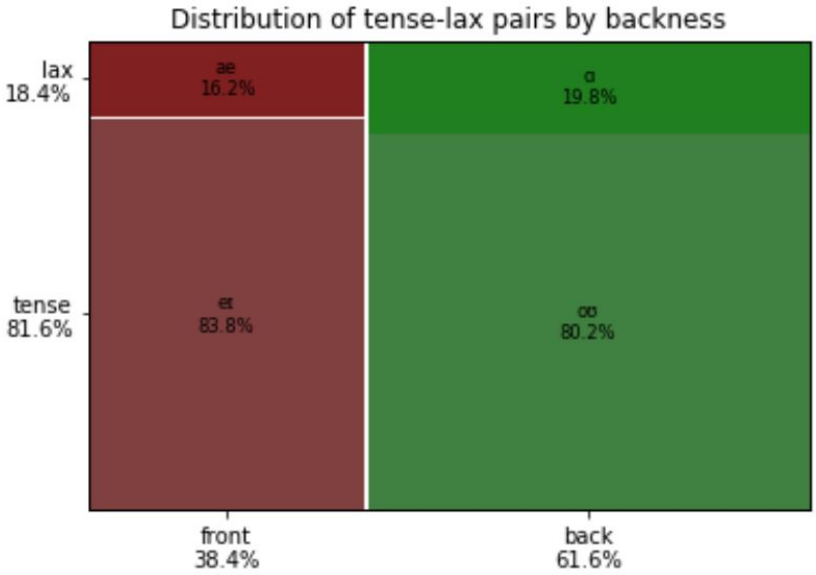

Figure 2.3: *Distribution of tense-lax pairs in the __CiV context (OED)*

2.2.6 Results: Interactions

We have seen that the proportion of tense responses is dependent on the tenseness of the unaffixed stem. We will now turn our attention to other factors and ask what effect, if any, do factors like backness and stress level have on proportion of tense responses. Looking ahead, we will see that backness and stress level each moderate the relationship between a stem's unaffixed tenseness and

[9]I restricted the set of vowels to [eɪ, æ, ɑ (& ɔ), oʊ] since these were the only target vowels used in the experiment.

[10]The /ɔ/ vowel does not exist in most varieties of American English, including the varieties prevalent in California, where most of the experimental participants were from. Since the /ɔ/ and /ɑ/ vowels have merged in General American English, I collapsed the /ɔ/ and /ɑ/ vowels into the /ɑ/ class. Regarding the experimental stimuli, the speaker produced all instances of /ɑ/ as [ɑ]. There were no stimuli with /ɔ/.

the proportion of tense responses.

Recall from §2.2.2.1 that that there were three predictors: tenseness (ə-stems, tense-stems, lax-stems), backness (front, back), and stress-level (none, primary, secondary). Tenseness was fully crossed with backness . Stress-level was also fully crossed with backness.

Tenseness and stress-level could not be fully crossed because all ə-stems necessarily bore no stress (/ə/ is an unstressed vowel). Nevertheless, within the full-voweled stems (tense-stems & lax-stems), tenseness was fully crossed with stress-level (primary & secondary).

Back vowels amplified the retention effect of full-voweled stems (Figure 2.4). In a mixed

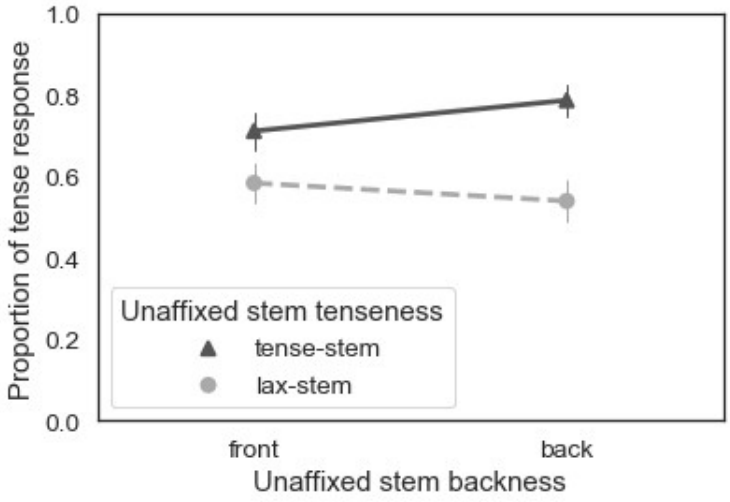

Figure 2.4: *Back vowels amplify the retention effect.*

effects logistic model[11] with random intercepts for item and subject, and with reference values set to "tense-stem, back, and secondary stress", the intercept was found to be significant ($\beta_0 = 1.26$, $p < .001$). Laxness ($\beta = -.92$, $p < .001$) and frontness ($\beta = -.61$, $p = .02$) each decreased the proportion of tense responses. However, being both lax and front led to a smaller decrease than expected ($\beta = .80$, $p = .02$). This indicated that front vowels had a smaller retention effect that their back-voweled counterparts did. This model found no other main effects or interactions at $\alpha = .05$. This model's parameters are shown in Table 2.5.

[11]The mixed effects models in this section include all three independent variables (unaffixed stem tenseness, backness and stress-level) as predictors. In this regard, the models in this section differ from those in §2.2.4, which have the unaffixed stem tenseness as their sole independent variable.

Fixed effects	*Estimate*	*Std. error*	*p*
Intercept	1.2597	.2160	<.001***
Tenseness = ə (*vs.* tense)	-.2442	.2282	.285
Tenseness = lax (*vs.* tense)	-.9184	.2532	<.001***
Stress = primary (*vs.* secondary)	.2602	.2755	.345
Backness = front (*vs.* back)	-.6135	.2558	.017*
Tenseness = lax (*vs.* tense) & Stress = primary (*vs.* secondary)	-.5972	.3633	.100
Tenseness = ə (*vs.* tense) & Backness = front (*vs.* back)	.5075	.3124	.104
Tenseness = lax (*vs.* tense) & Backness = front (*vs.* back)	.7979	.3502	.023*
Stress = primary (*vs.* secondary) & Backness = front (*vs.* back)	.4151	.3772	.271
Tenseness = lax (*vs.* tense) & Stress = primary (*vs.* secondary) & Backness = front (*vs.* back)	-.3981	.5052	.431

Note: *p <.05, **p <.01, ***p <.001

Table 2.5: *Fixed effects (Reference level = tense, back & secondary-stressed stems)*

The retention effect was likewise amplified for stems whose target vowels were primary-stressed in the unaffixed form (Figure 2.5). I reformatted the model with reference levels now

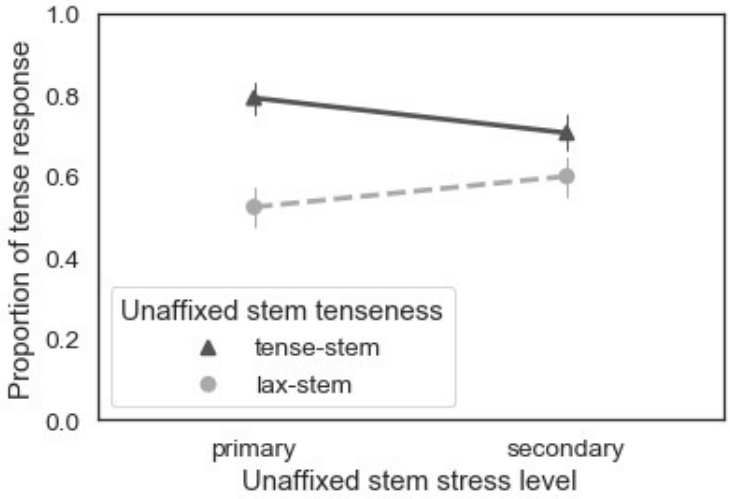

Figure 2.5: *Primary-stressed target vowels amplify the retention effect.*

set to "tense-stem, front, and primary stress". The intercept was significant ($\beta_0 = 1.32$, $p < .001$). In isolation, laxness ($\beta = -1.12$, $p < .001$) and secondary stress ($\beta = -.68$, $p = .01$) each re-

duced the proportion of tense responses. However, being both lax and secondary stressed led to a smaller reduction than expected ($\beta = 1.00$, $p = .005$), thus confirming the interaction between the unaffixed stem's tenseness & stress-level. In particular, having a secondary-stressed target vowel in the unaffixed stem resulted in a smaller retention effect. This model found no other main effects or interactions. The model parameters are shown in Table 2.6.

Fixed effects	***Estimate***	***Std. error***	***p***
Intercept	1.3216	.2182	<.001***
Tenseness = ə (*vs.* tense)	-.4120	.2291	.072
Tenseness = lax (*vs.* tense)	-1.1156	.2545	<.001***
Stress = secondary (*vs.* primary)	-.6754	.2576	.009**
Backness = back (*vs.* front)	.1984	.2772	.474
Tenseness = lax (*vs.* tense)	.9951	.3511	.005**
& Stress = secondary (*vs.* primary)			
Tenseness = ə (*vs.* tense)	.0925	.3301	.779
& Backness = back (*vs.* front)			
Tenseness = lax (*vs.* tense)	.4000	.3641	.271
& Backness = back (*vs.* front)			
Stress = secondary (*vs.* primary)	.4153	.3771	.271
& Backness = back (*vs.* front)			
Tenseness = lax (*vs.* tense)	-.3981	.5051	.431
& Stress = secondary (*vs.* primary)			
& Backness = back (*vs.* front)			

Note: *p <.05, **p <.01, ***p <.001

Table 2.6: *Fixed effects (Reference level = tense, front & primary-stressed stems)*

2.2.7 Discussion: Interactions

In §2.2.5, we saw that full-voweled stems retained their original tenseness. The interactions found in §2.2.6 suggest that factors like backness and stress-level moderate the retention of a stem's original tenseness. In particular, having a target vowel that is back or that bears primary stress amplifies[12] the retention effect. Nevertheless, backness and stress-level only affect the strength of the retention effect; they never reverse or eliminate it. The proportion of tense responses for tense-stems always remains higher than that of their lax-stem counterparts under all combinations

of conditions.

2.3 Experiment 2

Having ascertained in Experiment 1 that the basic generalization regarding CiV Lengthening that were set out in (3) are productive, we will now turn our attention to the triggering environment of CiV Lengthening.

2.3.1 Method

The triggering environment for CiV Lengthening has been described as __CiV (Steriade, 2019). This is an oddly specific environment. In particular, why is there only one consonant in the environment? What happens if the lone consonant is replaced with a sequence of consonants? This experiment sets out to investigate whether having just one consonant in the triggering environment of CiV Lengthening is justified.

2.3.1.1 Conditions

In this experiment, stems took one of two possible shapes: $C_1V_1C_2V_2C_3$ (single stem-final C) and $C_1V_1C_2V_2C_3C_4$ (two stem-final C's). One-third of the test stems had a single stem-final consonant, while the other two-thirds had two stem-final consonants. As in Experiment 1, the target vowel was V_2, with V_2 presenting two tenseness options upon affixation.

There were three stem-final consonant conditions for the test stems: single-C, onset double-C, and coda double-C. Test stems were evenly split between all three conditions. The single-C condition had a single stem-final consonant, which was a legal coda in English. The onset double-C condition featured legal onset consonant sequences in English (*e.g.* [sp, sk]), while the coda double-C condition had consonant sequences that could not function as onsets (*e.g.* [ŋk, mp, lf]).

[12]This could be due to more words that exhibit retention coming from the back group than those coming from the front group in English. There might also be more words exhibiting retention that come from the primary stressed group than those coming from the secondary stressed group.

Since stems were presented both in isolation and with a suffix, all stem-final consonant (sequences) had to also be legal English codas.

As with Experiment 1, the target vowel, V_2, was subject to three tenseness conditions (ə-stem, tense-stem, and lax-stem), and two backness conditions (front and back). The tenseness and backness conditions were fully crossed with the stem-final condition to produce the distribution of test stems in Table 2.7. There were a total of 108 test stems.

Tenseness	***Backness***	***Stem-final consonants***		
		Single-C	*Onset double-C*	*Coda double-C*
ə-stem	*Front*	6 [ˈkudəb]	6 [ˈpɪləsk]	6 [ˈtɑləŋk]
	Back	6 [ˈziləv]	6 [ˈgufəsk]	6 [ˈzipənʃ]
Tense-stem	*Front*	6 [jəˈteɪk]	6 [kəˈdeɪsk]	6 [səˈfeɪŋk]
	Back	6 [pəˈboʊk]	6 [nəˈdoʊsk]	6 [kəˈboʊmp]
Lax-stem	*Front*	6 [səˈdæl]	6 [gəˈwæsp]	6 [kəˈkælv]
	Back	6 [ɹəˈlɑn]	6 [kəˈmɑsk]	6 [səˈɹɑlf]

Table 2.7: *Distribution of test stems across conditions (Expt 2).*

Since [ə] was unable to bear stress in English, the target vowel for all ə-stems were necessarily unstressed. The target vowels for tense-stem and lax-stem items all bore primary stress.

Excepting V_2 and the stem-final consonant(s) discussed above, all other segments were irrelevant to experimental manipulation.

2.3.1.2 Materials

For the single-C condition, all stems for each of the six sub-conditions in Table 2.7 (*i.e.* front ə-stem, back ə-stem back, front (primary-stressed) tense-stem, back (primary-stressed) tense-stem,

front (primary-stressed) lax-stem, back (primary-stressed) lax-stem) were recycled from Experiment 1. This produced a total of 36 recycled test stems.

To create the onset double-C and coda double-C stem-final conditions, 72 other stems representing the six sub-conditions in Table 2.7 were randomly chosen from the test items used in Experiment 1. These stems had their final consonant removed and replaced with a consonant cluster drawn from [sp, sk] or [lm, ln, lf, lv, lθ, lʃ, lp, lb, lk, mf, mθ, mʃ, nθ, nʃ, nʒ, ŋθ, mp, mb, nk, ŋk][13] for the onset double-C and coda double-C stem-final conditions respectively. The frequency at which a particular consonant cluster was picked depended on its word-final frequency in the Oxford English Dictionary (OED).

In Experiment 1, the frequency at which a non-experimentally-manipulated sound was picked depended on its unigram frequency. For Experiment 2, consonant clusters (bigrams) were introduced. In order to have the consonant clusters reflect real-world frequencies as closely as possible, these bigrams needed to meet several criteria (9).

(9) a. Be a legal English coda.

b. Match real-world coda frequency.

Amongst other constraints, English codas must minimally have a decreasing sonority profile. To achieve (9a), I wrote a script that generated a set of bigram combinations with decreasing sonority profiles, B. Criterion (9b) was more challenging to achieve. I could not find a dictionary or corpus that reliably transcribed syllable boundaries, which were necessary to identify codas. In the absence of a large corpus with transcribed syllables boundaries, I decided to rely on word boundaries as a proxy for syllable boundaries. I chose the OED, which had 456,890 non-obsolete words. The phonological transcriptions for these words were available via the Oxford Dictionaries API. I wrote a script that collected the word-final frequencies for each of the bigrams in B.

Mirroring Experiment 1, the stem-final consonants [t, d, s, z, t͡ʃ, d͡ʒ] were reserved for filler stems. In Experiment 2, this meant that bigrams ending with [t, d, s, z, t͡ʃ, d͡ʒ] were removed from

[13]This list includes all the unique word-final bigrams in the Oxford English Dictionary (OED), which do not include the reserved sounds [t, d, s, z, t͡ʃ, d͡ʒ]. Some of the bigrams in this list are rather odd (*e.g.* [mb, nk]). Nevertheless, the odd bigrams have low frequencies and a check of the frequency-informed randomly generated stimuli showed that none of the test stems had any odd codas.

the set of bigrams for test stems. The frequencies of bigrams that were suitable for test stems are shown in Table 2.8. Figure 2.6 gives a visual representation of the bigrams that could only serve as codas (not onsets), ordered by decreasing sonority. Only bigrams with non-zero frequencies are shown.

Bigram	*Frequency*	*Percent*
sk	354	86.13
sp	57	13.87
Total	411	100.00

(a) *May be onset*

Bigram	*Frequency*	*Percent*
ŋk	626	38.74
mp	361	22.34
nθ	92	5.69
lf	85	5.26
lk	73	4.52
lv	69	4.27
nʃ	68	4.21
lp	47	2.91
lm	44	2.72
mf	37	2.29
lθ	23	1.42
lʃ	21	1.30
mb	13	.80
ŋθ	12	.74
ln	12	.74
lb	9	.56
nʒ	9	.56
nk	8	.50
mθ	6	.37
mʃ	1	.06
Total	1,616	100.00

(b) *Only coda*

Table 2.8: *Word-final bigram frequencies (OED)*
(Bigrams not ending with [t, d, s, z, t͡ʃ, d͡ʒ])

The same procedure from Experiment 1 (5) was used to create the 2AFC affixed options from their test stem counterparts in Experiment 2.

There were 54 filler stems, which was half the number of test stems. The stem-final consonant for these fillers was restricted to [t, d, s, z, t͡ʃ, d͡ʒ]. As with Experiment 1, the target vowels were [ə] or the [aɪ]∼[ɪ] and [aʊ]∼[ʌ] pairs. The distribution of filler stems is presented in Table 2.9.

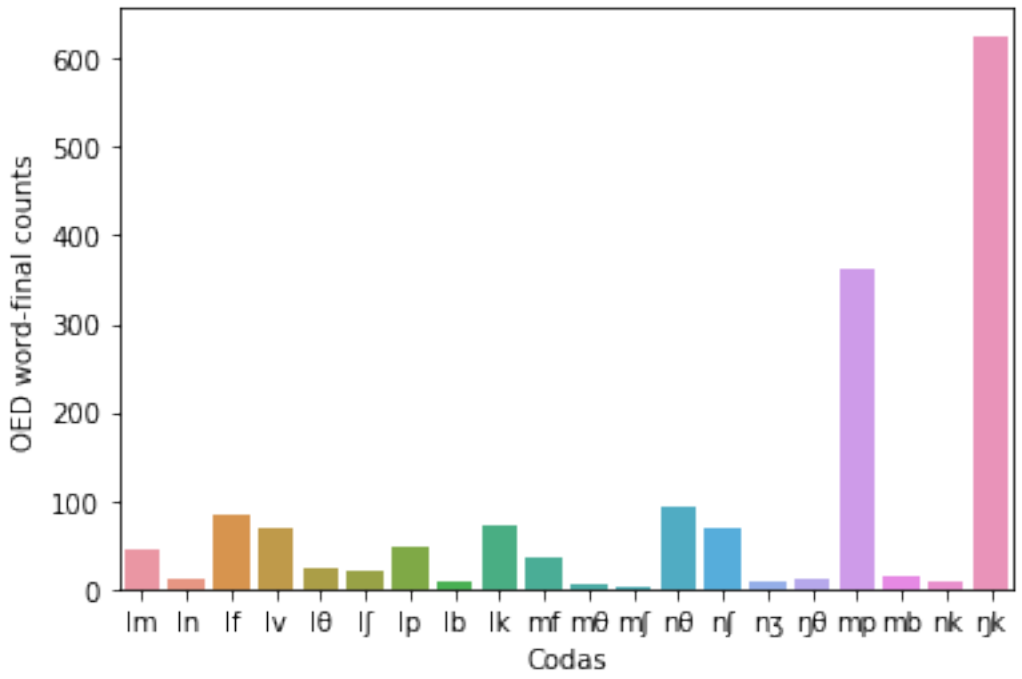

Figure 2.6: *Word-final bigram frequencies from OED (only codas)*

Stem		***Affix Tenseness***	
Tenseness	***Stem-final***	*Tense*	*Lax*
ə	*Onset*	1	2
	Coda	2	1
	One	1	2
Tense	*Onset*	1	2
	Coda	2	1
	One	1	2
Lax	*Onset*	1	2
	Coda	2	1
	One	1	2

Tenseness: retained , alternated

(a) Front

Stem		***Affix Tenseness***	
Tenseness	***Stem-final***	*Tense*	*Lax*
ə	*Onset*	2	1
	Coda	1	2
	One	2	1
Tense	*Onset*	2	1
	Coda	1	2
	One	2	1
Lax	*Onset*	2	1
	Coda	1	2
	One	2	1

(b) Back

Table 2.9: *Distribution of filler stems across conditions (Expt 2)*

For the single-C stem-final condition, 18 filler stems were recycled from Experiment 1. For the onset double-C and coda double-C stem-final conditions, 36 other filler stems were selected. These filler stems had their final consonant removed and replaced with bigrams that ended with [t, d, s, z, t͡ʃ, d͡ʒ]. The rate at which a particular bigram was picked reflected its word-final frequency in the OED (Table 2.10).

Following Experiment 1, the applications of yod-coalescence and yod-dropping were the distractor tasks. The 2AFC affixed options were created accordingly (6a-b). As with Experiment 1, the target vowels of both affixed options for each filler stem were the same, and the four target vowels [aɪ, ɪ, aʊ, ʌ] were evenly split evenly among conditions (6d). I also endeavored to distribute the retained *vs.* alternating vowels between conditions as evenly as possible (6c).

The stimuli were recorded by the same speaker who recorded the Experiment 1 stimuli.

2.3.1.3 Procedure

This procedure was identical to that of Experiment 1.

2.3.1.4 Participants

Participants were recruited via the UCLA Psychology Subject Pool. Those who were not native English speakers (4 participants) or who had taken more than three linguistics classes (1 participant) were excluded. 13 participants remained after exclusions.

2.3.2 Predictions

Experiment 2 is designed to determine whether the environment for CiV Lengthening is indeed __CiV; specifically whether the environment has to be restricted to only one consonant or whether CiV Lengthening takes place even when the environment contains multiple consonants. If the environment as described in the literature is correct, then we expect to see a higher proportion of tense responses in the single-C stem-final condition than in the onset double-C and coda double-C stem-final conditions. If the number (whether it be one or two) consonants do not matter for

Bigram	*Frequency*	*Percent*
st	5,873	100.00
Total	5,873	100.00

(a) *May be onset*

Bigram	*Frequency*	*Percent*
nt	5,145	34.02
nd	2,969	19.63
ns	1,624	10.74
ld	1,388	9.18
zd	894	5.91
lt	381	2.52
ft	371	2.45
ʃt	331	2.19
md	323	2.14
nz	252	1.67
vd	231	1.53
ʒd	203	1.34
nd͡ʒ	200	1.32
d͡ʒd	193	1.27
t͡ʃt	131	.87
lz	103	.68
ŋd	72	.48
mz	68	.45
ðd	49	.32
ls	49	.32
θt	39	.26
ŋz	36	.24
lt͡ʃ	23	.15
nt͡ʃ	19	.13
ld͡ʒ	13	.09
sd	7	.05
mt	5	.03
θd	2	.01
ms	2	.01
ʃd	1	.01
Total	15,124	100.00

(b) *Only coda*

Table 2.10: *Word-final bigram frequencies (OED)*
(Bigrams ending with [t, d, s, z, t͡ʃ, d͡ʒ])

the triggering environment, then we expect there to be no difference in the proportion of tense responses between the single-C stem-final condition and the two double-C stem-final conditions[14].

2.3.3 Results

The proportion of tense responses for the three stem-final conditions is shown in Table2.11. The

	Stem-final consonants			
Tenseness	*Single-C*	*Onset double-C*	*Coda double-C*	***Total***
Tense-stem	.71	.70	.69	.70
ə-stem	.61	.45	.53	.53
Lax-stem	.42	.33	.33	.36
Total	.58	.49	.51	.53

Table 2.11: *Proportion of tense responses grouped by Tenseness and Stem-finality.*

data is presented graphically in Figure 2.7. The barplots suggest that the differences between the three stem-final conditions are amplified when only reduced-voweled stems (*i.e.* ə-stems) are considered. This is not unexpected since we have already learned from Experiment 1 that the full-voweled stems (*i.e.* tense-stems and lax-stems) exhibit a retention effect, where there is a pressure to retain the tenseness value of the unaffixed stems' tenseness. For ə-stems, the retention effect cannot serve as a confounding factor, so any effect that stem-final consonants have on the proportion of tense responses would best show up in this condition.

To test the prediction laid out in §2.3.2, I submitted the proportion of tense responses to a mixed-effects logistic regression model. Stem-finality was the only independent variable in this model. The single-C stem-final condition was set as the reference level. Random intercepts for subjects and items were also included. This model found that having a plausible onset (*e.g.* [sp, sk])

[14]English coda consonants contribute to the weight of a syllable. Accordingly, the coda doucle-C condition is predicted to have a smaller proportion of tense responses compared to the onset double-C condition. The ideal weight of a stressed syllable is for the syllable to be heavy. If the rime of a syllable already contains a coda, adding a tense vowel makes it extra-heavy, which is sub-optimal. In contrast, a rime consisting of a coda with a lax vowel is heavy. Since the syllable itself is stressed, the heavy syllable is the most optimal syllable, so a lax vowel is preferable in the onset double-C condition. For the onset double-C condition, no codas are present to contribute to syllable weight. Hence, the tense vowel is prefered because a rime that consists of only a tense syllable is heavy (optimal) while a rime consisting of only a lax syllable is light (sub-optimal).

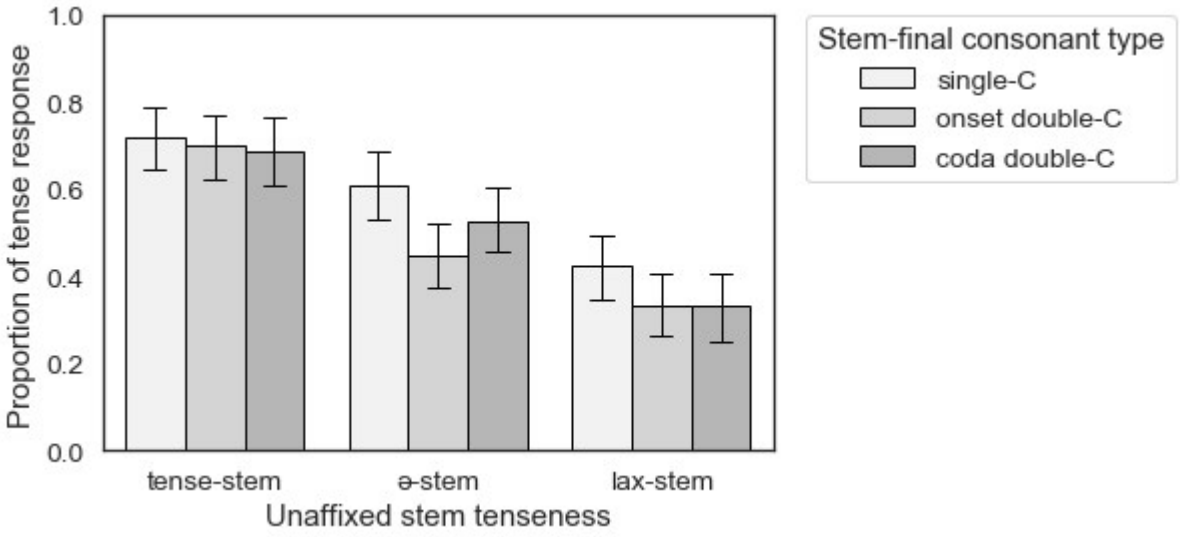

Figure 2.7: *Proportion of tense responses by stem-final consonant type & unaffixed stem tenseness.*

as the stem-final consonant cluster marginally reduced the proportion of tense responses compared to having just a single stem-final consonant ($\beta = -.40,\ p = .07$). However, having a stem-final consonant cluster that could only serve as a coda (*e.g.* ŋk, mp, lf) had no effect ($\beta = -.31,\ p = .16$) on the proportion of tense responses. This model's parameters are shown in Table 2.12.

Fixed effects	***Estimate***	***Std. error***	***p***
Intercept	.3809	.1767	.0311*
Stem-final = coda double-C (*vs.* single-C)	-.3129	.2250	.1644
Stem-final = onset double-C (*vs.* single-C)	-.4043	.2249	.0723

Note: *p <.05, **p <.01, ***p <.001

Table 2.12: *Fixed effects, all stems included (Reference level = single stem-final consonant).*

The barplots in Figure 2.7 suggested that the stem-final effects were amplified when the unaffixed stem's target vowel was a reduced vowel, as in the ə-stem condition. This was because the retention effect could not serve as a confounding factor for ə-stems. Hence, I followed up with another mixed-effects logistic regression model, this time including only the ə-stems. The fixed and random variables that I used for this model were otherwise identical to that of the model presented in Table 2.12 (*i.e.* The model that included tense-stems and lax-stems as well.). This model found that having a sequence of stem-final consonants that may be resyllabified as onsets (*i.e.* the onset

double-C stem-final condition. *e.g.* [sp, sk]) does indeed reduce the proportion of tense responses when compared to having a single stem-final consonant ($\beta = -.69$, $p = .02$). However, there was no change in the proportion of tense responses when the stem-final consonants could not both be resyllabified into onsets (*e.g.* [ŋk, mp, lf]), as in the coda double-C stem-final condition ($\beta = -.37$, $p = .23$). This model's parameters are presented in Table 2.13.

Fixed effects	***Estimate***	***Std. error***	***p***
Intercept	.4764	.2276	.0363*
Stem-final = coda double-C (*vs.* single-C)	-.3673	.3084	.2337
Stem-final = onset double-C (*vs.* single-C)	-.6940	.3093	.0248*

Note: *p <.05, **p <.01, ***p <.001

Table 2.13: *Fixed effects, only ə-stems (Reference level = single stem-final consonant).*

2.3.4 Discussion

When the target vowel was a reduced vowel in the unaffixed stem (*i.e.* the ə-stem condition), having a double consonant sequence that re-syllabified to an onset cluster (*e.g.* [sp, sk]) resulted in a reduction of the proportion of tense responses. Although not significant, a double stem-final consonant sequence that could not be fully resyllabified to an onset (*e.g.* [ŋk, mp, lf]) trended in the direction of reducing the proportion of tense responses. There was thus some support for having only one consonant in the __CiV environment[15]. This was especially so when both stem-final consonants could be resyllabified to onsets upon suffixation (*e.g.* bigrams like [sp, sk]).

It did seem rather curious that a stem-final sequence like [sp], which could be fully resyllabified to an onset, behaved differently than a sequence like [mp], which could not be fully resyllabified to an onset upon suffixation. To better visualize the results, I plotted the proportion of tense responses

[15]In Steriade's view, this odd single consonant environment could be due to the reduced weight of the pre-hiatus vowel in combination with weight being assessed on the vowel-to-vowel interval (Farnetani and Kori (1986), Kato *et al.* (2003), McCrary (2004)) and the ability of both onset and coda consonants to contribute to weight (Ryan, 2000). When there is only one consonant, the underlined vowel-to-vowel interval in $\underline{V_1CV}$.V is underweight when V_1 is a lax vowel, and is at a more optimal weight when V_1 is a tense vowel. When there are two consonants: $\underline{V_1CCV}$.V, the additional consonant helps to make up for the underweight problem, so there is less need for V_1 to be heavy (*i.e.* tense).

against all three independent variables – unaffixed stem tenseness ("tenseness"), backness, and the stem-final consonant type (Figure 2.8).

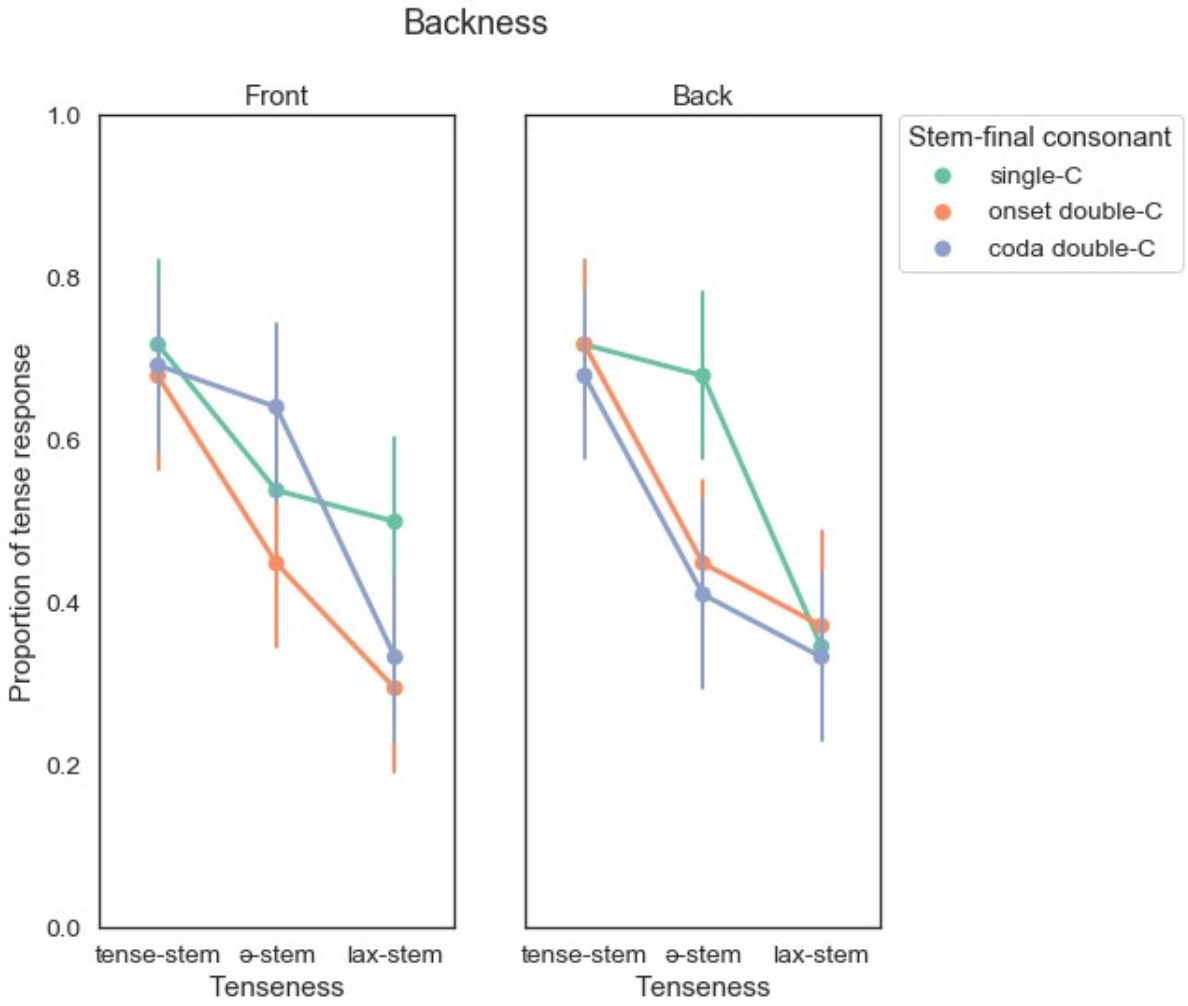

Figure 2.8: *Affixed tense preference grouped by backness, tenseness & stem-final consonant type.*

As expected, we generally saw the tenseness retention effect at play for the tense-stems and the lax-stems. Take for instance the front tense-stem condition. We see in Figure 2.8, a clustering of all three stem-final consonant conditions around 70%. This clustering of the three stem-final consonant conditions is also seen for the back tense-stems and the back lax-stems.

Let us now turn our attention to the ə-stems, where this clustering was not present, so the effect of the stem-final consonant(s) on the proportion of tense responses could not be obscured by a potentially conflicting tenseness retention effect. The picture was clear for the back ə-stems (right panel of Figure 2.8). Having two stem-final consonants (rather than one stem-final consonant)

decreased the proportion of tense responses for the back target vowels. For example, the tense [ˌziˈloʊviən] was a better affixed choice for [ˈziləv] than the tense [ˌziˈpoʊnʃiən] was for [ˈzipənʃ] because the former stem had only one stem-final consonant while the latter stem had two stem-final consonants.

Nevertheless, the picture was not so straightforward for the front vowels (left panel of Figure 2.8). Compared to the single-C stem-final condition (*e.g.* [ˈkudəb]→[ˌkuˈdeɪbiən, ˌkuˈdæbiən]), having two stem-final consonants that could both be resyllabified to onsets "onset double-C" (*e.g.* [ˈpɪləsk]→[ˌpɪˈleɪskiən, ˌpɪˈlæskiən]) appeared to likewise decrease the proportion of tense responses. However, when only one of those two consonants could be resyllabified to an onset "coda double-C" (*e.g.* [ˈtɑləŋk]→[ˌtɑˈleɪŋkiən, ˌtɑˈlæŋkiən]), the proportion of tense responses unexpectedly increased.

To confirm these findings, I performed two mixed-effects logistic regressions, one for the back ə-stems and another for the front ə-stems. These models' parameters are shown in Table 2.14 and Table 2.15 respectively.

Fixed effects	***Estimate***	***Std. error***	***p***
Intercept	.7709	.2798	.0059**
Stem-final = coda (double-C *vs.* single-C)	-1.1422	.3879	.0032**
Stem-final = onset double-C (*vs.* single-C)	-.9821	.3862	.0110*

Note: *p <.05, **p <.01, ***p <.001

Table 2.14: *Fixed effects, only back ə-stems (Reference level = single stem-final consonant).*

Fixed effects	***Estimate***	***Std. error***	***p***
Intercept	.1609	.2747	.558
Stem-final = coda double-C (*vs.* single-C)	.4372	.3888	.261
Stem-final = onset double-C (*vs.* single-C)	-.3737	.3842	.331

Note: *p <.05, **p <.01, ***p <.001

Table 2.15: *Fixed effects, only front ə-stems (Reference level = single stem-final consonant).*

The model for the **back** ə-stems found that having two stem-final consonants significantly reduced the proportion of tense affixed responses compared to their single stem-final consonant counter-

parts. (When only one consonant resyllabified ("coda double-C"; [ŋk, mp, lf]): $\beta = -1.14$, $p = .0032$. When both consonants resyllabified ("onset double-C"; [sp, sk]): $\beta = -.98$, $p = .011$.) For the **front** ə-stems, however, having two stem-final consonants produced no effect compared to having just one stem-final consonant ($\beta = .44$, $p = .26$ when only one consonant resyllabified ("coda double-C"; [ŋk, mp, lf]). When both consonants resyllabified ("onset double-C"; [sp, sk]): $\beta = -.37$, $p = .33$).

In general, having two stem-final consonants rather than a single stem-final consonant tended to decrease the proportion of tense responses. This was the case for three out of the four conditions above – namely the "back coda double-C" (*e.g.* ['zipənʃ]→[ˌzi'poʊnʃiən, ˌzi'pɑnʃiən]), "back onset double-C" (*e.g.* ['gufəsk]→[ˌgu'foʊskiən, ˌgu'fɑskiən]), and "front onset double-C" (*e.g.* ['pɪləsk]→[ˌpɪ'leɪskiən, ˌpɪ'læskiən]) conditions (although only in the two "back" conditions did the effect turn out to be significant at $\alpha = .05$). The front ə-stem coda double-C condition (*e.g.* ['tɑləŋk]→[ˌtɑ'leɪŋkiən, ˌtɑ'læŋkiən]), with only one stem-final consonant resyllabifying, was the only one that showed a mild increase in the proportion of tense responses (though again failing to meet the criterion for a significance effect). A schematic of the alternation choices for these four double-C conditions is offered in Table 2.16. The exceptional condition is found in the third row ("front coda").

			2AFC options	
Backness	***Stem-final type***	***Stem***	*Tense*	*Lax*
back	*coda*	…məŋk	…mo͡ʊŋ.kiən	…mɑŋ.kiən
	onset	…məsk	…mo͡ʊ.skiən	…mɑ.skiən
front	*coda*	…məŋk	…me͡ɪŋ.kiən	…mæŋ.kiən
	onset	…məsk	…me͡ɪ.skiən	…mæ.skiən

Note: Coda – only 1 stem-final C resyllabifies. Onset – both stem-final C's resyllabify

Table 2.16: *Schematic of ə-stem alternations.*

In the exceptional front onset double-C condition, the participant had to choose between the tense […me͡ɪŋ.kiən] and the lax […mæŋ.kiən]. The [æŋ] sequence, however, was not legal Californian English, where the vowel in /æŋ/ was often raised and the sequence realized as [eŋ]. Since [ŋk]

was the most common word-final bigram in the OED and stem-final bigrams were randomly drawn according to their corpus frequency, [ŋk] formed 83% of the stems for this particular condition. It was not surprising, then, that this front coda double-C condition boosted the proportion of tense responses compared to baseline (*i.e.* the single-C condition). This increased preference for the tense affixed form, however, was probably due to a specific surface preference for [eŋ] over [æŋ] sequences, rather than the __CCiV environment boosting a preference for tenseness compared to the __CiV environment.

In fact, the back version of the coda double-C condition also appeared to be affected by a specific phonotactic quirk. In this condition, participants had to choose between the tense [...mo͡ʊŋ.kiən] & the lax [...mɑŋ.kiən]. The tense [o͡ʊŋ] sequence was dispreferred in English. By random draw, the relevant [ŋk] sequence formed 50% of the items for this condition. Thus, the decreased proportion of tense responses in the back coda double-C condition (compared to its back single-C counterpart) could not be isolated to the __CiV environment booting tense affixed responses compared to __CCiV.

The two problematic conditions discussed above both concerned the stem-final consonants that could not be fully resyllabified into onsets (*i.e.* "coda double-C"; __C.CiV context). In contrast, the "onset double-C" conditions (*i.e.* __.CCiV context) did not suffer from these complications. Regarding whether the single-C (*i.e.* __.CiV) condition increased the proportion of tense responses compared to the onset double-C condition, we saw mixed results coming from the back versus the front target vowels (Table 2.14 and Table 2.15 respectively). In the back condition, having two stem-final consonants (*e.g.* [ˈgufəsk]→[ˌguˈfoʊskiən, ˌguˈfɑskiən]) rather than one stem-final consonant (*e.g.* [ˈziləv]→[ˌziˈloʊviən, ˌziˈlɑviən]) decreased the proportion of tense responses. In contrast, no such significant effect was found for the front condition (*e.g.* [ˈkudəb]→[ˌkuˈdeɪbiən, ˌkuˈdæbiən] *vs.* [ˈpɪləsk]→[ˌpɪˈleɪskiən, ˌpɪˈlæskiən]).

Why did the front vowels (unlike their back counterparts) fail to exhibit a significant reduction in the proportion of tense responses in the onset double-C condition compared to the single-C condition? One explanation could be that the proportion of tense responses in the onset single-

C condition[16] was uncharacteristically low in Experiment 2 for the front vowels but not for the back vowels. Comparing the proportion of tense responses for the **front** single-C condition across Experiment 1 and Experiment 2, Figure 2.9 shows that the proportion of tense responses was much lower in Experiment 2 compared to Experiment 1 (70.1% in Experiment 1; 53.8% in Experiment 2). In contrast, the proportion of tense responses for the single-C condition remained

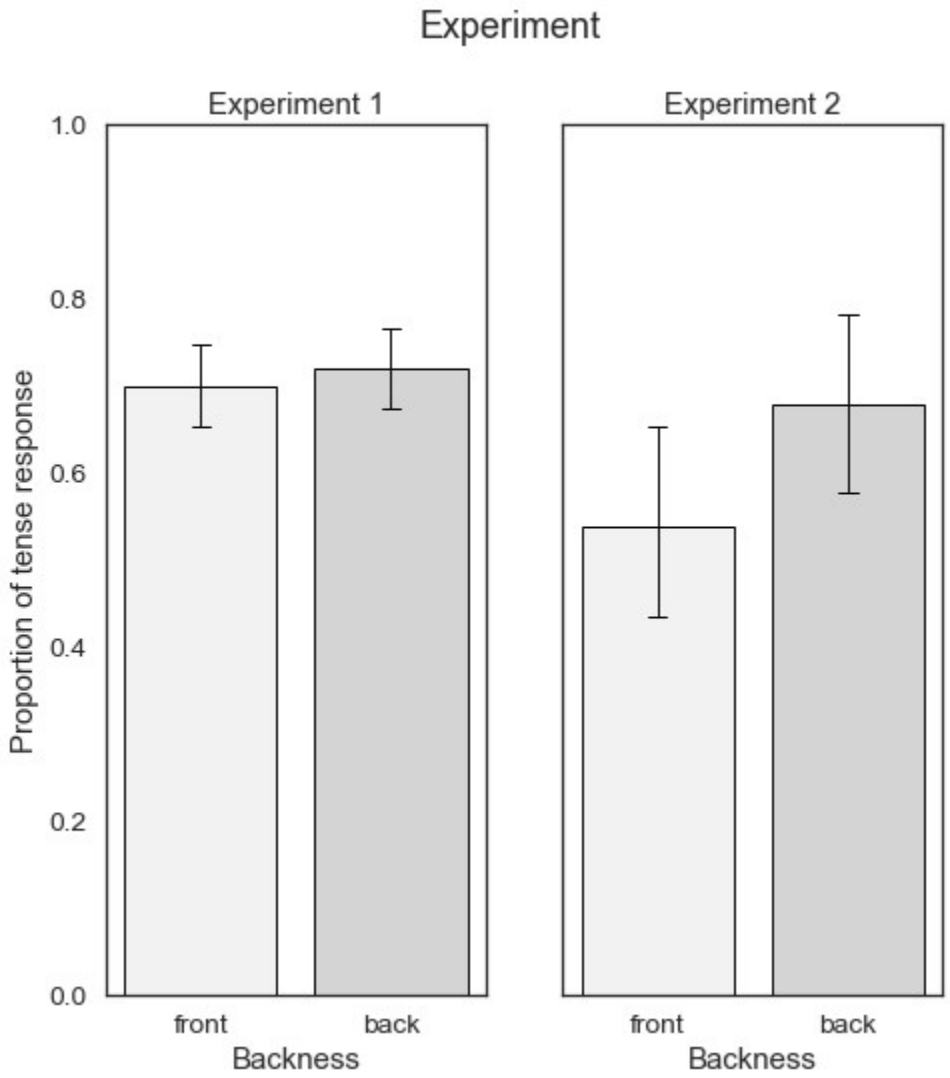

Figure 2.9: *Proportion of tense response for front single-C ə-stems is usually low in Experiment 2.*

stable across Experiment 1 and Experiment 2 for **back** vowels (72.1% in Experiment 1; 67.9% in

[16]In Experiment 1, all test items had a single stem-final consonant.

Experiment 2)[17]. For the **back** vowels, where the proportions of tense responses for the single-C condition remained similar across Experiment 1 and Experiment 2, the predicted increase in the proportion of tense responses for the single-C condition[18] (*i.e.* __CiV context) versus the onset double-C condition[19] (*i.e.* __.CCiV environment) turned out to be a significant effect. The uncharacteristically low proportion of tense responses for the **front** single-C items in Experiment 2 (*e.g.* [ˈkudəb]→[ˌkuˈdeɪbiən, ˌkuˈdæbiən]) might have dampened the predicted increase in the proportion of tense responses for the the single-C condition[20] (*i.e.* __CiV context) versus the onset double-C condition[21] (*i.e.* __.CCiV context) in Experiment 2.

There was reason to believe that had the **front** single-C condition (*e.g.* [ˈkudəb]) in Experiment 2 attained the same proportion of tense responses as it had in Experiment 1, the effect of the single-C condition (*i.e.* __.CiV environment) on increasing the tense preference versus the double-C onset condition (*i.e.* __.CCiV environment) would have come out significant for the **front** onset double-C ə-stems too (and thus would mirror the effect found for their back counterparts).

So, was the single C in the __CiV environment justified? Experiment 2 provided evidence that under certain conditions (namely when the target vowel was reduced in the unaffixed stem, and alternated to a back vowel upon affixation), having two consonants rather than one did indeed reduce the preference for the tense affixed vowel. This might be due to a general preference for tense back vowels in the __CiV environment but not in the __CCiV environment, or it might be due to a more specific dispreference against particular phonotactic sequences (such as an [o͡ʊŋ] rime), or some combination of both. Regardless of which specific mechanisms caused the reduction in tense preference, Experiment 2 supported having a single C in the triggering environment for CiV

[17]These two conditions (*i.e.* the front single-C condition and the back single-C condition) were essentially the same across Experiment 1 & Experiment 2. In fact, a randomly chosen subset of the items from these two conditions in Experiment 1 were used for the very same conditions in Experiment 2.

[18]*e.g.* [ˈziləv]→[ˌziˈloʊviən, ˌziˈlɑviən].

[19]*e.g.* [ˈgufəsk]→[ˌguˈfoʊskiən, ˌguˈfɑskiən].

[20]*e.g.* [ˈkudəb]→[ˌkuˈdeɪbiən, ˌkuˈdæbiən]

[21]*e.g.* [ˈpɪləsk]→[ˌpɪˈleɪskiən, ˌpɪˈlæskiən]

lengthening when a reduced [ə] alternated to a back vowel.

In the case of the reduced [ə] alternating to a front vowel, Experiment 2 found no evidence that having two C's compared to having one C in the triggering environment had any effect on preferring a tense vowel upon affixation.

2.4 Summary of findings

A summary of the results from the two experiments discussed in this chapter is presented below:

- When the target vowel was a reduced vowel in the unaffixed stem, the target vowel preferred to become tense when affixation placed it into the __CiV environment.
  - *i.e.* The proportion of tense affixed responses was higher than chance.
  - *e.g.* For the ə-stem [ˈkudəb], the rate of picking tense [ˌkuˈdeɪbiən] over lax [ˌkuˈdæbiən] was higher than 50%.
- However, when affixation placed the reduced vowel into the __CCiV environment, the tenseness preference disappeared when [ə] alternated to a back vowel. The picture was less clear for the alternation to front vowels.
- When the target vowel was a full vowel in the unaffixed stem, there was pressure to retain the tenseness value of the unaffixed target vowel in the affixed form. (*i.e.* A tenseness retention effect.)
  - *e.g.* There was pressure for the tense-stem [pəˈboʊk] to prefer the tense [pəˈboʊkiən] over the lax [pəˈbɑkiən].
  - *e.g.* Likewise, there was pressure for the lax-stem [ˈtɪˌdɑn] to prefer the lax [ˌtɪˈdɑniən] over the tense [ˌtɪˈdoʊniən].
- Backness and degree of stress were moderating factors for the tenseness retention effect. More concretely, having a back or a primary-stressed target vowel in the unaffixed form increased the tenseness retention effect in the affixed form.
  - *e.g.* [pəˈboʊk], having a back target vowel, would have a stronger tenseness retention effect than [jəˈteɪk], which had a front target vowel.
  - *e.g.* [jəˈteɪk], having a primary-stressed target vowel, would have a stronger tenseness retention effect than [ˈzuˌseɪm], whose target vowel was secondary-stressed.

Taken together, these experimental findings indicate that CiV Lengthening is productive. In addi-

tion, there is some tentative evidence that its rather complex triggering environment – particularly relating to the number of consonants – is justified.

2.5 A new analysis of CiV Lengthening

2.5.1 CiV Lengthening as the Emergence of the Unmarked

In this section, I show that it is possible to model CiV Lengthening without any reference to derived environment effects. That is to say, I make reference to neither morphological derived environments (*e.g.* morpheme boundaries) nor to phonological derived environments (*e.g.* newly stressed vowels). Instead, my analysis hinges on the informativeness (or the lack thereof) of reduced vowels, and what it means to be faithful to a neutralized vowel. In particular, I propose that surface [ə], which serves as a reduced vowel for the non-high full vowels in English, is uninformative because the vowel quality of the full vowel cannot be recovered from surface [ə]. To operationalize the distinction between the tenseness properties of full vowels and reduced ones, I will use [$+/-$tense] for the tenseness values of full vowels and [0tense] for reduced [ə].

I propose that CiV Lengthening is driven by a low-ranked markedness constraint: the Stress-to-Weight Principle (SWP) (Myers (2003), Riad (1992)). The SWP states that all stressed syllables must be heavy. The effect of this markedness constraint only emerges when the target vowel of the unaffixed form is a reduced vowel. When the target vowel of the unaffixed form is a full vowel, a higher ranking faithfulness constraint hides the effect of this markedness constraint.

Let us first consider the case of full-voweled stems. CiV Lengthening suffixes require primary stress to land on the antepenultimate syllable. I assume the presence of an undominated markedness constraint that rules out candidates that do not have primary stress on the antepenult, and so will not include this constraint or its corresponding candidates in the tableaux. The FAITHFULNESS constraint IDENT-OO [tense] requires that the tenseness value of a vowel in the BASE (here: the unaffixed stem) be identical to the tenseness value of its correspondent in the suffixed form. Thus, when lax-stems and tense-stems are suffixed, IDENT-OO [tense] acts to ensure that the target vowel remains unchanged, as in *Chadian* in Table 2.17 and *Ukrainian* in Table 2.18. The

lower-ranked markedness constraint, SWP, has no say in picking the winning candidate. Hence, full-voweled stems retain the tenseness value of their unaffixed form upon affixation with a CiV Lengthening suffix.

[ˈtʃæd] + *ian*$_{suff}$	*ə́	IDENT-OO/IO [tense]	SWP
a. ˈtʃeɪdiən		∗!	
☞ b. ˈtʃædiən			∗
c. ˈtʃədiən	∗!	∗	∗

Table 2.17: *Tableau for Chadian, a full-voweled lax-stem.*

[ˌjuˈkɪeɪm] + *ian*$_{suff}$	*ə́	IDENT-OO/IO [tense]	SWP
☞ a. ˌjuˈkɪeɪmiən			
b. ˌjuˈkɪæniən		∗!	∗
c. ˌjuˈkɪəniən	∗!	∗	∗

Table 2.18: *Tableau for Ukrainian, a full-voweled tense-stem.*

Let us now turn our attention to reduced-voweled stems. In English, reduced vowels like [ə] result from neutralizing-type vowel reductions. That is to say, the [+/− tense] distinction is lost when a vowel gets reduced to [ə]. Hence, I propose that the correspondence between [ə] and [+tense] is as equally good (or bad) as the correspondence between [ə] and [−tense]. One way to formalize this is to treat [ə] is [0tense], so corresponding [+tense] and [−tense] vowels are equally bad with respect to [ə]. (An alternative, if [ə] is treated as [−tense], is to replace IDENT-OO [tense] with IDENT-OO-V, a constraint that is satisfied only if the vowel in the suffixed form is featurally identical to the vowel in the BASE.)

When a ə-stem word, like Can[ə]da, is suffixed, the CiV Lengthening suffix again requires primary stress to land on the antepenult (Table 2.19). For ə-stems, this antepenultimate syllable's nucleus is [ə] in the unaffixed form (*i.e.* the BASE). Candidate (6c) merely shifts stress onto the target vowel while remaining faithful to the vowel quality. Candidate (2.19c) is ill-formed and ruled out by an undominated markedness constraint against stressed schwas. The two remaining candidates (2.19a) and (2.19b) do equally badly on the next highest ranking constraint, IDENT-OO [tense]. The tie between these two candidates is broken by the emergent markedness constraint, SWP, which requires that stressed syllables be heavy. Thus, the unmarked heavy stressed syllable

emerges only when the target vowel was [ə] in the unaffixed surface form.

[ˈkænədə] + *ian*$_{suff}$	*ə́	IDENT-OO/IO [tense]	SWP
☞ a. kəˈneɪdiən		*	
b. kəˈnædiən		*	*!
c. kəˈnədiən	*!		*

Table 2.19: *Tableau for Canadian, a reduced-voweled ə-stem.*

A final set of words remains to be accounted for: the monomorphemic words. CiV Lengthening is reported to not apply to monomorphemic words, which means that there is no restriction on lax vowels showing up in the _CiV environment for these words. OO-Corr constraints are irrelevant for monomorphemic words, so this time, we will use IDENT-IO [tense][22] instead of IDENT-OO in the tableau in Table 2.20. Since IDENT-IO [tense] outranks SWP, the underlying lax /æ/ surfaces faithfully in ˈc[æ]meo.

/ˈkæmiˌoʊ/	*ə́	IDENT-OO/IO [tense]	SWP
a. ˈkeɪmiˌoʊ		*!	
☞ b. ˈkæmiˌoʊ			*

Table 2.20: *Tableau for cameo, a monomorphemic word with a lax target vowel.*

2.5.2 Discussion of re-analysis of CiV Lengthening

I have provided a model of CiV Lengthening that falls out from the interactions of traditional markedness and faithfulness constraints. Nevertheless, my analysis does face a limitation. While my model of CiV Lengthening provides an account of vowel tenseness in the _CiV environment, it cannot predict other vowel qualities such as backness and height. For example, whilst the following target vowels are all tense, why do we get the mid front [eɪ] in Can[eɪ]dian, the high front [i] in Hand[i]lian, and the mid back [oʊ] in cust[oʊ]dial? Orthography is likely to play an important role in determining the backness and height of the target vowel since surface [eɪ, i, oʊ] are most closely

[22] IDENT-OO [tense] and IDENT-IO [tense] are two distinct constraints that may be ranked differently. Here, both of these constraints happen to rank below *ə́ and above SWP. Hence, I placed both if them in the same column to illustrate that they share the same ranking with respect to the two markedness constraints. However, this should not be taken as a claim that these two IDENT constraints are one and the same.

associated with orthographic <a, e, o> respectively. A fuller account of CiV Lengthening would likely have to integrate orthographic effects into the analysis.

In my proposed analysis, CiV Lengthening is modeled via the interactions of simple markedness and faithfulness constraints. This contrasts with a more traditional derived environment effect analysis of CiV Lengthening (Steriade, 2019)[23]. In constraint-based models such as Optimality Theory (Prince and Smolensky, 1993/2004), derived environment effects have been analyzed via constraint conjunction (Lubowicz, 2002) or comparative markedness (McCarthy, 2003). Both constraint conjunction and comparative markedness increase the complexity of learning the relevant constraints of one's language. In the case of constraint conjunction, the child needs to learn which constraints should be conjoined. Comparative markedness constraints are more complex than traditional markedness constraints. While traditional markedness constraints only look at the output[24] in order to assess constraint violations, comparative markedness constraints need to look at both the output and the input in order to assess violations. In addition, both of these approaches also multiply the number of constraints whose rankings need to be learned.

Nevertheless, simplifying the constraint machinery required to model CiV Lengthening does not come for free. Under my analysis, the featural representation increases in complexity. More specifically, the [tense] feature goes from being a binary-valued feature ($+/-$) to being a ternary-valued feature ($+/-/0$). In the case of reduced vowels, I have argued for the [0tense] status of [ə], given that it is a neutralized vowel in which the tenseness distinction is lost. That is to say, given a surface [ə], the underlying tenseness value cannot be recovered. Regarding non-reduced vowels, I forward that non-reduced (*i.e.* "full") vowels may only bear [$+/-$tense] (*i.e.* not [0tense]) because tenseness is contrastive in English. When affixation with a CiV Lengthening pushes stress onto [ə], [ə] must be repaired to a non-reduced vowel since stress is incompatible with [ə]. The

[23]Tableaux illustrating Steriade's analysis of CiV Lengthening are presented in Table 2.25 and Table 2.26. To differentiate between the tense-lax Trisyllabic Shortening pairs, Steriade uses the length contrast (rather than the tenseness contrast). This is a minor operational detail that has no bearing on the constraint violations and consequently the winning candidates.

[24]Here I use "output" to refer to the candidates in a tableau. The "input" refers to the representation in the top-left cell of a tableau. The "input" may be variously an underlying representation (as in the tableau in (7)) or the BASE+suffix (as in the tableaux in (4-6)).

ternary-valued [tense] allows for [ə] ([0tense]) to be faithful to neither [+tense] nor to [−tense]. Since faithfulness is unable to decide between the [+tense] and [−tense] full vowels, the winning [+tense] vowel (*e.g.* ˈCan[ə]da∼Caˈn[eɪ]dian) provides a rare opportunity to see the low-ranking SWP picking the winning candidate. In other words, for ə-stem words like *Canada*, the ternary-valued [tense] feature renders the higher-ranking faithfulness constraint impotent, thus enabling the oft-hidden preference for heavy stressed syllables to emerge.

As alluded to in §2.5.1, my analysis of CiV Lengthening, which treats this phenomenon as a case of the emergence of the unmarked Weight-to-Stress Principle can be formalized in at least two ways. The first "[0tense]" formalization treats [ə] as [0tense]; the second formalization uses the IDENT-V faithfulness constraint. In the second "IDENT-V" option, the IO and OO versions of IDENT-V would replace their repective IO and OO versions of the IDENT [tense] constraint. Tableaux illustrating how the "IDENT-V" formalization works for the affixed words *Chadian*, *Ukrainian*, *Canadian*, and *cameo* are shown in Table 2.21 through Table 2.24.

[ˈtʃæd] + ian_{suff}	*ə́	IDENT-OO/IO-V	SWP
a. ˈtʃeɪdiən		∗!	
☞ b. ˈtʃædiən			∗
c. ˈtʃədiən	∗!	∗	∗

Table 2.21: IDENT-OO-V *tableau for Chadian, a full-voweled lax-stem.*

[ˌjuˈkɪɛm] + ian_{suff}	*ə́	IDENT-OO/IO-V	SWP
☞ a. ˌjuˈkɪɛmiən			
b. ˌjuˈkɪæniən		∗!	∗
c. ˌjuˈkɪəniən	∗!	∗	∗

Table 2.22: IDENT-OO-V *tableau for Ukrainian, a full-voweled tense-stem.*

[ˈkænədə] + ian_{suff}	*ə́	IDENT-OO/IO-V	SWP
☞ a. kəˈneɪdiən		∗	
b. kəˈnædiən		∗	∗!
c. kəˈnədiən	∗!		∗

Table 2.23: IDENT-OO-V *tableau for Canadian, a reduced-voweled ə-stem.*

The faithfulness constraint IDENT-V is only satisfied when the vowel in the "input" is identical

/ˈkæmiˌoʊ/	*ə́	IDENT-OO/IO-V	SWP
a. ˈkɛɪmiˌoʊ		∗!	
☞ b. ˈkæmiˌoʊ			∗

Table 2.24: IDENT-IO-V *tableau for cameo, a monomorphemic word with a lax target vowel.*

to the vowel in the "output"[25]. In short, under the IDENT-V formalization, vowel quality is retained unless doing so would result in a stressed [ə]. For *Chadian*, *Ukrainian*, and *cameo*, retaining the vowel quality of the target vowel in the BASE or the underlying form does not violate undominated constraint against stressed [ə], so the vowel quality of the target vowel is retained. For *Canadian*, however, the vowel quality of the target vowel in the BASE is [ə]. Affixation would shift stress onto [ə], except that doing so violates the the undominated constraint against stressed [ə]. To avoid having a stressed [ə], the vowel quality changes. However, the next highest ranked constraint, IDENT-OO-V, is unable to decide between the tense [eɪ] and the lax [æ] because neither of these vowels are identical to [ə]. The emergent markedness constraint, Stress-to-Weight, decides in favor of the tense [eɪ], in order to satisfy its requirement that stressed syllables must be heavy.

It should be noted that the [0tense] and the IDENT-V formalizations of my emergence of the unmarked SWP analysis are very similar. The ranking of the faithfulness constraint with respect to the markedness constraints are the same in both formalizations: *ə́ ≫ IDENT [tense] ≫ SWP in the former and *ə́ ≫ IDENT-V ≫ SWP in the latter. This falls out from IDENT [tense] under the [0tense] featural system and IDENT-V performing exactly the same work. If we compare the IDENT [tense] version of a tableau with its IDENT-V counterpart, we see that both of these constraints assign exactly the same violations.

One difference between the [0tense] formalization and the IDENT-V formalization lies in where the complexity increases. The "IDENT-V" formalization does not suffer from the increase in the complexity of the featural representation that the "[0tense]" formalization does. Nevertheless, the IDENT-V constraint can be argued to be more complex than its IDENT [tense] counterpart. In order to assess constraint violations, the IDENT [tense] constraint only needs to compare between

[25]Here, I'm using "input" and "output" loosely. In the Output-Output version of this constraint, the "input" would be the BASE, and the "output" would be the candidates for the derived form. In the Input-Output version of this constraint, the "input" would be the underlying form and the "output" would be the candidates for the surface form.

the singular [tense] feature of the input and the output. If there is a mismatch, then a violation is assigned; otherwise, no violation is assigned, and the assessment terminates. Segments can be thought of as a shorthand for a bundle of features that are relevant for English vowels. Hence, V is shorthand for the 5-7[26] vowel quality features. In order to assess constraint violations for an input-output pair, the IDENT-V constraint needs to compare between the input and output values of each of these 5-7 vowel features in turn. Once a mismatch is found, the process may terminate; however, if no mismatch is found, the process only terminates when all 5-7 vowel features have been considered. In other words, compared to the "[0tense]" formalization, the "IDENT-V" formalization trades-off the increase in the complexity of the featural representation (from a binary to a ternary [tense] feature) for a slight increase in the complexity of the faithfulness constraint (from a faithfulness constraint that is concerned with only one feature to a faithfulness constraint that is responsible for 5-7 features).

Let us take stock of the two analyses and four formalizations discussed so far. CiV Lengthening is a phenomenon that has been considered by previous authors (Chomsky and Halle, 1968; Baković, 2013; Steriade, 2019) to be a derived environment effect. To model a phonological phenomenon as a derived environment effect, complex constraint machineries such as Lubowicz's conjoined constraints or McCarthy's comparative markedness constraints are required. I have proposed a novel analysis of CiV Lengthening, in which I have re-analyzed it as the emergence of the unmarked Weight-to-Stress Principle. I have provided two ways to formalize my emergence of the unmarked analysis. Both of these formalizations also feature an increase in complexity (when compared to conventionally accepted feature values and faithfulness constraints), albeit on a smaller scale. For the "[0tense]" formalization, the 0 value was introduced for the [tense] feature. For the "IDENT-V" formalization, the IDENT-V constraint, being concerned with all vowel quality features, is more complex than the run-of-the-mill IDENT-$\mathcal{F}$ constraint, where $\mathcal{F}$ refers to a single feature.

CiV Lengthening is a complicated phenomenon, and we have seen that all four of the formalizations discussed require the introduction of new material. This new material could come in the

[26]Depending on how one counts.

form of new constraint machinery, 0-valued features, or the IDENT-V constraint. I forward that the increase in complexity caused by the introduction of new material is smaller in the emergence of the unmarked analysis than in the derived environment analysis. I base my argument on the observation that the former introduces new material that is essentially a change in **quantity** while the new material introduced in the latter is a **qualitatively** different than conventional phonological constraints.

Let us first consider the two formalizations associated with the emergence of the unmarked analysis. Phonological features are conventionally binary-valued. Adding the 0 value to the [tense] feature makes it ternary-valued, so the "[0tense]" analysis presents an increase in the quantity of feature values. Likewise, going from a conventional IDENT-$\mathcal{F}$ faithfulness constraint to the IDENT-V constraint just means that the faithfulness constraint has to now account for multiple features rather than just a single feature.

We will now consider the two formalizations associated with the derived environment analysis. Lubowicz's conjoined constraints require the conjunction of a markedness and a faithfulness constraint. Constraint conjunction does not exist in the original formulation of Optimality Theory, so the introduction of conjoined constraints ushers in an entirely new class of constraints. In a similar vein, McCarthy's comparative markedness constraints are also a new class of constraints. Conventional markedness constraints look only at the output while the privilege of looking at both the input and the output belongs to the faithfulness constraints. That is to say, conventional markedness constraints judge how well-formed a sound (or sound sequence) is by looking only at the output. Conventional markedness constraints do not care about the derivational history[27] of any particular sound (or sound sequence). In short, conventional markedness constraints do not need to look at the input in order to judge the well-formedness of an output. In contrast, McCarthy's comparative markedness constraints need to look at the input in order to determine whether a sound (or sound sequence) in the output is well-formed. Thus, McCarthy's comparative markedness constraints are also qualitatively different than the conventional markedness constraints.

To illustrate how McCarthy's comparative markedness constraints work to produce derived environment effects, the tableaux for the derived word *Canadian* and the non-derived word *cameo* are shown in Table 2.25 and Table 2.26 respectively. In these tableaux, the crucial work is done by

the comparative markedness constraint $_{N}$SWP. The subscript in $_{N}$SWP indicates that a markedness violation needs to be "new" in order for an input-output pair to pick up a violation of $_{N}$SWP. In other words, an output only violates $_{N}$SWP if the SWP violation is present in a particular syllable of the output, but absent from its corresponding syllable in the input. In the tableau[28] for the derived word *Canadian* (Table 2.25), the second syllable in the BASE (*i.e.* [ˈkænədə]) vacuously meets the requirement of the Stress-to-Weight Principle since it is a stressless syllable. Candidate (2.25b)

[ˈkænədə] + *ian*$_{suff}$	$_{N}$SWP	IDENT-OO/IO [tense]
☞ a. kəˈneɪdiən		*
b. kəˈnædiən	*!	
c. kəˈnədiən	*!	

Table 2.25: *Comparative markedness tableau for Canadian, a reduced-voweled ə-stem.*

[kəˈnædiən] and candidate (2.25c) [kəˈnədiən] do not meet the requirement of the Stress-to-Weight Principle because they have a stressed syllable that is light. Hence, these two output candidates pick up a violation of $_{N}$SWP since the SWP violation is present in their derived forms, but absent from their corresponding syllables in the BASE. These two candidates are eliminated because they violate the undominated $_{N}$SWP constraint, leaving the path for the tense [kəˈneɪdiən] to win.

The tableau for the non-derived word *cameo* is shown in Table 2.26. As with the *Canadian*

/ˈkæmiˌoʊ/	$_{N}$SWP	IDENT-OO/IO [tense]
a. ˈkeɪmiˌoʊ		*!
☞ b. ˈkæmiˌoʊ		

Table 2.26: *Comparative markedness tableau for cameo, a monomorphemic word with a lax target vowel.*

tableau in Table 2.25, the candidate with the lax [æ] (*i.e.* candidate (2.26b) [ˈkæmioʊ]) does not meet the requirement of the Stress-to-Weight Principle because it has a stressed syllable that is light. However, unlike *Canadian*, the corresponding syllable in the underlying form /ˈkæmioʊ/ violates the Stress-to-Weight Principle because it has the stressed light syllable [ˈkæ]. Since the

[28] In this tableau, the input consists of the BASE+*ian*, and the output candidates are derived forms.

[28] Or the BASE in cases where the "input" is another surface form.

SWP violation is present in corresponding syllables of both the input and the output, candidate (2.26b) [ˈkæmioʊ] does not violate the undominated $_{\text{N}}$SWP, so lax vowels can surface faithfully in non-derived words.

Somewhat interestingly, the $_{\text{N}}$SWP constraint fails to produce the correct surface form for *Chadian* (Table 2.27). As with the *Canadian* tableau in Table 2.25, the candidates with the lax vowels[29]

[ˈtʃæd] + *ian*$_{suff}$	$_{\text{N}}$SWP	IDENT-OO/IO [tense]
☞ a. ˈtʃeɪdiən		∗
☜ b. ˈtʃædiən	∗!	
c. ˈtʃədiən	∗!	

Table 2.27: *Comparative markedness tableau for Chadian, a full-voweled lax-stem.*

[æ, ə] (*i.e.* candidate (2.27b) [ˈtʃædiən] and candidate (2.27c) [ˈtʃədiən]) do not meet the requirement of the Stress-to-Weight Principle because they have a stressed syllable that is light. Here, the corresponding BASE syllable [ˈtʃæd] meets the requirement of the Stress-to-Weight Principle because the stressed syllable is heavy. Since the SWP violations are present only in the outputs and not in their corresponding BASE syllables, candidate (2.27b) [ˈtʃædiən] and candidate (2.27c) [ˈtʃədiən] both violate $_{\text{N}}$SWP. At this juncture, a problem arises. The rightful surface form [ˈtʃædiən] violates the undominated $_{\text{N}}$SWP constraint, so it loses to [ˈtʃeɪdiən].

It turns out that under a formalization of CiV Lengthening that uses the $_{\text{N}}$SWP constraint, the predicted winning candidates for *Can*[eɪ]*dian* and *Ch*[æ]*dian* will either both be tense or both be lax. This is because the violation profiles of *Canadian*'s and *Chadian*'s three candidates (*i.e.* The candidates where the target vowels are [eɪ, æ, ə]) are identical. A schematic tableau showing the violation profiles of the three candidates is offered in Table 2.28. To illustrate why the $_{\text{N}}$SWP constraint is violated for the candidates with the lax target vowels [æ, ə], let us consider the case of the candidate whose target vowel is [æ]. For the output [... .ˈ(C)æ. ...] (*e.g.* [kə.ˈnæ.di.ən, ˈtʃæ.di.ən]), the syllable that contains the target vowel violates the SWP because it is stressed but light (the rime consists of a lax vowel and lacks codas). For *Canadian*, the input syllable with

[29] I am assuming the most basic featural system for this analysis. Under the most basic featural system, the [tense] feature is binary-values, so both [æ, ə] are [−tense].

	$_{N}$SWP	IDENT-OO/IO [tense]
a. .ˈeɪ.		*
b. .ˈæ.	*!	
c. .ˈə.	*!	

Table 2.28: *Schematic tableau for Canadian and Chadian, using the comparative markedness constraint $_{N}$SWP.*

the target vowel [ˈkæn̲ə̲də] vacuously[30] satisfies the SWP because this syllable is stressless. For *Chadian*, the input syllable with the target vowel [ˈt̲ʃæd̲] satisfies the SWP because it is a stressed syllable that is also heavy. For both *Canadian* and *Chadian*, the [... .ˈ(C)æ. ...] candidate picks up a violation of $_{N}$SWP because the the SWP violation is present in an output syllable but absent from its corresponding input syllable. Having *Can*[eɪ]*dian* and *Ch*[æ]*dian* share essentially identical tableaux results in a constraint ranking conflict. *Can*[eɪ]*dian* wants the candidate with the tense [eɪ] to win, necessitating the constraint ranking: $_{N}$SWP $\gg$ IDENT-OO [tense]. However, *Ch*[æ]*dian* wants the candidate with the faithful lax [æ] to win, which requires the opposite constraint ranking: IDENT-OO [tense] $\gg$ $_{N}$SWP.

2.5.3 Conclusion

Strictly speaking, only words like *Canadian* and *cameo* form the core of the derived environment phenomenon within CiV Lengthening. The comparative markedness constraint $_{N}$SWP is able to produce the correct surface forms when the world consists of only these two words, which shows the success of comparative markedness constraints at modeling derived environment effects.

We have also seen in §2.2.5 that the tenseness retention effect exemplified by *Chadian* and *Ukrainian* is an important part of the CiV Lengthening landscape and that this effect is synchronically active. Unfortunately, with the inclusion of *Chadian*, a constraint ranking conflict arises between *Canadian* and *Chadian*.

This suggests that the derived environment view of CiV Lengthening is illusory and dependent on a restricted set of data. When more data (*e.g. Chadian*) is included, the derived environment

[30]The Stress-to-Weight Principle requires that stressed syllables be heavy. It makes no requirement of stressless syllables.

view gives way to an emergence of the unmarked view.

Under the emergence of the unmarked view, the default is for the vowel quality of the input (*i.e.* the BASE (for derived words) or the underlying form (for non-derived words)) to be retained. Only when the retention of this vowel would result in a banned stressed [ə] does the pressure for stressed vowels to be heavy emerge. This pressure is encoded by the Stress-to-Weight Principle, and its effect is seen when the bad stressed [ə] is repaired to a tense vowel (rather than a lax vowel).

CHAPTER 3

The model

3.1 Introduction

In order to fully acquire language, a child has to learn both the representations and the grammar of her language from observed surface forms. Representations include underlying forms, metrical structures, morphological boundaries within words, *etc.* All representations (save the surface representation[1]) are absent from the observed data that the child receives, and are thus termed hidden structure. In this chapter, I propose a computational model that places the hidden structure(s) and the grammar on equal footing. The learning algorithm that I employ simultaneously adjusts the weights of the parameters that control the hidden structure(s) and the weights of the parameters that control the grammar. This means that neither the hidden structure(s) nor the grammar are privileged as both are learned concurrently.

The model and the learner are introduced in §3.2 and §3.3 respectively. In §3.4.1, a simple Voicing Assimilation data set is used to illustrate the set-up of the training data, the hidden structures (*i.e.* URs) under consideration, and the parameters (*i.e.* phonological constraints and UR constraints). In §3.4.2, a more complex version of the Voicing Assimilation data set is used to demonstrate the generalizaing ability of the trained models. In the rest of §3.4, the properties of the trained models are studied: the ability to generalize like human speakers (§3.4.3), the learned parameter weights (§3.4.5), the learned lexicon (*i.e.* the URs learned for each MORPHEME in the lexicon; §3.4.6), and the learned grammar (§3.4.7). §3.6 summarizes the modeling results.

In §3.5, I propose an interpretation of the modeling results that takes into account the prop-

[1] I use the surface representation as a proxy for the pronounced form that a child hears.

erties of the set of "best" trained models as a whole. Namely, I hypothesize that the distribution of qualitatively different analyses amongst global maxima is a model of inter-speaker variation. In the vast majority of phonological modeling, where the task is to model only the UR-SR mapping, generalization tasks are redundant (Goldwater and Johnson, 2003). In §3.7, I discuss why generalization tasks become relevant when modeling hidden structure.

3.2 Model

The knowledge whose acquisition will be investigated is knowledge of a particular distribution over WORD-UR-SR triples (*e.g.* <CROC-PL, /kɹɑk+z/, [kɹɑks]>: 49%; <CROC-PL, /kɹɑk+z/, [kɹɑkz]>: .003%; <CROC-PL, /kɹɑk+s/, [kɹɑks]>: .002%; ...; <DOG-PL, /dɑg+z/, [dɑgz]>: 49%; <DOG-PL, /dɑk+z/, [dɑks]>: .002%; ...). WORD represents a sequence of morphemes, and morphemes are represented with uppercase letters. WORD is also abbreviated WD. The probability of a triple can be rewritten as:

$$Pr(WD, UR, SR) = Pr(SR|WD, UR) * Pr(WD, UR) \tag{3.1}$$

The first term, $Pr(SR|WD, UR)$, is the probability of an SR for a given WORD-UR pair, and is determined by the traditional phonological constraint grammar. For instance, if Pr([bæŋks]|BANK-PL, /bæŋk+z/) = 0.9, then we should interpret it to mean that the WORD-UR pair <BANK-PL, /bæŋk+z/> is realized as SR [bæŋks] 90% of the time. The model proposed here does not condition the UR-SR mapping on the word. Using the example above, this means that Pr([bæŋks]|BANK$_1$-PL, /bæŋk+z/) = Pr([bæŋks]|BANK$_2$-PL, /bæŋk+z/) = Pr([bæŋks]|BANK$_3$-3SG.PRES, /bæŋk+z/), where BANK$_1$ is the financial institution concept, BANK$_2$ is the river concept, and BANK$_3$ is the concept of turning at an angle[2]. Consequently, $Pr(SR|WD, UR)$ = $Pr(SR|UR)$, and the probability of the WORD-UR-SR triple can be simplified to equation (3.2):

$$Pr(WD, UR, SR) = Pr(SR|UR) * Pr(WD, UR) \tag{3.2}$$

Such probabilistic mappings of SRs conditioned on URs (*i.e.* $Pr(SR|UR)$) are computed by virtually all probabilistic constraint-based grammars (*e.g.* probabilistic OT, probabilitic versions of Harmonic Grammar, *etc.*) The current study uses a MaxEnt model, which is a weighted constraint grammar.

Following the traditional phonological MaxEnt model, each UR-SR pair (x, y) is associated with a feature vector, $\vec{v}(x, y)$, which captures the pair's properties. For UR-SR pairs, there are two classes of relevant properties. The first class concerns the form that the SR takes. For example, a feature may be used to track how many pairs of adjacent obstruents of an SR have different voicing values. Such features are known as markedness constraints. The second class of features concerns the mapping between the UR and the SR, and are most commonly used to penalize any changes between the two. These features are conventionally known as faithfulness constraints. Each feature has an associated weight, and the feature weights can be organized into the weight vector $\vec{w}$. The features of the UR-SR pair (x, y) are linearly combined (as in equation (3.3)) to produce its harmony score, $h(x, y)$. $h(x, y)$ is essentially the weighted sum of the UR-SR pair (x, y)'s features, and is a scalar (rather than a vector). Notice the negative sign in equation (3.3). This is because a UR-SR pair is active for a phonological constraint when it **violates** the requirements of that constraint. This in turn should **reduce** a pair's conditional probability (following convention where phonological constraint weights must be non-negative), hence the negative sign.

$$h(x, y) = -(\vec{w} \cdot \vec{v}(x, y)) \tag{3.3}$$

[2]In cases where apparently identical underlying forms get pronounced differently, as in /lif+z/∼[livz] 'leaves' (plural noun) and /lif+z/∼[lifs] 'leafs' (3rd-person singular verb, as in 'she leafs through the book'), it is not because of a difference in the morphemes LEAF1 *vs.* LEAF2. Rather, it is because of some systematic structural difference, like /lif+z/='leaves' is also tagged as being a noun, and the fricative-voicing constraint only applies to nouns, or because of some exceptionality mechanism (*e.g.* /lif+z/='leaves' bears a diacritic that makes it exceptionally susceptible to some voicing constraint), or because the UR of 'leaves' is actually /livz/, *etc.* (The reader is free to imagine their favorite theory.) My examples will avoid these complications and so I wil not need to commit to a mechanism; it only matters that whatever the mechanism is, the information that differentiates LEAF1+PL from LEAF2+3.SG reside in the UR.

The MaxEnt model then maps each pair's harmony score to its probability (equation (3.4)).

$$Pr(SR = y|UR = x) = \frac{e^{h(x,y)}}{Z(x)} \tag{3.4}$$

Since the traditional phonological MaxEnt grammar is a conditional ("discriminative") model, the partition function $Z(x)$ sums over all UR-SR pairs that share the same UR (equation 3.5).

$$Z(x) = \sum_{y' \in \mathcal{Y}_x} e^{h(x,y')} \tag{3.5}$$

In equation (3.5), $\mathcal{Y}_x$ is the set of all SRs that are compatible with UR x. This has the effect of normalizing the probability of a particular UR-SR mapping among only all other mappings from the same UR.

The second term in equation (3.2), $Pr(WD, UR)$, is the joint probability of a WORD-UR pair. Take for instance English voicing assimilation. The child notices that adjacent obstruents agree in voicing. So for a word like DUCK, the surface sequence [ks] could have arisen from any of the following UR sequences {/k+s/, /k+z/, /g+s/, /g+z/}. To learn that DUCK ends in /k/ and the plural morpheme is /-z/, we'd expect to see a higher probability for Pr(/dʌk+z/, DUCK-PL). At the same time, Pr(/dʌk+s/, DUCK-PL), Pr(/dʌg+z/, DUCK-PL), *etc.* have to be very much lower. For the morpheme DUCK, the learner needs to choose between 2 possible stem-final segments: voiceless /k/ and voiced /g/. For the plural morpheme, the learner needs to choose between voiceless /s/ and voiced /z/. Consequently, there are four potential URs that the learner considers for the word DUCK-PL (Table 3.1). Table 3.1 also shows the four features for each of the four variants that the learner has to choose among. These features represent the strength of association between a particular morpheme and an aspect (*e.g.* morpheme-final obstruent voicing) of its UR. Within phonology, such features are also known as UR constraints (Zuraw, 2000; Boersma, 2001). Similar to its UR-SR counterpart, there is a feature vector $\vec{u}(w, x)$ for each WORD-UR pair (w, x). Likewise, the UR constraint weights can be organized into a vector $\vec{\theta}$. The harmony score for each WORD-UR pair is computed as per equation (3.6). Notice that there is no negative sign in equation (3.6). This is because a WORD-UR pair is active for a particular UR

WORD	**UR$_{\text{WORD}}$**	(DUCK, /dʌk/)	(DUCK, /dʌg/)	(PL, /-s/)	(PL, /-z/)
DUCK-PL	/dʌk+s/	1	0	1	0
	/dʌg+s/	0	1	1	0
	/dʌk+z/	1	0	0	1
	/dʌg+z/	0	1	0	1

Table 3.1: *UR constraints for the word DUCK-PL.*

constraint when it contains the morpheme, segment, *etc.*, required by that constraint. That is, a pair is active for a constraint when it **meets** the requirements of that constraint[3]. This in turn **increases** the pair's probability when the UR constraint has a positive weight.. Hence the sign difference between equations (3.3) and (3.6).

$$g(x, y) = \vec{\theta} \cdot \vec{u}(x, y) \tag{3.6}$$

The harmony score of a WORD-UR pair is then mapped to its probability (equation 3.7).

$$Pr(WD = w, UR = x) = \frac{e^{g(w,x)}}{Z} \tag{3.7}$$

In contrast to the UR-SR model described above, the WORD-UR model is not conditional. The normalization takes place over all WORD-UR pairs (equation (3.8)).

$$Z = \sum_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}_w} e^{g(w,x)} \tag{3.8}$$

In equation (3.8), $\mathcal{W}$ is the set of words, and $\mathcal{X}_w$ is the set of all URs that are compatible with word w. This normalization produces a generative distribution over WORD-UR pairs, which in turn produces the generative distribution over WORD-UR-SR triples of equation (3.2). This departs from the models in Staubs and Pater (2016) and Nazarov and Pater (2017), which are discriminative models. A generative model is capable of describing differences in the frequencies of various words, in addition to the relationship between words and their realizations, whereas a discriminative model only does the latter.

[3]The weights of the UR constraints were allowed to be in the range $(-\infty, \infty)$. The weights of UR constraints

3.3 Learning

The observed data that the model takes in is a set of WORD-SR **pair** frequencies (*e.g.* {<DUCK-PL, [dʌks]>: 50; <DUCK-PL, [dʌkz]>: 0; ...}). Meanwhile, the model produces a probability distribution over WORD-UR-SR **triples** (*e.g.* <DUCK-PL, /dʌk+z/, [dʌks]>: 99%; <DUCK-PL, /dʌk+z/, [dʌkz]>: .003%; <DUCK-PL, /dʌk+s/, [dʌks]>: .002%; ...). So how to get the model's output (triples) to be compatible with its input (pairs)? Notice that the triple probability defined in Section 3.2 implicitly defines a distribution over WORD-SR pairs as well. We just need to sum over URs for each WORD-SR pair. More concretely, the probability of pairs can be computed from the probability of triples via this summation:

$$Pr(WD = w, SR = y) = \sum_x Pr(WD = w, UR = x, SR = y) \tag{3.9}$$

$$= \sum_x Pr(SR = y|UR = x) * Pr(WD = w, UR = x) \tag{3.10}$$

$$= \sum_x \frac{e^{h(x,y)}}{Z(x)} * \frac{e^{g(w,x)}}{Z} \tag{3.11}$$

The learner's goal is to train a model that produces the best fit to the observed data, which consists of the frequencies of WORD-SR pairs. The likelihood function

$$L(\theta|WD, SR) = \prod_i \prod_j Pr(WD_i, SR_j|\theta)^{fr(WD_i, SR_j)} \tag{3.12}$$

is a function of the parameter vector θ, which is a combination of the parameters $\vec{w}$ and $\vec{t}$ that govern the WORD-UR and UR-SR mappings respectively. The likelihood function takes the parameter vector θ as an argument, and is a measure of how well the parameter weights fit the fixed data. In order to find the values in θ that produce a trained model that fits the observed data well as possible, the learner traverses the search space that is defined by the likelihood function. The best

should be interpreted as follows: A positive weight for (PL, /-s/) means that it is good for the plural morpheme to be underlying /-s/. A negative weight for (DUCK, /dʌk/) means that is is bad for the morpheme DUCK to be underlying /dʌk/.

solution (*i.e.* best fit to the observed data) occurs at the highest likelihood. Thus, learning consists of finding the maximum likelihood estimate of θ.

In practice, the log-likelihood

$$\ell(\theta|WD, SR) = \sum_i \sum_j fr(WD_i, SR_j) ln(Pr(WD_i, SR_j|\theta)) \tag{3.13}$$

is often used in place of the likelihood. Since the individual probabilities are less than one, the product of many such small numbers can quickly result in numerical underflow problems. Using the logarithm allows us to sum the logarithms of the probabilities rather than to take the product of the probabilities. This has the advantage that during the computation, the log-likelihood decreases at a much slower rate than its likelihood counterpart. Since the logarithm is a monotonically increasing function, maximizing the logarithm of a function, f, is equivalent to maximizing f itself.

In Equation 3.12 and Equation 3.13, the term $fr(WD_i, SR_j)$ refers to the frequency of each unique WD-SR pair. Writing the equations in this manner indicates that we iterate through only unique WD-SR pairs, and treat their frequencies appropriately (*e.g.* by multiplying in the observed frequencies in (Eq 3.13)). The alternative is to write these equations at the level of each data point, allowing the (log-)probability of a unique WD-SR pair to be sum/multiplied in for each occurrence. The difference is non-consequential numerically. But presenting the equations in this manner makes the comparison with the "expected frequencies" in Expectation-Maximization much clearer.

As a final note, the learner's sole objective was to seek the values of θ that maximized the log-likelihood. Experimentation showed that regularization terms did not improve performance in fitting to test data that was withheld from training. Furthermore, the lack of regularization would make the results more interpretable. Hence, the learner did not have to balance multiple priorities such as model fit (operationalized via the log-likelihood) with any biases that a regularization term might introduce.

3.3.1 The EM algorithm

In order to assess the values in the parameter vector θ that will be found by the learner, I use the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm is useful when latent (*i.e.* hidden / unobserved variables) are present in the model. In general, latent variable models consist of three important ingredients: observed variables, hidden variables, and parameters. In the model that was defined in §3.2, there were two observed variables (the WORD and the SR), there was one hidden variable (the UR), and the parameters consisted of the weights of the UR constraints that parameterized the WORD-UR mapping as well as the weights of the markedness and faithfulness constraints that parameterized the UR-SR mapping.

As previously mentioned, the goal of the learner is to find the maximum (log-)likelihood solution (*i.e.* finding the values in the parameter vector θ that produce the highest log-likelihood). In this case, we'd like to maximize the log-likelihood in Equation 3.13. However, direct analytical maximization of the log-likelihood function is generally impossible when latent variables are present. The EM algorithm provides a numerical method to traverse the solution space in search of a local maximum.

The EM algorithm consists of two steps: the Expectation-step and the Maximization-step. Recall that the model produces a probability distribution over WD-UR-SR triples while the observed data consists of frequencies of WD-SR pairs. During the Maximization-step, the complete data (*i.e.* the complete frequencies over WD-UR-SR triples) must be present in order for maximization to take place. However, such data is unavailable from the observed data because the latent variable (*i.e.* the UR) is unobserved. Thus, the role of the Expectation-step is to fill in this "missing data" by making an educated guess.

The goal of the Expectation-step is to produce the expected value of the log likelihood function, $Q(\theta|\theta^t)$. This term represents the log-likelihood function when the "missing data" is filled in by the best-guess efforts given our current parameter weights, θ^t, at the current time, t. Similar to the likelihood function we have previously seen in (Eq 3.13), this log likelihood function, $Q(\theta|\theta^t)$, is a function of the parameter vector θ. The term θ^t refers to the fixed values in the parameter vector θ at the current time t. Thus, the expected value of the log likelihood function, $Q(\theta|\theta^t)$, depends

on the current parameter values θ^t. The expected frequencies of each of the WD-UR-SR triples go into the computation of the expected value of the log likelihood function, $Q(\theta|\theta^t)$. The expected frequency of the WD-UR-SR triple (w, x, y) is

$$E[w, x, y] = Pr(x|w, y) \times fr(w, y), \tag{3.14}$$

where $Pr(x|w, y)$ is the conditional probability of the unobserved UR instance x given the observed WD-SR pair (w, y). The term $fr(w, y)$ refers to the absolute frequency of the WD-SR pair (w, y). The value of $Pr(x|w, y)$ can be calculated as follows:

$$Pr_{\theta^t}(UR = x|WD = w, SR = y) = \frac{Pr(WD = w, UR = x, SR = y|\theta^t)}{\sum'_x Pr(WD = w, UR = x', SR = y|\theta^t)}. \tag{3.15}$$

Notice how the value of $Pr(x|w, y)$ depends on the parameter vector θ^t, whose values have been fixed at time t. If this were the very first Expectation-step, then the values in θ^t would be randomly initialized. If this were a subsequent Expectation-step, then the values in θ^t would be inherited from the previous Maximization-step. The expected value of the log likelihood function, $Q(\theta|\theta^t)$, is thus calculated:

$$Q(\theta|\theta^t) = \sum_{w \in \text{WD}} \sum_{x \in \text{UR}} \sum_{y \in \text{SR}} E[w, x, y] ln(Pr(w, x, y|\theta^t)), \tag{3.16}$$

where the term $E[w, x, y]$ is the expected frequency of the WD-UR-SR triple (w, x, y) (Eq 3.14) and $Pr(w, x, y|\theta^t)$ is the predicted probability of the (w, x, y) triple given the current parameter weights θ^t.

If we were to compare the likelihood function we are now working with (Eq 3.16) with one that we had originally wanted to optimize (Eq 3.13, re-written in Eq 3.17), we see that they are indeed very similar.

$$\ell(\theta) = \sum_{w \in \text{WD}} \sum_{y \in \text{SR}} fr(w, y) ln(Pr(w, y|\theta)) \tag{3.17}$$

The key difference is that the term $E[w, x, y]$ has now replaced the observed frequency $fr(w, y)$. We are now working with completed data. We have "completed" the data by substituting the

expected frequencies of the WD-UR-SR triples (w for WD, x for UR and y for SR) for the actual observed frequencies of WD-UR-SR triples, which we do not have. These expected frequencies (Eq 3.14) are the actual observed frequencies, $fr(w, y)$, weighted by the conditional probability of the unobserved variable (UR) given the observed variables (WD, SR), $Pr(x|w, y; \theta^t)$. They thus represent the best-guess frequency of each WD-UR-SR triple given the current weights.

The goal of the Maximization-step is to find the values of θ that maximize the new log-likelihood function, $Q(\theta|\theta^t)$

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t). \tag{3.18}$$

That is, we want to find the new parameter weights, θ^{t+1}, that can get us as close to the expected frequencies of the "completed" data as possible. Since we have hallucinated expected frequencies in order to complete the data, there are effectively no hidden variables in $Q(\theta|\theta^t)$. So, we can now find a local maximum in just the same way that we would find a local maximum for a log-likelihood function that does not have hidden data. In other words, learning algorithms that are appropriate for the latter can be used in the Maximization-step.

Once the best parameter weights are found for the current best-guess data, the Maximization-step is complete. The updated parameter weights are then fed back into the Expectation-step, which produces an even better-guess of the WD-UR-SR triple frequency, and an updated $Q(\theta|\theta^{t+1})$. This process continues until a Maximization-step in which no better fit to $Q(\theta|\theta^{t+n})$ can be found. Thus, by interatively hallucinating complete data and maximizing the hallucinated likelihood function, the EM algorithm finds a maximum (log-)likelihood estimate of θ for the marginal distribution $Pr(WD, SR|\theta)$.

To recap, the original incomplete-data log-likelihood function that we really want to optimize (*i.e.* $\ell(\theta|WD, SR)$). However, this log-likelihood function cannot be easily optimized. Nevertheless, by substituting the expected frequency of the WD-UR-SR triples for their unavailable observed frequencies, we can obtain a hallucinated log-likelihood, $Q(\theta|\theta^t)$ that is much easier to optimize. The hallucinated log-likelihood function, $Q(\theta|\theta^t)$, is (1) equivalent to the original log-likelihood function at θ^t, and (2) always upper-bounded by the original log-likelihood function. These two properties allow the learner to jump from point to point in the otherwise inaccessi-

ble solution space of the original log-likelihood function because each Expectation-step puts the learner back on a point in the original log-likelihood function. After arriving at an updated point in the solution space of the original log-likelihood function, a new log-likelihood function is then hallucinated. The upper bound property ensures that the newly hallucinated log-likelihood function cannot have a better (higher) log-likelihood than the original log-likelihood function at all parameter values. Furthermore, the EM algorithm converges because each iteration is guaranteed to improve the log-likelihood (Wu, 1983). Another way to put it, the point that we land on in the original log-likelihood's solution space is always guaranteed to have at least as good a log-likelihood as the previous iteration's point (*i.e.* $\ell(\theta^{t+1}) \geq \ell(\theta^t)$).

A couple of caveats should be noted. First, there is no guarantee that the solution space of the function that we really want to optimize (*i.e.* $\ell(\theta|WD, SR)$) is concave; this means that non-global local maxima are likely to be present. Second, the EM algorithm is only guaranteed to halt learning at a point in the solution space where the gradient is 0. While global maxima certainly have a gradient of 0, non-global local maxima and saddle points also meet the criterion of having a gradient of 0. Both of these considerations indicate that multiple simulations that begin at different starting points in the solution space are needed in order to increase our chances of finding global maxima.

3.3.2 Implementational details

The optimization problem in the Maximization-step was solved using the L-BFGS-B method as implemented in `SciPy`'s `optimize.minimize` function. The L-BFGS-B method (Zhu *et al.* (1997)) was chosen because it allowed the User to specify bounds for parameter weights. This is important in phonology because phonological constraints (*i.e.* features/parameters that model the UR-SR mapping) are not allowed to have negative weights.

In order to speed up the learning process, I wrote the necessary functions to analytically calculate the gradients, rather than relying on `optimize.minimize`'s built-in numerically approximated gradients. Bounds were set to $[0, +\infty)$ for phonological constraints and $(-\infty, +\infty)$ for UR constraints.

3.3.3 Interpretation of learning outcomes for language acquisition (Preliminary)

Notice that $Pr(WD, SR)$ is a marginal distribution. This marginal probability was produced by the summation in equation (3.9). The likelihood function of marginal distributions is not guaranteed to be convex, so each EM run finds a local maximum. I take the highest of these local maxima to identify the predicted outcome of learning. In practice, for all the phenomena that were modeled in this dissertation, the highest of the local maxima always coincided with the global maxima. Hence, this approach was equivalent to sampling the solution space to collect a set of global maxima. This concept will be elaborated further in §4.3.

3.4 English Voicing Assimilation

In English voicing assimilation, adjacent obstruents with different voicing values are resolved with suffixes assimilating their voicing value to that of the stem (Table 3.2). The learner observes that

stem	*singular* *stem*-Ø	*plural* *stem*-PL
'duck'	[dʌk]	[dʌks]
'dog'	[dɑg]	[dɑgz]

Table 3.2: *English voicing assimilation*

surface obstruent sequences agree in voicing, but cannot observe the underlying voicing values of these sounds. For example, the learner observes that the word DOG-PL is realized as [dɑgz]. The [gz] sequences could have arisen from one of four logical combinations: /ks, kz, gs, gz/, none of which are directly observed. The first consonant in the cluster belongs to the stem-final obstruent, while the second comes from the suffix. Hence, the underlying voicing value of both stem-final obstruents and suffixes constitute the hidden structures.

3.4.1 Experiment 1

3.4.1.1 Experimental setup

The input to the learner consisted of WORD-SR pair frequencies (Table3.3). The following words

WORD-SR	*Frequency*
DUCK-PL∼[dʌks]	1
DUCK-PL∼[dʌkz]	0
DUCK-PL∼[dʌgs]	0
DUCK-PL∼[dʌgz]	0
DOG-PL∼[dɑks]	0
DOG-PL∼[dɑkz]	0
DOG-PL∼[dɑgs]	0
DOG-PL∼[dɑgz]	1
COW-PL∼[kaws]	0
COW-PL∼[kawz]	1

Table 3.3: *WORD-SR pair frequencies*

constituted the language: {DUCK-PL, DOG-PL, COW-PL}. Each training example was observed once[4]. Notice that a word here is defined as a sequence of one or more morphemes.

Let us now turn our attention to the candidate set. Both the stem-final obstruent and the suffix consonant constitute two different types of hidden structure. Thus, I considered both the voiced and voiceless variants for both of these segmental positions. This resulted in four URs as potential sources of hidden structure for a WORD containing an obstruent-final stem like DOG-PL. A WORD like COW-PL, which contained a sonorant-final stem, had two URs as potential sources of hidden structure. Likewise for surface forms, I considered the very same voicing variants as for the underlying forms. This resulted in a word like DOG-PL arising from four potential sources (UR), and being realized in four potential ways (SR). This combined to give $4 * 4 = 16$ WORD-UR-SR candidates for the word DOG-PL, and $2 * 2 = 4$ such candidates for the word COW-PL.

The feature set consisted of two types of parameters – (1) phonological constraints that either encoded some property of the SR alone (markedness) or encoded some property of the mapping between UR and SR (faithfulness), and (2) UR constraints that encoded the WORD-to-UR mapping.

[4]In this experiment, even though the model has the ability to incorporate the frequencies of different words, I am assuming here that all words are equally frequent in the learner's input.

There were three phonological constraints: AGREE(voice), IDENT$_{general}$, IDENT$_{stem}$. Definitions of these constraints are presented in (10).

(10) a. AGREE(voice): Adjacent obstruents must agree in voicing. Assess one violation for each pair of adjacent obstruents whose members have different voicing values.

b. IDENT$_{general}$: A segment in the UR must be identical to its corresponding segment in the SR. Assess one violation for each segment that does not meet this criterion.

c. IDENT$_{stem}$: A stem segment in the UR must be identical to its corresponding segment in the SR. Assess one violation for each stem segment that does not meet this criterion.

The lone markedness constraint, AGREE(voice), was included because all observed forms always had their adjacent obstruents agreeing in voicing. Moving on to faithfulness constraints, notice that for a word like DOG-PL, a single source like /dɑgs/ may map to four potential realizations [dɑks, dɑkz, dɑgs, dɑgz]. Hence at least two constraints were needed in order to differentiate the four possible mappings. Whilst the obvious way to encode the mapping would be to assign one feature to encode a changing stem-final segment and the other feature to encode a changing stem segment (Table 3.4), this isn't the most phonologically sound. Instead, I use the features IDENT$_{stem}$ and

/...k$_{stem}$+s$_{suffix}$/	IDENT$_{stem}$	IDENT$_{suffix}$
[...ks]		
[...kz]		1
[...gs]	1	
[...gz]	1	1

Table 3.4: *Tableau for distinguishing UR-SR mapping with separate features targeting stem & suffix changes.*

IDENT$_{general}$, which are equally capable of differentiating the four mapping possibilities (Table 3.5) and have phonological backing. IDENT$_{general}$ represents a general dispreference for discrepancies between the UR and the SR. This general constraint is believed to be learned before more specific ones which make reference to morpheme type. The inclusion of IDENT$_{stem}$ but not IDENT$_{suffix}$

/...k_{stem}+s_{suffix}/	IDENT$_{general}$	IDENT$_{stem}$
[...ks]		
[...kz]	1	
[...gs]	1	1
[...gz]	2	1

Table 3.5: *Tableau for distinguishing UR-SR mapping with 1 feature targeting general changes and a specific one targeting the stem.*

represents the cross-linguistic tendency to preserve properties of stem segments over their suffixal counterparts. While it is possible to include all three of these faithfulness constraints, doing so would in fact over-parameterize the UR-SR mapping, so the final set of faithfulness constraints excluded IDENT$_{suffix}$.

The second type of features were the UR constraints, which concerned the WORD-UR mapping. These constraints encoded particular UR properties of morphemes. Recall that the stem-final obstruent was allowed to vary in voicing, so the morpheme DOG had 2 UR variants /dɑk/ and /dɑg/. Consequently, there were two UR constraints for the morpheme DOG: (DOG, /dɑk/) and (DOG, /dɑg/). Likewise, there were two UR constraints for the morpheme DUCK: (DUCK, /dʌk/), (DUCK, /dʌg/). Since the plural suffix also had two UR variants, the following UR constraints were needed: (-PL, -S), (-PL, -Z).

Let's summarize and relate the features back to the equations in §3.2. The grammar had three phonological constraints, so $\vec{w}$ was 3-dimensional for this language (Eq 3.3). In addition, there were six UR constraints, making $\vec{\theta}$ 6-dimensional (Eq 3.6). The solution space was thus nine-dimensional.

3.4.1.2 Results

Recall that the model consisted of 10 logically possible WORD-SR pairs, of which three were observed in the training data. Each of the three observed pairs {(DUCK-PL, [dʌks]), (DOG-PL, [dɑgz]), (COW-PL, [kawz])} was only observed once. The learner sought the parameter values that maximized the likelihood of the training data. Five settings of the parameters are shown in Table

3.6. I found these by running the EM algorithm from 20 randomly initialized[5] starting points. The likelihood of the training data for each of the five parameter settings is $0.33 \times 0.33 \times 0.33 = 0.33^3$. These five settings have already hit the ceiling likelihood of training data. That is, they matched empirical relative frequencies (almost) perfectly. There are no other parameter settings that would be able provide a much better likelihood.

	Model number				
	2	*7*	*14*	*15*	*17*
AGREE(voice)	24.7	40.9	24.4	33.3	26.5
$\text{IDENT}_{\text{stem}}$	15.0	11.9	12.6	29.4	11.6
$\text{IDENT}_{\text{general}}$	11.9	18.5	12.0	18.5	13.6
(DOG, /dɑk/)	-43.2	-11.9	-30.0	-26.4	-11.6
(DOG, /dɑg/)	0.0	0.0	0.0	0.0	0.0
(DUCK, /dʌk/)	0.0	0.0	0.0	0.0	0.0
(DUCK, /dʌg/)	-16.3	-23.7	-20.5	-33.1	-13.3
(-PL, /-s/)	0.3	14.4	9.5	20.7	24.6
(-PL, /-z/)	91.0	56.8	76.6	83.0	77.4
Pr(DUCK-PL, [dʌks])	0.33	0.33	0.33	0.33	0.33
Pr(DOG-PL, [dɑgz])	0.33	0.33	0.33	0.33	0.33
Pr(COW-PL, [kawz])	0.33	0.33	0.33	0.33	0.33
Likelihood of training data	0.33^3	0.33^3	0.33^3	0.33^3	0.33^3
Log-likelihood of training data	-3.29585	-3.29585	-3.29585	-3.29584	-3.29586

Table 3.6: *Feature weights, probability of observed data, & likelihood of training data from the best five runs (English voicing assimilation).*

Note that in this paper, negative weights were only allowed for UR constraints. Weights for regular phonological constraints were not allowed to be negative. For this experiment, the UR constraint (COW, /kaw/) was excluded from the set of features, since the morpheme COW had only one underlying form under consideration[6]. This resulted in (DOG, /dɑg/) attaining 0 weight,

[5]Initial weights for all eight languages in the present study were drawn from a uniform distribution with range=[0.1, 5) for phonological constraints & range=[0, 100) for UR constraints.

[6]For the purposes of this experiment, excluding the UR constraint (COW, /kaw/) created no issues because it was the only such WORD with one UR under consideration. If multiple such WORDs were in the data set and they each had different frequencies, then the UR constraints of all such WORDs would have had to be included in order to express the differences in the frequencies of these WORDs.

which pushed (DOG, /dɑk/) to a negative weight[7]. Because it is the difference between weights rather than the actual value of the weights that matter, the negative weights do not have any meaningful impact on the results.

These five trained models did not distribute any probability mass to the non-observed WORD-SR pairs. Thus, we still have no evidence of whether my over-arching model is able to generalize the information it had learned to unobserved WORD-SR pairs in an appropriate manner. For example, if the model were trained only on DOG-PL∼[dɑgz] and CRAB∼[kɹæb], we would like to see it generalize by giving high probabilities to the WORD-SR pairs: DOG∼[dɑg] and CRAB-PL∼[kɹæbz]. Generalizing in such a manner would be consistent with humans. For example, if a human knew that the WORD DOG-PL was paired with the SR [dɑgz] and that the WORD CRAB was paired with the SR [kɹæb], then they should (1) know that the WORDs DOG and CRAB-PL exist, and (2) be able to pair these two WORDs with appropriate SRs. In Experiment 1, such generalization was not possible because all WORDs had the same morphemic structure: stem-PL. This was an artefact of the candidate set of WORDs, which subsequently affected the candidate WORD-UR-SR set and the candidate WORD-SR set. In order to produce trained models that generalize by assigning non-zero probabilities to data it was not trained on, I performed Experiment 2.

3.4.2 Experiment 2

3.4.2.1 Experimental setup

As in Experiment 1, the input to the learner consisted of WORD-SR pair frequencies (Table 3.7). Three new stems, CROC, CRAB, ROO, were introduced. The training data now consisted of six WORDs: {DUCK-PL, DOG-PL, COW-PL, CROC, CRAB, ROO}. As with Experiment 1, each

[7]The reader may have wondered why the weights that were learned for the two suffixal UR constraints were always positive. This is probably because the UR constraints were randomly initialized in the range [0, 100), which makes it more probable for these constraints to be randomly initialized with a positive weight. Since only the weight difference between these two suffixal UR constraints matter, both constraints are likely to remain in the positive region while the weight adjustment process takes place. This contrasts with the UR constraints of the stems. When (COW, /kaw/) is excluded from the set of UR constraints, (DOG, /dɑg/) and (DUCK, /dʌk/) have to have a weight of 0 in order for the three words DOGS, DUCKS and COWS to have the same predicted probability.

Group no.	*Description*	*WORD-SR*	*Frequency*
1	Suffixed Expt 1 stems	DUCK-PL∼[dʌks]	1
		DUCK-PL∼[dʌkz]	0
		DUCK-PL∼[dʌgs]	0
		DUCK-PL∼[dʌgz]	0
		DOG-PL∼[dɑks]	0
		DOG-PL∼[dɑkz]	0
		DOG-PL∼[dɑgs]	0
		DOG-PL∼[dɑgz]	1
		COW-PL∼[kaws]	0
		COW-PL∼[kawz]	1
2	Unaffixed new stems	CROC∼[kɹɑk]	1
		CRAB∼[kɹæb]	1
		ROO∼[ɹu]	1
3	Unaffixed Expt 1 stems	DUCK∼[dʌk]	0
		DOG∼[dɑg]	0
		COW∼[kaw]	0
4	Suffixed new stems	CROC-PL∼[kɹɑks]	0
		CROC-PL∼[kɹɑkz]	0
		CROC-PL∼[kɹɑgs]	0
		CROC-PL∼[kɹɑgz]	0
		CRAB-PL∼[kɹæps]	0
		CRAB-PL∼[kɹæpz]	0
		CRAB-PL∼[kɹæbs]	0
		CRAB-PL∼[kɹæbz]	0
		ROO-PL∼[ɹus]	0
		ROO-PL∼[ɹuz]	0

Table 3.7: *WORD-SR pair frequencies*

training example was observed once. However, unlike Experiment 1, the candidate set of WORDs included WORDs that were not observed: {DUCK, DOG, COW, CROC-PL, CRAB-PL, ROO-PL}. In other words, in Experiment 1, the WORD candidate set had the very same members as the set of observed WORDs. In contrast, the WORD candidate set is a proper superset of the set of observed WORDs in Experiment 2. This allows the model to assign some non-zero probability to WORDs it was not trained on (*i.e.* WORDs in the candidate set that had frequency 0), if it is indeed able to generalize.

Notice too that this modification produced an asymmetry in the WORD candidate set for Ex-

periment 2. In Experiment 1, observed WORDs were formed by combining all stems with the plural suffix. Thus, there was the only type of word – the suffixed one. In Experiment 2, observed WORDs were produced as follows: half of the stems were combined with the plural suffix while the other half remained unaffixed. There were thus two types of words in Experiment 2. The asymmetry arose because stems were not fully crossed with word type. This asymmetry should enable to model to transfer what it learned to the unseen WORD candidates. For example, the trained model should (ideally) have learned the following from the suffixed WORDs: (1) the grammar required for voicing assimilation, and (2) the possibility of affixing the plural suffix to a stem. From this, it should (ideally) generalize by assigning non-zero probability to the unseen WORD-SR pairs (CROC-PL, [kɹɑks]), (CRAB-PL, [kɹæbz]) and (ROO-PL, [ɹuz]). Likewise, the model should (ideally) have learned from the unaffixed WORDs that stems can be unaffixed. It should then (ideally) generalize by giving non-zero probability to the unseen words (DUCK, [dʌk]), (DOG, [dɑg]) and (COW, [kaw]).

Now that we have considered the candidate set for WORDs, let's move on the the candidate sets for URs and SRs. There are 4 groups of WORDs to consider here. First, the group consisting of observed suffixed forms (*e.g.* DOG-PL). This group had candidate URs and SRs generated using the same method as in Experiment 1. Second, the group consisting of observed unaffixed forms (*e.g.* CRAB-PL). Only one UR and SR was considered for each of these words. The underlying assumption is that when the morpheme only surfaces with one form and there is no possibility of an alternation having affected its realization (given the lack of affixation), the candidate SR is exactly what is heard (*e.g.* [kɹæb] for CRAB), and the candidate UR is identical to the SR (*e.g.* /kɹæb/) (Prince and Smolensky (1993/2004) on Lexicon Optimization). In other words, there is only one UR-SR pair per WORD for the second group.

The third and fourth groups consist of unseen WORDs. None of the morphemes are new, only the sequences in which they appear. This includes sequences of length 1. The candidate URs for these two groups thus follow from the candidate URs above. For example, the unseen unaffixed WORD, DOG, has two candidate URs /dɑk/, and /dɑg/ which it inherits from the word DOG-PL, which had the same candidate URs for its morpheme DOG. Similarly, it inherits the two SRs [dɑk] and [dɑg]. The fourth group, the unseen affixed WORDs, likewise inherit their candidate URs from

their observed counterparts. Take for instance the WORD CRAB-PL. This WORD contains two morphemes, CRAB and the plural suffix. It inherits the lone UR candidate /kɹæb/ for CRAB, and the two UR candidates /-s/ and /-z/ for its plural suffix. This produces $1 * 2 = 2$ candidate URs for CRAB-PL: /kɹæbs/ and /kɹæbz/. Regarding the SRs for these unseen WORDs, I allowed stem-final obstruents and the suffix to vary in voicing. The WORD DOG had thus two candidate SRs [dɑk], and [dɑg]. Similarly, four candidate SRs were considered for CRAB-PL: [kɹæps], [kɹæpz], [kɹæbs] and [kɹæbz].

Recall that the feature set consisted of two types of constraints: phonological constraints and UR constraints. The phonological constraints were the same three used in Experiment 1. In contrast, for the UR constraints, I included at least one UR constraint for each morpheme even if there was only one UR candidate available for that morpheme. This was for the sake of completeness. For example, the morpheme DOG has two UR candidates /dɑk/ and /dɑg/, so there were two UR constraints associated with this morpheme: (DOG, /dɑk/) and (DOG, /dɑg/). The morpheme CRAB, which only has one UR candidate, /kɹæb/, has only one UR constraint associated with it: (CRAB, /kɹæb/).

3.4.2.2 Results

The training candidate set consisted of 28 WORD-SR pairs, of which only six were observed. As with Experiment 1, the learner sought parameter values that maximized the likelihood of the six training data points. The learning problem was more difficult than in Experiment 1. In this experiment, $\frac{6}{28} = 0.214$ of the possible candidate WORD-SR pairs were observed while in Experiment 1, $\frac{3}{10} = 0.3$ of the possible such pairs were observed. Thus the number of random initializations was increased to 50 from 20. Of these 50 runs, 9 tied for equal-highest likelihood. Given that there were six unique WORD-SR pairs that were each observed once, the theoretical ceiling likelihood was $6 \times ln(\frac{1}{6}) = -10.751$. The highest likelihood found was -14.909. In order to attain the ceiling likelihood, the predicted distribution needs to match the distribution at training. Not meeting the ceiling likelihood meant that some probability was assigned to WORD-SR pairs that were not part of the training data. From Table 3.8 we can see that half the probability mass was assigned to the

training data: $0.0833+0.0833+0.0833+0.0833+0.0833+0.0833 = 0.5$. We also see that each of

	Model number								
Training example	*3*	*4*	*14*	*22*	*27*	*29*	*39*	*41*	*49*
DUCK-PL∼[dʌks]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
DOG-PL∼[dɑgz]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
COW-PL∼[kawz]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
CROC∼[kɹɑk]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
CRAB∼[kɹæb]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
ROO∼[ɹu]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
Log-likelihood of training data	-14.912	-14.909	-14.909	-14.910	-14.909	-14.910	-14.909	-14.909	-14.909

Table 3.8: *Predicted probability of training data from trained models that tied for equal-highest likelihood.*

the training WORD-SR pairs are equally probable – they each receive one-twelth (0.0833) of the probability. We know that the shape of the predicted distribution over training data matches that of the training distribution (*i.e.* all training WORD-SR pairs are equiprobable in both distributions); but half the probability mass is missing.

So where did the missing probability mass go? Table 3.9 shows the unobserved candidate WORD-SR pairs that received probability mass[8] in the predicted distribution together with their predicted probabilities. We see that the missing probability mass was assigned to the unaffixed

	Model number								
Unobserved WORD-SR pairs	*3*	*4*	*14*	*22*	*27*	*29*	*39*	*41*	*49*
DUCK∼[dʌk]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
DOG∼[dɑg]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
COW∼[kaw]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
CROC-PL∼[kɹɑks]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
CRAB-PL∼[kɹæbz]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
ROO-PL∼[ɹuz]	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833

Table 3.9: *Predicted probability of unobserved WORD-SR candidates from trained models that tied for equal-highest likelihood. Of 22 unobserved pairs, only those that received appreciable probability mass shown.*

counterpart of trained suffixed stems (*e.g.* training example DOG-PL resulted in some predicted probability assigned to DOG) as well as the suffixed counterpart of trained unaffixed stems (*e.g.* training example CRAB resulted in some predicted probability assigned to CRAB-PL). In fact, all 12 possible WORDs are equiprobable $(1/12 = 0.0833)$.

[8]Technically, all candidate WORD-SR pairs received some non-zero probability in the predicted distribution. This

To recap, Experiment 2 was performed because the trained models produced by the preceding experiment did not generalize in the sense of assigning (appreciable) non-zero probability to unseen WORD-SR pairs. This was concerning because equation (3.2) indicated that the model should produce a joint distribution over WORD-UR-SR pairs; joint models are known to be generative. Modifications were made to the candidate set of WORDs in Experiment 2 such that the model's generalizing ability (if it existed) should be plainly visible. Since half of the predicted probability mass was distributed amongst all six WORDs withheld during training, we can be confident that the model is indeed capable of generalizing by distributing probability mass to unobserved WORD-SR pairs.

Now that we have ascertained that the model is capable of generative behavior, we can move on to ask whether the models in Experiment 1 and Experiment 2 generalize in the same manner as human language learners.

3.4.3 Generalization

If the proposed model is a good model of the concurrent acquisition of hidden structure and grammar, then the behavior of the trained models should mimic that of human language learners. Note however, that not all trained models are equally good at matching the training data. For instance, the log likelihoods found for the 50 trained models of Experiment 2 ranged from a high of -14.909 to a low of -18.961. For now it suffices to state that I take the highest of these likelihoods to be the outcome of learning. The interpretation of this approach for human language learning is further elaborated in §4.3. In practice, multiple runs tie for equal-highest likelihood. In such cases, the set of these runs will be taken to be the outcome of learning. To exemplify, for the trained models of Experiment 2, I test only trained models whose log likelihoods are at (or very close to)

---

is, in fact, expected. Recall that the model produces a probability distribution over WORD-UR-SR triples. This model is a product of two log-linear models. The probability of a WORD-SR pair is obtained by summing over triples that share the same WORD and SR. A log-linear model cannot assign zero probability to an unobserved candidate; the best it can do is to assign it a vanishingly small probability. Since it is not possible for a log-linear model to assign zero probability to any candidate, it follows that no WORD-SR pair should receive zero probability. Nevertheless, the model is still able to give very small probabilities to particular WORD-UR-SR candidates. This in turn results in a good amount of WORD-SR pairs receiving at least 10^4 less probability than the more probable pairs (which receive approximate 0.0833 probability), so these very improbable pairs are excluded from Table 3.9.

-14.909. These models were tested for their ability to generalize to novel data in the way that human speakers do. I do not consider trained models whose likelihood isn't the highest.

Let's familiarize ourselves with how English speakers behave on an alternation task "wug test" when asked to pluralize a novel stem. Adult native speakers of English do not allow a tautosyllabic sequence of consonants to disagree in voicing. To avoid such prohibited consonant sequences, they perform voicing assimilation by changing the voicing value of the suffix to match that of the stem-final consonant(s) (Berko, 1958).

What information is observable from wug tests? Certainly, the SR is observable as it is the produced form. The WORD can also be inferred. Participants are asked to pluralize a particular stem, so we can infer the stem-suffix combination. For example, if we ask them to add a plural to the novel stem WUG, we can infer that the WORD is WUG-PL. In contrast, we don't actually know the UR of WUG-PL. The most we know is that the UR of the stem is /wʌg/, since participants are taught the unaffixed form. We cannot know for certain what the UR of the plural suffix is, since the participant draws on their previously learned underlying representation of the plural suffix, which is not directly observable. Thus, wug tests furnish us with information about the SRs that a human produces for a certain WORD; they do not tell us anything about the WORD's UR.

In order to mimic human behavior well, a trained model has to produce a similar distribution over SRs when presented with a certain WORD, as humans do. For example, a trained model that assigns most of the probability mass to [wʌgz] for WUG-PL is a better trained model than one that assigns only a small probability to it. URs play no role in this assessment.

The candidate WORD-UR pairs for the "wug test" style generalization task is shown in Table 3.10. There were three test words, representing the three types of stem conditions in Experiment 2. The UR is partially known. That is, the UR of the stem is known, as these stems are assumed to be taught to the model, mirroring actual wug tests. In contrast, the trained model has to fill in the UR of the suffix with knowledge that it previously acquired during training. Regarding candidate SRs, I allowed stem-final obstruents as well as the suffix's obstruent to vary in voicing. Hence, each WORD-UR pair which had an obstruent-final stem (*e.g.* CRAB-PL) was associated with four candidate SRs. Each WORD-UR pair which had a sonorant-final stem (*e.g.* ROO-PL) was

WORD-UR
CROC-PL∼/kɹɑk + s/
CROC-PL∼/kɹɑk + z/
CRAB-PL∼/kɹæb + s/
CRAB-PL∼/kɹæb + z/
ROO-PL∼/ɹu + s/
ROO-PL∼/ɹu + z/

Table 3.10: *Candidate WORD-UR pairs under consideration for wug test*

associated with only two candidate SRs.

I consider a trained model to have successfully generalized for a particular test WORD when the "correct" SR attains at least 99% predicted probability. The "correct" SR is the SR that is produced by humans. Human responses in wug tests serve as the ground truth, the standard that the trained models are held up to. For example, a trained model successfully generalizes for WUG-PL when $Pr([\text{wʌgz}]|\ \text{WUG-PL}) \geq 0.99$, because English speakers have been observed to overwhelmingly favor the voiced [gz] sequence for this word. Each of these successful generalizations counts as a "hit". The maximum number of "hits" that a trained model can obtain is three. There were five trained models from Experiment 1 and nine from Experiment 2 that tied for equal-highest likelihood in their respective experiments. Each of these trained models attained "hits" for all three test words. This indicates that all trained models that tied for equal-highest likelihood generalize in a way that mimics human speakers.

These generalization results suggest that the trained models that have successfully generalized to novel stems have learned the very same UR for the plural suffix in addition to the grammar that English speakers have. That is to say, the trained models are successful because they have learned the hidden structure relevant to the pluralization task as well as the grammar needed to apply voicing assimilation. In the following sections, we shall take a closer look at the parameter weights of these trained models in order to better understand exactly what underlying representations and grammars they have learned.

In Experiment 2, the candidate set of WORDs was increased in order to allow the trained models to assign some probability to unseen WORDs. Thus, an additional generalization task was

baked into this enlarged candidate set. For instance, I trained the model on the unaffixed CRAB. Since there was no reason for the stem-final obstruent to alternate, the UR was identical to its realized form: /kɹæb/. The suffixed counterpart, CRAB-PL was observed 0 times, but the model assigned some probability to it. In this way, generalization results can be obtained by looking at whether the "correct" SR obtained at least 99% probability for CROC-PL, CRAB-PL and ROO-PL, WORDs whose stems were only observed in their unaffixed versions in the training set. Each of the nine models that tied for equal-highest likelihood in Experiment 2 attained the perfect number of "hits" (three).

3.4.4 Interim summary

We have seen that the proposed model and learner is able to learn English voicing assimilation. From Experiment 1, we saw that five models were able to replicate the probability distribution that it was trained on. That is to say, these five trained models attained likelihoods of the training data that were at ceiling. In both Experiment 1 and Experiment 2, the trained models that tied for equal-highest likelihood were all able to generalize to novel stems in a way that mimicked English speakers. In the following sections, I present the parameter weights of the trained models that attained equal-highest likelihoods in order to ascertain the grammar(s) and plausible lexicon(s) that were learned.

3.4.5 Parameter weights

The weights attained by the trained models with equal-highest likelihoods for Experiment 1 and Experiment 2 are shown in Table 3.11 and Table 3.12 respectively. At first blush, the five trained models from Experiment 1 seem to be very different, having attained different values for each parameter. The same can be said of the nine trained models from Experiment 2. Nevertheless, identical patterns can be observed across these models.

Within a particular trained model, the weights of specific UR constraints are identical. For example, in Model 3 of Experiment 2, the weights of the following six constraints attained identical weights: (DUCK, /dʌk/), (DOG, /dɑg/), (COW, /kaw/), (CROC, /kɹɑk), (CRAB, /kɹæb/),

	Model number				
Parameter	*2*	*7*	*14*	*15*	*17*
AGREE(voice)	24.7	40.9	24.4	33.3	26.5
$\text{IDENT}_{\text{stem}}$	15.0	11.9	12.6	29.4	11.6
$\text{IDENT}_{\text{general}}$	11.9	18.5	12.0	18.5	13.6
(DOG, /dɑk/)	-43.2	-11.9	-30.0	-26.4	-11.6
(DOG, /dɑg/)	0.0	0.0	0.0	0.0	0.0
(DUCK, /dʌk/)	0.0	0.0	0.0	0.0	0.0
(DUCK, /dʌg/)	-16.3	-23.7	-20.5	-33.1	-13.3
(-PL, /-s/)	0.3	14.4	9.5	20.7	24.6
(-PL, /-z/)	91.0	56.8	76.6	83.0	77.4

Table 3.11: *Trained weights of tied highest-likelihood models (Voicing assimilation Expt 1).*

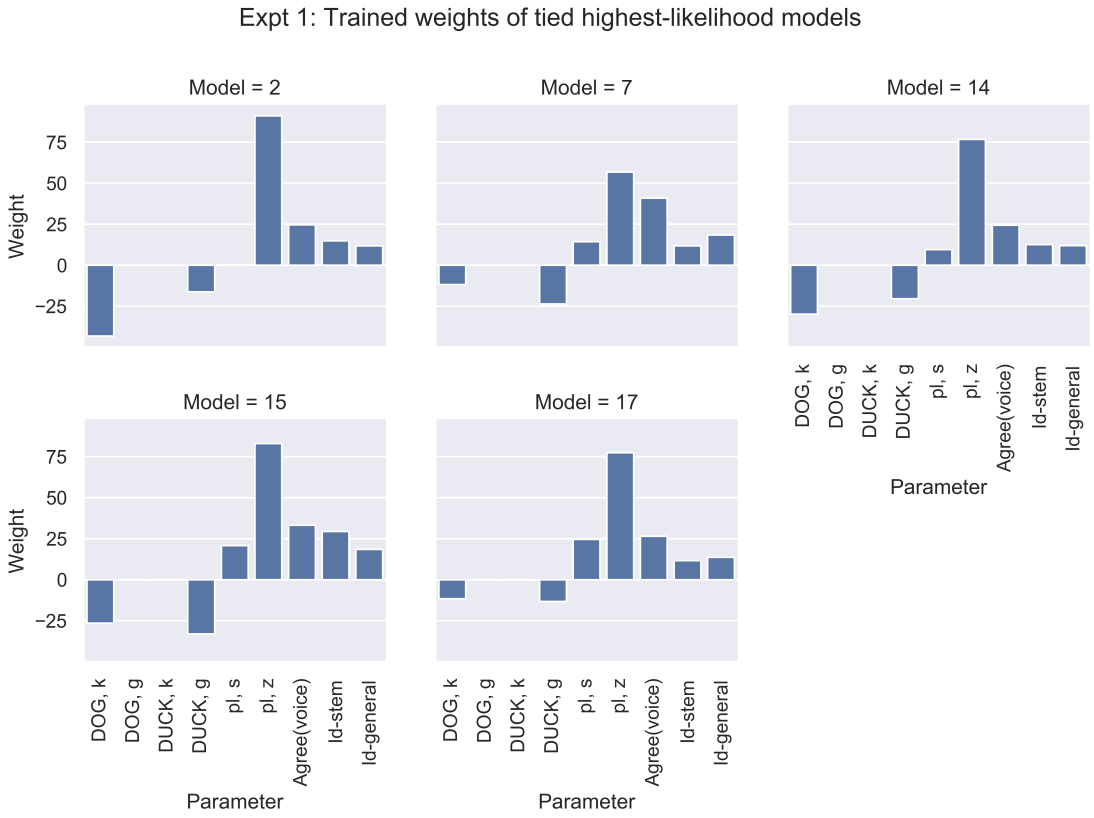

Figure 3.1: *Experiment 1 trained weights –*
(DOG, /g/), (DUCK, /k/), (-PL, /z/) attained higher weights than their counterparts.

(ROO, /ru/). This pattern is obtained by all eight other trained models from Experiment 2 that tied for equal-highest likelihood. In order to understand why this pattern holds, we have to remember that the WORD in which each of these stems occurred was observed once in the training

	Model number								
Parameter	*3*	*4*	*14*	*22*	*27*	*29*	*39*	*41*	*49*
AGREE(voice)	21.0	41.1	25.3	20.6	31.9	27.1	36.8	119.1	43.6
$\text{IDENT}_{\text{stem}}$	10.3	35.8	13.7	10.4	16.6	10.6	21.4	16.0	16.1
$\text{IDENT}_{\text{general}}$	10.3	18.1	13.8	10.4	15.6	13.1	18.8	23.1	14.8
(DOG, /dɑk/)	57.6	11.8	31.3	28.7	49.1	33.8	25 8	58.8	24.1
(DOG, /dɑg/)	68.4	58.0	46.1	40.7	71.4	48.3	53.6	76.2	71.7
(DUCK, /dʌk/)	68.4	58.0	46.1	40.7	71.4	48.3	53.6	76.2	71.7
(DUCK, /dʌg/)	22.5	7.17	13.9	29.4	0.9	20.6	37.9	-100.8	26.5
(COW, /kaʊ/)	68.4	58.0	46.1	40.7	71.4	48.3	53.6	76.2	71.7
(CROC, /kɹɑk/)	68.4	58.0	46.1	40.7	71.4	48.3	53.6	76.2	71.7
(CRAB, /kɹæb/)	68.4	58.0	46.1	40.7	71.4	48.3	53.6	76.2	71.7
(ROO, /ɹu/)	68.4	58.0	46.1	40.7	71.4	48.3	53.6	76.2	71.7
(-PL, /-s/)	-9.2	-18.4	-14.2	-12.8	-53.5	-11.5	-14.3	-71.9	-21.3
(-PL, /-z/)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3.12: *Trained weights of tied highest-likelihood models (Voicing assimilation Expt 2).*

data. Since the model produces joint probability, identical parameter values are required in order to match the observed frequencies.

A similar pattern is observed in Experiment 1. Both (DUCK, /dʌk/) and (DOG, /dɑg/) attained a weight of 0 across all five models. The learned weights in the models of Experiment 1 were more constrained in the values they arrived at than in Experiment 2. This was because the UR constraint (COW, /kaw/) was omitted from Experiment 1, which is the equivalent of setting the weight of this parameter to 0. In order to match the observed frequencies, the weights of (DUCK, /dʌk/) and (DOG, /dɑg/) also had to be 0.

Two of the above stems, DUCK and DOG, are each associated with two UR constraints[9]. For example, (DUCK, /dʌk/) and (DUCK, /dʌg/). The plural suffix, also has two UR constraints. Here, we observe a second pattern emerging. Within each trained model, (DUCK, /dʌk/), (DOG, /dɑg/) and (-PL, /-z/) always attained higher weights than the their counterpart with the differently-voiced morpheme-final segment. Within the proposed model, a high UR constraint weight leads to a more probable outcome for the candidates that are active for this particular UR constraint. The attained UR constraint weights thus indicate that these fourteen models learned

[9]The other four stems only have one UR constraint each.

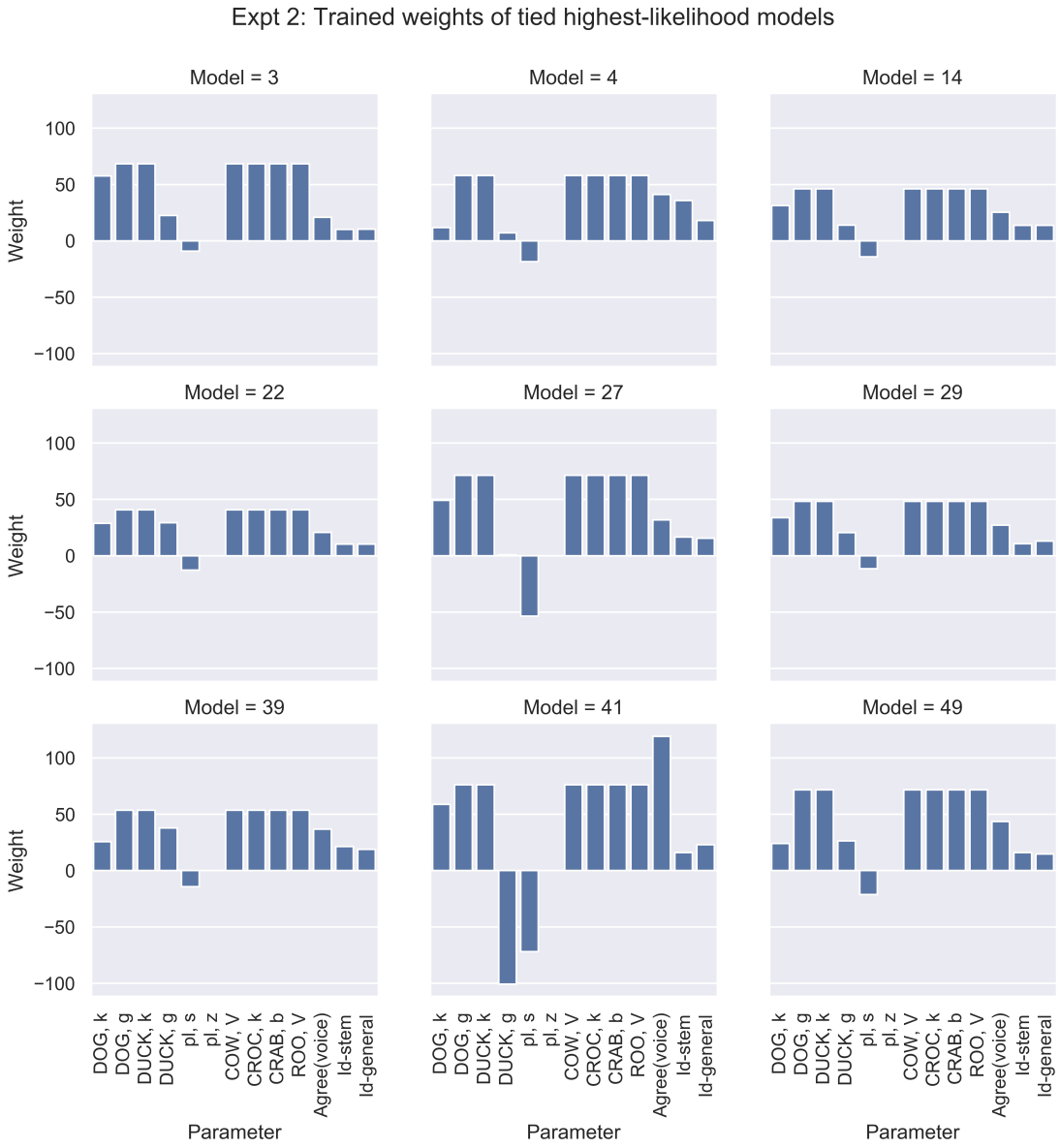

Figure 3.2: *Experiment 2 trained weights –*
(DOG, /g/), (DUCK, /k/), (-PL*, /z/) attained higher weights than their counterparts.*

that the URs of DUCK, DOG and the plural morpheme are more likely to be /dʌk/, /dɑg/ and /-z/ than their counterparts with opposite voicing values.

Even though the equal-highest likelihood trained models from each experiment are at different points in the solution space, similar patterns of parameter weights across the models of a given experiment suggest that these models may have effectively learned the very same morpheme URs (*i.e.* lexicon) and grammar. The acquired lexicons of these models are presented in §3.4.6, and the

grammars in §3.4.7.

3.4.6 Lexicon

In order to determine the UR of a given morpheme (*e.g.* DOG), we need to know how much more probable one morpheme-sized UR variant (*e.g.* /dɑg/) is than the other (*e.g.* /dɑk/). Conventional phonological theory would posit that the morpheme DOG has the UR /dɑg/ rather than /dɑk/. In order to evaluate whether the trained models learned the URs that are posited in traditional phonological theory for the morpheme DOG, we need to calculate $Pr(UR = /\text{dɑg}/ | MORPHEME = DOG)$.

The model proposed in §3.2 does not directly tell us how to calculate the conditional probability we are interested in. Nevertheless, the first log-linear model within the model, which produces a joint probability over WORD-UR pairs, is useful in helping us get to the conditional probability we want. Let's imagine a world in which the WORD, w, has only two unique word-length URs, u_i and u_j. w is composed of n morphemes, of which the URs of $n-1$ morphemes are held constant across u_i and u_j. Let the unique morpheme-sized UR in u_i be called v_i, and the one in u_j be called v_j. The first log-linear model calculates $Pr(w, u_i)$ as well as $Pr(w, u_j)$. Since both WORD-UR pairs share the same WORD, w, the conditional probability $Pr(u_i|w)$ can be calculated as follows: $Pr(u_i|w) = \frac{Pr(w,u_i)}{Pr(w,u_i)+Pr(w,u_j)}$. Note that since both word-length URs arise from the same WORD (*i.e.* the same string of morphemes), v_i and v_j must represent two UR variants of the same morpheme, m. If we imagine $n-1$ to be empty, then $w = m$ and $Pr(u_i|w) = Pr(v_i|m)$.

The UR constraint set was designed such that each UR variant for a given morpheme was coded by a unique parameter. This means that we need only look at all and only the UR constraints that indicate the UR variant of the particular morpheme we are interested in. The generalized equation when there are more than two UR variants is shown in Equation 3.20.

$$Pr(UR_M = i | MORPHEME = m) = \frac{Pr(UR_M = i, MORPHEME = m)}{\sum_{i' \in \mathcal{I}} Pr(UR_M = i', MORPHEME = m)} \quad (3.19)$$

$$= \frac{e^j}{\sum_{j' \in \mathcal{J}} e^{j'}} \quad (3.20)$$

$\mathcal{I}$ is the set of morpheme-sized UR variants of MORPHEME m. $\mathcal{J}$ is the set of UR constraints associated with MORPHEME m. j is the UR constraint (m, i).

To illustrate, trained model 3 of Experiment 2 learned the following weights for the UR constraints associated with the MORPHEME DOG: (DOG, /dɑg) = 68.4, (DOG, /dɑk/) = 59.6. The probability of the MORPHEME DOG having UR /dɑg/ is calcuated as follows:

$$Pr(UR_m = /\text{dɑg}/|MORPHEME = DOG) = \frac{e^{68.4}}{e^{68.4} + e^{59.6}} = 0.99998 \tag{3.21}$$

Let's define successful learning of the lexicon to be attaining the "correct" UR variant for a morpheme at at least 99%. The "correct" UR variant is the one that is posited in traditional phonological analysis. Since each morpheme only had 2 UR variants, the required weight difference between the two relevant UR constraints is just the natural logarithm of the Boltzmaan factor (Equation 3.22). $C_i - C_j$ represents the weight difference between UR constraints, C_i and C_j, needed in order for the UR variant coded by C_i to get probability P_i and the UR variant coded by C_j to get probability P_j. A positive difference indicates $P_i > P_j$ and a negative difference indicates $P_i < P_j$.

$$ln\frac{P_i}{P_j} = ln\frac{e^{C_i}}{e^{C_j}} = ln(e^{C_i - C_j}) = C_i - C_j \tag{3.22}$$

Substituting the minimum requirements for successful UR learning at the lexicon (*i.e.* morpheme) level, the required weight difference comes out to 6.91 (Equation 3.23).

$$\text{Minimum weight difference} = C_i - C_j = ln\frac{0.99}{0.01} = 6.91 \tag{3.23}$$

A quick look at Figures 3.1 and 3.2 confirms that the weight differences between UR constraints for the same MORPHEME indeed exceed 6.91 for all three relevant MORPHEMEs: DUCK, DOG and -PL. More importantly, the UR constraint of each pair that attained the greater weight indeed matches the URs that are posited in traditional phonological theory.

3.4.7 Grammar

For a given set of URs (*i.e.* lexicon), particular weight inequalities need to hold for specific phonological constraints in order for the observed SRs to surface. The set of required inequalities can be thought of as the grammar. The first two required weight inequalities are illustrated in Table 3.13). The markedness constraint, AGREE(voice), needs to outweigh $\text{IDENT}_{\text{general}}$. This makes

/dʌk + z/	AGREE(voice)	$\text{IDENT}_{\text{general}}$	$\text{IDENT}_{\text{stem}}$
[dʌks]		1	
[dʌkz]	1		
[dʌgs]	1	2	1
[dʌgz]		1	1

Table 3.13: *Tableau for DUCK-*PL *with UR /dʌk + z/*

retaining UR voicing values when the UR has adjacent disagreeing obstruents (*i.e.* faithful [dʌkz]) more costly than devoicing the suffix obstruent in order to resolve the voicing conflict (*i.e.* repaired [dʌks]). The second inequality requires $\text{IDENT}_{\text{stem}}$ to be greater than 0. This makes voicing the stem obstruent [dʌgz] a more costly repair than devoicing the suffix obstruent [dʌks]. The third

/kaw + z/	AGREE(voice)	$\text{IDENT}_{\text{general}}$	$\text{IDENT}_{\text{stem}}$
[kaws]		1	
[kawz]			

Table 3.14: *Tableau for COW-*PL *with UR /kaw + z/*

required weight inequality is illustrated in Table 3.14. $\text{IDENT}_{\text{general}}$ needs to be greater than 0 in order for faithful [kawz] to be preferred. The plots in Figure 3.1 and Figure 3.2 show that these three weight inequalities were indeed learned by all the trained models that tied for equal-highest likelihood in Experiment 1 and Experiment 2 respectively.

Here, a couple of notes about the grammar is in order. While the grammar in a strict sense would include every combination and permutation of weight inequalities, the vast majority of these weight inequalities are irrelevant. Since the role of the grammar is to map URs to SRs, the relevance of any particular weight inequality depends on the URs, which are in turn built from morpheme-sized URs (*i.e.* the lexicon). When evaluating whether multiple trained models have learned the same grammar, I therefore include only the relevant weight inequalities in the

definition of the "grammar". For example, the relative weights of the two IDENT constraints are irrelevant given the lexicon. Hence, Model 2 and Model 7 of Experiment 1 are considered to have learned the same grammar despite $\text{IDENT}_{\text{general}}$ outweighing $\text{IDENT}_{\text{stem}}$ in Model 7, and the opposite weight inequality holding in Model 2. Recall that these two constraints are in a subset-superset relationship (any candidate that violates $\text{IDENT}_{\text{stem}}$ also violates $\text{IDENT}_{\text{general}}$). The weight of $\text{IDENT}_{\text{stem}}$ determines how much *more faithful* stems are than affixes, regardless of whether this weight is more or less than that of $\text{IDENT}_{\text{general}}$.

Since the role of the grammar is to map URs to SRs, an argument can be made that the differing magnitude of a given inequality across the trained models means that the probability that a particular UR is mapped to a particular SR differs across models, and thus the learned grammars are all different. To illustrate the argument that the magnitude of the inequality matters, consider the tableaux in Table 3.15. Both the left and right tableaux have the same inequality, where C1

UR	Pred prob. (%)	C1 $w = 3.1$	C2 $w = 3$		UR	Pred prob. (%)	C1 $w = 10$	C2 $w = 3$
SR_1	52.5		1		SR_1	99.9		1
SR_2	47.5	1			SR_2	0.1	1	

Table 3.15: *Magnitude of weight inequality affects distribution of SRs*

outweighs C2. Nevertheless, they produce very different distributions over the two competing SRs, with the left tableau indicating a variational pattern and the right an essentially categorical one. The difference between variation and categoricalness is determined by the magnitude of the inequality. So how do we know that the differing magnitudes across trained models do not have any meaningful impact on the distribution of SRs? The generalization results in §3.4.3 indicate that these equal-highest likelihood trained models all produced identical categorical outcomes with the criterion for categoricalness set at 99%. Categoricalness is attained when the magnitude is sufficiently large. So despite the inequalities across models differing in magnitude, the magnitudes are large enough such that the differences do not make any meaningful impact on the distribution over SRs. That is to say, while the grammars attained by these trained models are not strictly identical, the differences are so minor that these models can be considered to have learned the same grammar.

But why does the difference in magnitude for a given required weight inequality exist at all?

There are two possible contributing factors. First comes the mismatch in categoricalness between the model's input and output. Input data is categorical while the model can only approximate, but not achieve, categorical outcomes. With each EM loop, the model inches ever closer to categorical outcomes by increasing the magnitude of the required weight inequalities. The learning algorithm halts when then improvement to the objective function falls below a certain threshold. The journey towards categorical outcomes is arrested while close to but not quite at the peak, and depends greatly on the steepness of the slope. That is to say, minor differences in predicted probability are expected between runs considered to be at the global maximum due to the way the learning algorithm halts. These minor differences are reflected in the differing magnitude of required weight inequalities. Second, and perhaps more importantly, ever larger magnitudes are required the learner appraoches categorical outcomes. The same change in magnitude that results in a large difference in predicted probability when close to equiprobable outcomes produces only a small difference when close to categorical outcomes. In other words, the minor differences between predicted probabilities at near-categorical outcomes would appear amplified when looking at the differences in feature weights.

Let's take stock of what we have learned from this deep dive into the lexicon and grammar learned by the equal-highest likelihood trained models. Despite the variations in the actual trained parameter weights, all trained models that tied for equal-highest likelihood shared:

- The same lexicon
  - *i.e.* The URs of the set of morphemes needed to model the language.
  - In addition, the lexicon coincided with that posited within traditional phonological theories.
- The same grammar

3.5 Interpretation of learning outcomes for language acquisition

The assumption I work under is that when there is a single global maximum, all learners will converge at that point. If the solution space includes multiple global maxima, then I predict that

the distribution of qualitatively different analyses is distributed amongst the global maxima. More specifically, I hypothesize that the distribution of qualitatively different analyses amongst global maxima is a model of inter-speaker variation. Such inter-speaker variation can be probed with wug tests, which is a type of generalization task.

In order to elucidate the shape of the solution space, I initialized multiple runs for each experiment. The results indicated that the landscape of the solution space featured multiple global maxima. This wasn't unexpected. Recall that the model produced a distribution over WORD-UR-SR triples while the training data consisted of the frequencies of WORD-SR pairs. In order to make the two compatible, the marginal distribution, $Pr(WORD, SR)$, was obtained by summing over the probabilities of WORD-UR-SR triples that shared the same UR (Equation 3.9). The solution space was the likelihood function of this marginal distribution. There is no guarantee that the likelihood function of any given marginal distribution is convex. That is to say, multiple local maxima may exist. Multiple of these local maxima may also tie for the global maximum.

For the simulations of English voicing assimilation in Experiment 1 and Experiment 2, all models that tied for equal-highest likelihood behaved identically in the wug-style generalization task in §3.4.3. Further analysis of the learned lexicon and grammar confirmed that these trained models indeed learned the same lexicon and grammar. I thus conclude that these models all converged on the same qualitative analysis, which was available at multiple points in the solution space, with these multiple points all being global maxima. Following my hypothesis, English speakers are predicted to behave identically on wug tests, producing voiceless [s] for the plural of stems that end with voiceless segments and voiced [z] otherwise.

It might seem that it raises some questions like the following: Does the human learner really run multiple random initializations? Let me explain why this is not really an issue.

It is important to remember that I am **proposing a model** that produces a distribution over WORD-UR-SR triples. To proceed from the output of the proposed model to my hypothesis regarding inter-speaker wug test variation, the following steps were taken:

1. Collect a random sample of the local maxima.

    - Each EM run finds a local maximum.

2. Treat the subset of local maxima that tie for equal-highest likelihood as global maxima.

    - Each global maximum represents a particular qualitative analysis.
    - Analyses are treated as qualitatively different if they produce different generalization results.

3. **Hypothesis**: the distribution of generalization results produced by these qualitatively different trained models should mirror the distribution of generalization results observed across different human learners in generalization task – the wug test.

Notice that my hypothesis concerns only the maxima of the solution space. In other words, it is the landscape of the solution space that matters, rather than the way the algorithm moves around the landscape to land at a local maximum. The multiple random initializations are merely a method to collect a random sample of objects (here: global maxima) that I'm interested in.

I make no claim as to how human learners explore the solution space. What is important is that human learners are able to arrive at a global maximum. Whether each learner find a global maximum by initializing at multiple random points and finding the nearest hill, or via a more efficient way of surveying the space makes no difference for my hypothesis. All that matters is that the learner knows when they have found a grammar that is a global maximum[10].

In short: human learners do not need to run multiple random initializations for my hypothesis to hold. The multiple random initializations were a tool I used to randomly sample the global maxima since the set of global maxima could not be found analytically.

3.6 Summary

The experiments in this chapter have shown that the lexicon and grammar can be learned simultaneously even when the learner had not encountered both the unaffixed and suffixed versions of the same stem. In Experiment 2, the learner encountered both unaffixed singulars and suffixed plurals

[10]Recall that learners know whether they are at (or essentially at) a global maximum because learners know the WORD-SR distribution.

of different stems. In Experiment 1, the learner input to the learner was even more impoverished, with only suffixed plurals. All runs that reached the global maximum were able to generalize in a way that mimicked human learners, and were found to have done so via learning the lexicon and grammars posited in traditional phonological analysis.

3.7 Final remarks

Since global maxima play a crucial role in my hypothesis, it is important to know that the set of local maxima that tie for equal-highest likelihood that I treat as "global maxima" are indeed global maxima. One way to do this is to make the ceiling likelihood attainable (*e.g.* Experiment 1), so we will know whether the highest-likelihood maxima that I am treating as global maxima are indeed the true global maxima. The rest of the phenomena modeled in this dissertation will take this approach so that there will be no uncertainty on whether the "global maxima" we are looking at are indeed true global maxima.

Before moving on, a discussion on the generalization task is in order. A generalization task such as the one performed in §3.4.3 is generally redundant in phonological modeling (Goldwater and Johnson, 2003). Most phonological modeling assumes a fully known UR, so the only unknown that the models need to learn is the grammar – the grammar being the set of phonological constraint weights. In such models, generalization to a test set is redundant because the grammar's performance on the test set would be identical to its performance on the training set, assuming that both training and test sets have been appropriately chosen. In contrast, the present modeling task concerns the WORD-SR mapping via an as-yet-undetermined intermediary: the UR. There are potentially multiple paths from the WORD to the SR via different URs. (*cf.* The UR-SR mapping where only one path exists since there is no intermediary between these two representations.) The WORD-UR mapping is governed by the set of UR constraint weights. Both the WORD-UR mapping and UR-SR mapping interact to produce the distribution we truly want to match: the WORD-SR mapping. Since I leave it to the model to learn which UR[11] to use for each WORD,

[11] or more precisely which distribution over URs

there is no guarantee that the performance of the test set matches that of the training set. In the following chapters, we will encounter cases where trained models tying for the global maximum produce different generalization results; thus, indicating that generalization tasks are very relevant for hidden structure modeling particularly when the presence of a morpheme boundary is in question.

CHAPTER 4

Velar Softening

4.1 English Velar Softening

4.1.1 The phenomenon

English features the following [k]∼[s] alternation: *electri*[k]∼*electri*[s]-*ity*, *electri*[s]-*ize*, *electri*[s]-*ism*. This alternation has been analyzed as Velar Softening, where the underlying velar /k/ is softened to an [s] before [ɪ] or [aɪ] when a morpheme-boundary intervenes. Velar softening is an instance of the derived environment effect (DEE)[1] because its triggering environment requires the presence of a morpheme boundary. For example, the underlying /k/ in the word *kitty* does not alternate to [s]*ity* because no morpheme boundary intervenes between the problematic /k/ and /ɪ, aɪ/. This contrasts with the alternation observed in the suffixed words above.

In an additional wrinkle, velar softening is only triggered by specific morphemes. For example, the *-ity* morpheme triggers velar softening, but the *-ish* morpheme does not, despite both beginning with the same high front vowel [ɪ] (Table 4.1). Suffixes thus fall into two separate categories with

	-ity	*-ish*
electri[k]	*	
electri[s]		*

Table 4.1: *Realization of stem-final /k/ as [k,s] is arbitrary.*

respect to VS-triggering behavior: those that trigger VS, and those that do not. This arbitrary

[1] Although Velar Softening can be characterized as a derived environment effect, the phonological constraints and UR candidates I use in my analysis are more akin to a morpheme-specific phonology analysis. Velar Softening is both phonologically and morphologically conditioned, so it is true that it is not a pure case of DEE. It is also possible that DEEs are not a monolithic concept (*cf.* Chong (2019)).

behavior of English suffixes is not predictable from any other information that is stored in the lexical entry. We have seen that the VS-triggering distinction is not predictable from the UR, since the very same stem *electri*/k/ sometimes undergoes velar softening and does not at other times. Neither is it predictable from the part-of-speech of the stem that the suffix attaches to. For instance, *-ity* and *-ish* both combine with adjectives. Thus, an English learning child needs to store the arbitrary VS-triggering behavior in the lexical entry of each suffix[2].

Velar softening is not driven purely by phonotactics. The [kɪ] and [kaɪ] sequences are legal in English. In the general corpus, the percentage of [kɪ] and [kaɪ] out of the following sequences: [kɪ], [kaɪ], [sɪ], [saɪ], regardless of whether a morpheme boundary is present, is 31.9%. But the percentage of [k] becomes severely under-represented (7.2%) when the same measurement is taken, this time with the restriction that the second segment of the sequence (*i.e.* [ɪ, aɪ]) must belong to a VS-triggering suffix. Examples of the 7.2% include *catechism, monarchist, etc.* These counts were based on type frequency of data from the CMU Pronouncing Dictionary. A χ^2-test showed that the distribution of [k]∼[s] indeed differed in the general corpus *vs.* the restricted case when only VS-triggering suffixes were considered ($\chi^2 = 19.1978$, $df = 1$, $p < .01$). This suggests that: (1) The ban against [kɪ] and [kaɪ] when a VS-triggering suffix is involved is real, and (2) This specific ban should not arise from any general ban against [kɪ] and [kaɪ]. I will model this asymmetry by utilizing a lexically-specified markedness constraint that prohibits the [kɪ] and [kaɪ] sequences if any of its segments arise from a VS-triggering suffix.

Lexically-specified markedness constraints do not apply automatically since they make reference to specific affixes (Alderete 1999; Ito & Mester 1999). Previous work utilizing lexically-specified constraints (*e.g.* Ito & Mester 1999) assumed that lexical information had to be learned before lexically-specified constraints could make reference to them in the grammar. Examples of lexical information that are relevant for velar softening include learning (1) Which suffixes should have a VS-triggering tag?, and (2) What the lexical entry itself should be. That is, should there be a separate lexical entry for *electricity* or is having the stem *electric* and the suffix *-ity* in the

[2]Since only two values are possible (either VS-triggering or not), a single split is enough two distinguish the two behaviors. It may be more economical to record in the lexical entry only the VS-triggering status of VS-triggering suffixes while recording nothing additional for the other suffixes.

lexicon good enough? Both of these questions concern the learning of morphology. Hence, one implication of having such lexically-specified constraints is that the learning of when /k/ becomes [s] goes hand in hand with the learning of morphology. In other words, in order to learn that velar softening is triggered by *-ity*, the child has to first learn that there is a morpheme boundary in *electric+ity*. But in order to learn that *electricity* has a morpheme boundary, it helps to know that velar softening exists.

In order to acquire velar softening, the child has to figure out the following information that cannot be directly observed:

1. Phonological information
    - Is the stem-final consonant underlying /k/ or /s/?
2. Morpho-(phono)logical information
    - Is a morpheme boundary present? *i.e.* Is a word stored as a lexical entry, or should it be assembled from component morphemes?
    - Does a particular suffix trigger velar softening?

In Ch 3 we saw that the proposed learner could learn hidden phonological information and the grammar simultaneously. In this chapter, I extend the learner to simultaneously learn hidden morpho-(phono)logical information alongside the grammar and hidden phonological information.

4.2 The experiment

This section is organized as follows. I describe the training data, potential hidden structure candidates (*i.e.* URs), phonological constraints, and UR constraints in §4.2.1. The determination of which trained models make the criterion for a global maximum model, and the generalization results of the set of global maxima models is presented in §4.2.2 to §4.2.4. In §4.2.5 and §4.2.6, we will take a close look at the grammars and lexicons learned by the global maxima models in order to understand the generalization results.

4.2.1 Experimental setup

4.2.1.1 Inputs to the learner: WORD-SR pairs

As with the previous chapter, the input to the learner consisted of the type frequencies of WORD-SR pairs (Table 4.2). The following words constituted the language: {SONIC, SONICITY, SONICISH, KITTY, SECURE, SECURITY, SMALL, SMALLISH}. The learner observed the unaffixed stem-final [k] in SONIC [sɑnɪk]. Two affixed words with the stem SONIC were provided: SONICITY [sɑnɪsɪɾi], which was meant to be an instance of velar softening; and SONICISH [sɑnɪkɪʃ], which was meant to be an instance in which velar softening did not apply. Notice, however, that the learner does not receive any explicit information that SONIC is related in any way to SONICITY or SONICISH. Neither was there any information regarding morpheme boundaries in the input to the learner. Such information was "hidden", and the learner had to figure these out for itself. To give the learner some clues that *-ity* and *-ish* might be suffixes, I had included the WORDs SECURE & SECURITY as well as the WORDs SMALL & SMALLISH (again, drawing no obvious relationship between these pairs of WORDs in the input). I'd also included the WORD KITTY [kɪɾi], to indicate that the surface sequence [kɪ] was legal in English.

WORD-SR	*Frequency*
SONIC∼[sɑnɪk]	1
SONIC∼[sɑnɪs]	0
SONICITY∼[sɑnɪkɪɾi]	0
SONICITY∼[sɑnɪsɪɾi]	1
SONICISH∼[sɑnɪkɪʃ]	1
SONICISH∼[sɑnɪsɪʃ]	0
KITTY∼[kɪɾi]	1
KITTY∼[sɪɾi]	0
SECURE∼[sɪkjʊɚ]	1
SECURITY∼[sɪkjʊɚɪɾi]	1
SMALL∼[smɑl]	1
SMALLISH∼[smɑlɪʃ]	1

Table 4.2: *WORD-SR pair frequencies (velar softening)*

4.2.1.2 Potential sources of hidden structure: URs

Let us now turn to the candidate set. The candidates consist of WORD-UR-SR triples. Since URs arise from WORDs, we'll have to consider the types of hidden structure for each of the eight WORDs above. Recall that there are three main types of hidden structure:

1. Is there a suffix in the WORD?
2. For a surface [kɪ] or [sɪ] sequence, is the first segment underlying /k/ or /s/?
3. Does a particular suffix trigger velar softening?

For the first type of hidden structure, each surface [kɪ] or [sɪ] sequence was treated as having potentially arisen from either /kɪ/ or /sɪ/. This applied to the WORDs: SONICITY, SONICISH, and KITTY. However, since there is no known process for /s/ to be realized as [k] in English, and the underlying /k/ in KITTY was uncontroversial, I removed UR candidates for KITTY that had underlying /s/ in the interest of speeding up the learning process. Although SONIC itself does not have a surface [kɪ] of [sɪ], it can serve as a potential stem for SONICITY and SONICISH, which do contain these sequences. Thus, I included both /sɑnɪk/ and /sɑnis/ as candidate URs for SONIC. The segmental options for the UR of each WORD is shown in Table 4.3.

The second type of hidden structure concerned the presence or absence of suffixes in the WORD. In other words, can the WORD be analyzed into smaller morphemes? Since I was only interested in segmenting out potential *-ity* and *-ish* suffixes, the relevant WORDs were the ones that ended in ...ITY or ...ISH: SONICITY, SONICISH, KITTY, SECURITY, and SMALLISH. When crossed with the segmental options, this produced the UR candidates in Table 4.4.

The third type of hidden structure was concerned with whether a particular suffix triggered velar softening. This applied to the same set of WORDs concerned with the presence/absence of suffixes (SONICITY, SONICISH, KITTY, SECURITY, and SMALLISH), since these were the only WORDs that could potentially contain suffixes.

As mentioned previously, all sources of hidden structure were crossed to create candidate URs. The full set of UR candidates arising from each WORD is shown in Table 4.5.

WORD	*UR Candidate*
SONIC	/sɑnɪk/ /sɑnɪs/
SONICITY	/sɑnɪkɪɾi/ /sɑnɪsɪɾi/
SONICISH	/sɑnɪkɪʃ/ /sɑnɪsɪʃ/
KITTY	/kɪɾi/
SECURE	/sɪkjʊəˠ/
SECURITY	/sɪkjʊəˠɪɾi/
SMALL	/smɑl/
SMALLISH	/smɑlɪʃ/

Table 4.3: *UR candidates with only segmental options shown*

WORD	*UR Candidate*
SONIC	/sɑnɪk/ /sɑnɪs/
SONICITY	/sɑnɪkɪɾi/ /sɑnɪk+ɪɾi/ /sɑnɪsɪɾi/ /sɑnɪs+ɪɾi/
SONICISH	/sɑnɪkɪʃ/ /sɑnɪk+ɪʃ/ /sɑnɪsɪʃ/ /sɑnɪs+ɪʃ/
KITTY	/kɪɾi/ /k+ɪɾi/
SECURE	/sɪkjʊəˠ/
SECURITY	/sɪkjʊəˠɪɾi/ /sɪkjʊəˠ+ɪɾi/
SMALL	/smɑl/
SMALLISH	/smɑlɪʃ/ /smɑl+ɪʃ/

Table 4.4: *UR candidates with segmental and morpheme boundary options crossed*

WORD	*UR Candidate*
SONIC	/sɑnɪk/ /sɑnɪs/
SONICITY	/sɑnɪkɪɾi/ /sɑnɪk+ɪɾi$_{\text{vs-suff}}$/ /sɑnɪk+ɪɾi/ /sɑnɪsɪɾi/ /sɑnɪs+ɪɾi/ /sɑnɪs+ɪɾi$_{\text{vs-suff}}$/
SONICISH	/sɑnɪkɪʃ/ /sɑnɪk+ɪʃ/ /sɑnɪk+ɪʃ$_{\text{vs-suff}}$/ /sɑnɪsɪʃ/ /sɑnɪs+ɪʃ/ /sɑnɪs+ɪʃ$_{\text{vs-suff}}$/
KITTY	/kɪɾi/ /k+ɪɾi/ /k+ɪɾi$_{\text{vs-suff}}$/
SECURE	/sɪkjʊɚ/
SECURITY	/sɪkjʊɚɪɾi/ /sɪkjʊɚ+ɪɾi/ /sɪkjʊɚ+ɪɾi$_{\text{vs-suff}}$/
SMALL	/smɑl/
SMALLISH	/smɑlɪʃ/ /smɑl+ɪʃ/ /smɑl+ɪʃ$_{\text{vs-suff}}$/

Table 4.5: *Complete set of UR candidates for velar softening*

4.2.1.3 Phonological constraints

The /k/ → [s] process is a UR-SR mapping, so it was parameterized by the phonological constraints defined in (11).

(11) a. *kɪ$_{\text{vsSuff}}$: [kɪ] sequences are disallowed if any one of its segments belongs to an exception-tagged morpheme. Assess one violation for each such sequence.

b. *kɪ: [kɪ] sequences are disallowed. Assess one violation for each such sequence.

c. FAITH: Each segment in the UR must be identical to its corresponding segment in the SR. Assess one violation for each segment that does not meet this criterion.

There were two versions of the Markedness constraint against the [kɪ] sequence. The constraint *kɪ$_{\text{vsSuff}}$ specifically targeted sequences whose segments were associated with an exception-tagged morpheme. The general *kɪ applied to all instances of the [kɪ] sequence. Two constraints were needed because in English Velar Softening, the [k]∼[s] alternation only occurred when specific suffixes were part of the prohibited sequence. Data from the CMU Pronouncing Dictionary (Table 4.6) also provided evidence that two distinct constraints prohibiting the [kɪ] sequence were needed. A chi-square goodness of fit test was performed to determine whether the proportion of [k]-to-[s]

	Sound		
Environment	k	s	***Percentage k out of k+s***
General	2159	4143	34.3
Before *-ity* type suffix	5	64	17.9

Table 4.6: *Distribution of surface [k] and [s] before the [ɪ] vowel.*

was the same in the general environment before [ɪ] versus the specific case in which the [ɪ] came from an *-ity* type suffix. The proportion of [k]-to-[s] did differ according to whether the sound appeared in a general environment before [ɪ] or whether the [ɪ] belonged to an *-ity* type suffix, $\chi^2(1, N = 6371) = 16.8286$, $p < .001$. Specifically, [k] was less likely to occur before [ɪ] when the [ɪ] was from an *-ity* type suffix. Therefore, two separate constraints against the [kɪ] sequence were needed: (1) a lexically-specified *kɪ constraint that only applied when VS-triggering suffixes were present, and (2) a general *kɪ constraint that always applied. A general FAITH constraint was also included to guard against extraneous alternations. The tableau in (Table 4.7) shows the

violations picked up by the UR /sɑnɪk+*ɪti/ ‘sonicity’ for these three constraints.

/sɑnɪk+*ɪti/	*kɪ$_{\text{vsSuff}}$	*kɪ	FAITH
sɑnɪkɪti	1	1	
sɑnɪsɪti			1

Table 4.7: *Tableau illustrating phonological constraints for Velar Softening.*

4.2.1.4 UR constraints

Unobserved URs arise from WORDs. I use UR constraints to parameterize the WORD-to-UR mapping. The WORD-to-UR mapping could be thought of as proceeding via an intermediary: the MORPHEME sequence. That is, a WORD consisted of a sequence of MORPHEMEs, and each MORPHEME was in turn associated with a UR. This meant that two distinct types of UR constraints were needed to parameterize the WORD-to-UR mapping: one type parameterized the WORD-to-MORPHEME sequence mapping; the second type parameterized the MORPHEME-to-WORD mapping.

The presence/absence of boundaries within a WORD was parameterized by the first type of UR constraint, since this was essentially a WORD-to-MORPHEME mapping. WORDs whose morpheme boundaries were in question each had three such UR constraints associated with them:

1. a feature for when the MORPHEME consisted of the entire WORD
2. a feature for a MORPHEME consisting of a potential suffix
3. a feature for a MORPHEME consisting of what remained of the WORD when the suffixed was removed

For example, the WORD SONICITY was associated with the UR constraints: {SONICITY, -ITY, SONIC}. The WORD-UR pair SONICITY∼/sɑnɪk-ɪti/ was active for the UR constraints SONIC and -ITY because it contained these two MORPHEMEs. It was inactive for the UR constraint SONICITY because it did not contain a MORPHEME consisting of the full WORD.

The tableaux in Table 4.8 show which WORD-to-MORPHEME features are active for the WORD-UR pairs associated with SONICITY and KITTY.

		SONIC	SONICITY	K	KITTY	-ITY
WD:SONICITY	/sɑnɪk-ɪti$_{\text{vs-suff}}$/	1				1
	/sɑnɪk-ɪti/	1				1
	/sɑnɪɪs-ɪti$_{\text{vs-suff}}$/	1				1
	/sɑnɪs-ɪti/	1				1
	/sɑnɪkɪti/		1			
	/sɑnɪsɪti/		1			
WD:KITTY	/k-ɪti$_{\text{vs-suff}}$/			1		1
	/k-ɪti/			1		1
	/kɪti/				1	

Table 4.8: *Tableaux illustrating UR constraints parameterizing the WORD-to-MORPHEME mapping for the WORDs: SONICITY, KITTY.*

The second type of UR constraint encodes particular UR properties of a MORPHEMEs (*i.e.* MORPHEME-to-UR type features). We've already encountered this type of UR constraint in Chapter 3. Recall that one type of hidden structure was concerned with whether a particular segment was underlyingly /k/ or /s/. This was relevant for the WORDs: SONIC, SONICITY and SONICISH. Thus, for each of these three WORDs, each WORD-to-MORPHEME feature associated with it sprouted two MORPHEME-to-UR type features. For example, the WORD-to-MORPHEME feature SONIC resulted in the two MORPHEME-to-UR features: (SONIC, /k/) & (SONIC, /s/).

The final type of hidden structure was concerned with whether a particular suffix was exception-tagged. Exception-tagged suffixes trigger velar softening while their untagged counterparts do not. This applied to both suffixes (-ITY and -ISH), since we want the learner to figure out whether each suffix is exception-tagged or not. Hence each WORD-to-MORPHEME type feature for the suffixes was associated with two MORPHEME-to-UR features. For example, the WORD-to-MORPHEME feature -ITY was associated with the exception-tagged MORPHEME-to-UR feature (-ITY, /-ity$_{\text{vs-suff}}$/) and the untagged one (-ITY, /-ity/).

The tableau in Table 4.9 shows the relevant UR constraints for the six URs of the WORD SONICITY.

WD: SONICITY	SONIC	SONICITY	-ITY	(SONIC, /k/)	(SONICITY, /s/)	(SONICITY, /k/)	(SONICITY, /s/)	(-ITY, /-ity$_{\text{vs-suff}}$/)	(-ITY, /-ity/)
/sanɪk-ɪti$_{\text{vs-suff}}$/	1	0	1	1	0	0	0	1	0
/sanɪk-ɪti/	1	0	1	1	0	0	0	0	1
/sanɪs-ɪti$_{\text{vs-suff}}$/	1	0	1	0	1	0	0	1	0
/sanɪs-ɪti/	1	0	1	0	1	0	0	0	1
/sanɪkɪti/	0	1	0	0	0	1	0	0	0
/sanɪsɪti/	0	1	0	0	0	0	1	0	0

Table 4.9: *Tableau illustrating the UR constraints parameterizing the WORD-to-UR mapping for the WORD SONICITY.*

4.2.2 Results

I performed 2000 randomly initialized simulations. The distribution of the final log likelihoods of all 2000 trained models is shown in Figure 4.1. This is a violin plot; broader regions have a greater

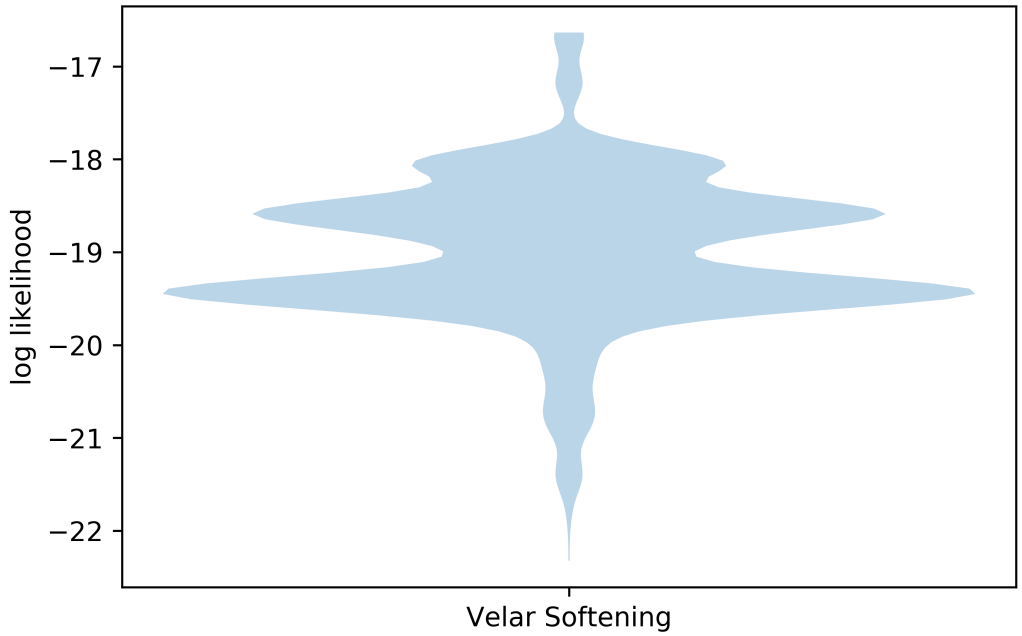

Figure 4.1: *Distribution of all final log likelihoods (Velar Softening)*

density of models. There appears to be a cluster of log likelihoods above the -18 region. To get a closer look at the distribution of above -18 log likelihoods, I perform a finer-grained visualization (Figure 4.2).

Figure 4.2 is a histogram. Each bin width was set to 0.1, with the right-most bin starting at the value of the highest final log likelihood (16.635). The two right-most bins formed a cluster, so I

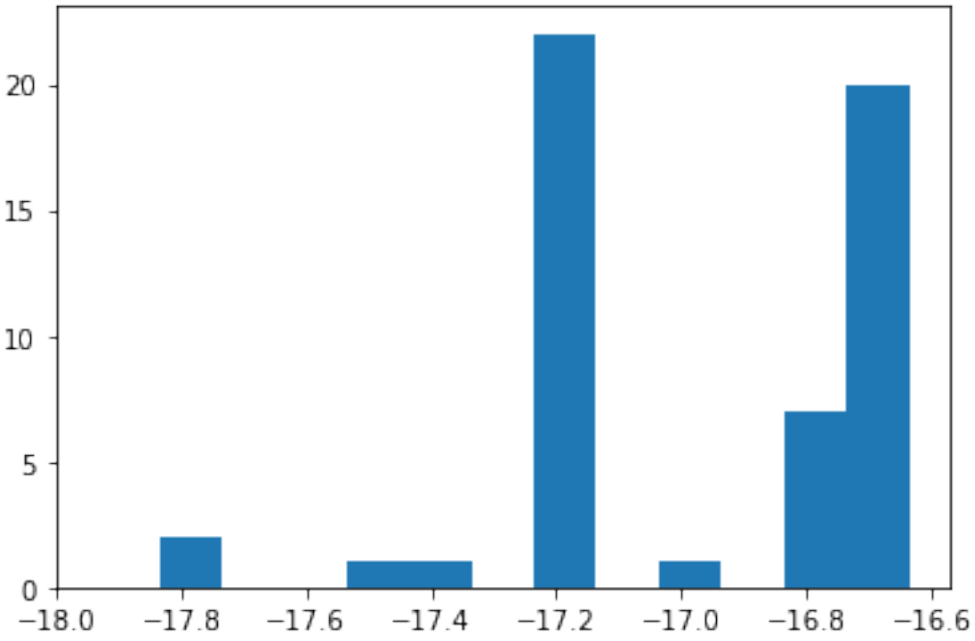

Figure 4.2: *Distribution of final log likelihoods over -18 (Velar Softening).*

considered the runs whose final log likelihood fell above -16.835 ($= -16.635 - 0.1 - 0.1$) as tying for equal-highest final log likelihood. 27 runs met this criterion.

The theoretical ceiling was -16.635 ($= 8 \times log(\frac{1}{8})$), so these 27 runs had indeed reached the ceiling log likelihood. That is, these 27 trained models were able to match the training data perfectly. Recall that the training data consisted of frequencies of WORD-SR pairs. Matching the training data perfectly[3] meant that these 27 trained models were able to produce the correct surface forms[4] for words that they had encountered before. From a linguistics point of view, this model predicts that all humans will be able to learn the Velar Softening surface pattern for words that currently exist in English.

[3]With a small allowance of -0.2 for minor differences in probability.

[4]More precisely, these 27 trained models were able to match the observed $\frac{1}{8}$ probability for each WORD-SR pair (and give essentially zero probability mass to unobserved pairs).

4.2.3 Generalization

While these 27 trained models were able to produce the surface pattern for English Velar Softening, we can still ask whether they are indeed good models of Velar Softening for actual human speakers. Notice that URs are missing from the observed data (WORD-SR pairs) that the full WORD-UR-SR model predicts. These models are free to use any UR that enables them to match the surface pattern (WORD-SR frequencies). Models that utilize different URs might also require different grammars (*i.e.* weights of phonological constraints) to handle the UR-to-SR mapping. So it is theoretically possible for these trained models to have URs and grammars that look nothing like the URs and grammars that we would expect humans to have. It is not possible to know exactly what URs and grammars English speakers have because such information cannot be directly observed. Nevertheless, tests can be designed to probe the grammar and at least some portions of URs.

Consider the following wug task. Pierrehumbert (2006) provided participants with novel stems ending in /k/, and had participants produce the *-ity* suffixed form. Such a wug test probes whether participants had acquired the necessary grammar for changing /k/ to [s] in the presence of a Velar Softening triggering suffix. It also probes whether participants had learned that *-ity* was a Velar Softening triggering suffix.

In Pierrehumbert (2006)'s study, she found that two in 10 participants never applied Velar Softening while the other eight did. That is, two participants never changed the stem-final /k/ to [s] upon *-ity* suffixation. This suggests that English speakers may arrive at different analyses for Velar Softening (*i.e.* UR and grammar combinations) that nevertheless allow them all to produce the same surface Velar Softening pattern for existing words.

I subjected these 27 ceiling likelihood models to an identical wug task. I included only these 27 models because I assume that actual English speakers have the correct surface forms for words that they have heard before like *electricity, kitty, etc.* The models were faced with the complex word CLEMIC-ITY, which consisted of the novel stem CLEMIC and a familiar suffix -ITY. Since the stem was entirely novel, I provided its UR /klɛmɪk/[5]. The models had to fill in what they had learned about the -ITY suffix: Namely, does this suffix trigger Velar Softening? They also had to fill in what they had learned about when to change /k/ to [s]: Always change /k/ to [s]? Only

apply /k/→[s] in the presence of a Velar Softening suffix? Never change /k/ to [s]?

There were four candidate WORD-UR-SR triplets for this wug test (Table 4.10). Each model

WORD-UR-SR
CLEMIC-ITY∼/kləmɪk + ɪti$_{\text{VS-suff}}$/∼[kləmɪkɪɾi]
CLEMIC-ITY∼/kləmɪk + ɪti$_{\text{VS-suff}}$/∼[kləmɪsɪɾi]
CLEMIC-ITY∼/kləmɪk + ɪti/∼[kləmɪkɪɾi]
CLEMIC-ITY∼/kləmɪk + ɪti/∼[kləmɪsɪɾi]

Table 4.10: *Candidate WORD-UR pairs under consideration for wug test (Velar Softening)*

produced a probability distribution over these four candidates. A model was considered to have generalized if the probability of the spirantized SR was at least 99% (*i.e.* $P([\text{kləmɪsɪɾi}]|\text{CLEMIC-ITY}) \geq 99\%$). 21 of the 27 models met this criterion for generalization, changing the /k/ of the novel stem to [s] upon *-ity* suffixation at a rate greater than 99%. The six that did not predicted that the novel stem alternates to [s] upon *-ity* suffixation at a rate smaller than 0.01%.

These 27 ceiling runs represent a random sample of the global maxima. Overall, we have seen that 78% of the sample of global maxima were able to generalize Velar Softening to a novel *k*-final stem. In contrast, the remaining 22% could not perform this generalization. There are thus (at least[6]) two qualitatively different analyses represented in this set of global maxima, with the analysis that allows generalization taking up approximately 78%, and the one that does not taking up approximately 22%. The overall prediction of the model is that 78% of English speakers will extend Velar Softening to a novel *k*-final stem. This prediction finds a parallel in Pierrehumbert's wug study where 80% of English speakers did in fact extend Velar Softening to a novel *k*-final stem.

[5]More specifically, the only UR that I provided for the affixed form was /kləmɪk/, since I did not train the models on stress-induced vowel reduction.

[6]Looking ahead, a careful study of the lexicons and grammars of these 27 models show that there are in fact three distinct analyses. The present wug test is only able to make a two-way distinction between analyses.

4.2.4 Additional generalization tests

In the previous section, I performed a generalization task with the novel *k*-final stem *clemic* and the familiar *-ity* suffix. In traditional phonological analysis, the *-ity* suffix would be exception-tagged because the /k/→[s] process is triggered only by a small minority of suffixes.

By changing whether (1) the stem or the suffix was novel and (2) the exceptional status of the suffix, I produced four logically possible test WORDs (Table 4.11). I subjected the 27 trained

	VS-suffix	No VS-suffix
$\text{stem}_{\text{familiar}}\text{-suffix}_{\text{novel}}$	*sonic-ism*	*sonic-y*
$\text{stem}_{\text{novel}}\text{-suffix}_{\text{familiar}}$	*clemic-ity*	*clemic-ish*

Table 4.11: *Test words (Velar Softening).*

models to the same generalization task with the three new test words. There were two SRs for each WORD: one with the stem-final consonant surfacing as [k] and the other as [s]. For the novel suffixed WORD SONIC-$\text{ISM}_{\text{VS-suffix}}$, the goal was to produce a surface [s]. In contrast, for the test WORDs with normal suffixes (*e.g.* SONIC-Y, CLEMIC-ISH), the goal was to produce a surface [k]. A model was considered to have successfully generalized on a particular test word if it produced the correct SR at a rate $\geq 99\%$. As before, I provided the UR of the novel morpheme, while the model had to fill in what it had learned about the familiar morpheme.

All 27 models produced the correct SR for the test WORDs with normal suffixes. The same six models that failed to generalize to [s] for CLEMIC-$\text{ITY}_{\text{VS-suffix}}$, likewise failed to generalize to [s] for SONIC-$\text{ISM}_{\text{VS-suffix}}$. The remaining 21 models successfully generalized to [s] for SONIC-$\text{ISM}_{\text{VS-suffix}}$. These results are summarized in Table 4.12. We see that the distinction between

Generalized on clemi[s]-ity	*soni[s]-ism*	*soni[k]-y*	*clemi[k]-ish*
Yes	yes	yes	yes
No	no	yes	yes

Table 4.12: *Test words (Velar Softening).*

the two qualitatively different analyses holds for the novel word *sonic-ism*, which like *clemic-ity*, requires a surface [s]. In contrast, the novel words requiring a surface [k] cannot distinguish between the two qualitatively different analyses.

No single test WORD is able to fully probe what the trained models had learned about URs. Instead, each of these four test WORDs probe different portions of the URs. Understanding what URs these models learned will in turn inform us about the grammars that they learned. The $\text{stem}_{\text{familiar}}$-$\text{suffix}_{\text{novel}}$ test words *sonic-ism*$_{\textit{VS-suffix}}$ and *sonic-y* both probe what the model learned about the UR of the familiar (*i.e.* trained) MORPHEME SONIC. Since all models had produced the correct [k] for *sonic-y*, whose novel (*i.e.* given) suffix does not trigger Velar Softening, we can deduce that all models had learned that the UR for the MORPHEME SONIC ended in /k/. This is because there is no markedness constraint that prohibits [s] from surfacing, so there is no grammar available to change /s/ to [k]. Consequently, all surface [k]'s must have arisen from underlying /k/'s.

For the test WORD SONIC-$\text{ISM}_{\text{VS-suffix}}$, the models had to fill in the UR that they had learned for SONIC. They were told that the novel suffix -$\text{ISM}_{\text{VS-suffix}}$ was a VS-triggering one. We have already settled that all models leaned that SONIC had stem-final /k/ underlyingly, so all models must have used the UR /sɑnɪk+*izm/ for *sonic-ism*.

It can be further deduced that the six models that failed to generalize to [s] for SONIC-$\text{ISM}_{\text{VS-suffix}}$ were unable to do so because they lacked the grammar that produced /k/→[s] in the presence of a VS-suffix. We can tell that the lack of generalization can be attributed to the grammar rather than a difference in URs between generalizing and non-generalizing models because (1) all models used the same stem for both test *words sonic-ism*$_{\textit{VS-suffix}}$ and *sonic-y*, and (2) from test *word sonic-y*, we know that all models had learned the very same UR for the morpheme SONIC: /sɑnɪk/.

We now have an inkling that the models that failed to generalize did so because they lacked the grammar to change /k/ to [s] in the presence of a VS-triggering suffix. The generalization results, however, give no insight into the URs that the models learned. The non-generalizing models may well differ from the generalizing ones with respect to their learned URs too. In the upcoming subsections, we'll take a look at the parameter weights of these 27 models in order to gain a better understanding of the grammars and URs that they had learned.

4.2.5 Grammar

Under a traditional phonological analysis, the grammar that produces Velar Softening consists of two weight inequalities. Since the [kɪ] and [kaɪ] sequences are generally allowed in English, a general Faithfulness constraint, FAITH, that promotes the retention of all underlying segmental features should outweigh the general Markedness constraint *kɪ that always bans [kɪ] and [kaɪ] sequences. The tableau for the word *kitty* illustrates this (Table 4.13).

/kɪti/	FAITH	*kɪ
kɪɾi	1	
sɪɾi		1

Table 4.13: *Weight-inequality 1:* FAITH $>$ **kɪ.*

The second weight inequality is important for *word*s like *electricity* and *sonicity*, whose stem-final /k/ is realized as [s] when a VS-triggering suffix is present. In Table 4.14, the stem-final consonant is /k/. The faithful candidate [sɑnɪkɪɾi] violates two Markedness constraints: both the general one (*i.e.* *kɪ) that always bans the [kɪ] and [kaɪ] sequences, and the specific one (*i.e.* *$\text{kɪ}_{\text{vsSuff}}$) that bans these sequences only when one of its segments arises from a VS-triggering suffix.

/sɑnɪk+ɪti$_{\text{vs-suff}}$/	*$\text{kɪ}_{\text{vsSuff}}$	*kɪ	FAITH
sɑnɪkɪɾi	1	1	
sɑnɪsɪɾi			1

Table 4.14: *Weight-inequality 2:* *kɪ_{vsSuff} + *kɪ $>$ FAITH.

I considered a weight-inequality to be learned if (1) the correct inequality was present, and (2) the difference in weights between the expressions on both sides of the inequality was at least 5. The first weight inequality was learned by all 27 global maxima models, generalizing and non-generalizing ones alike. In contrast, the second weight inequality showed a very clear split between generalizing models and non-generalizing ones. All generalizing models learned the second weight inequality while all non-generalizing ones did not. In fact, all non-generalizing models had learned the opposite of the second weight-inequality[7].

[7]They smallest difference between weights was 9.25 (in the wrong direction). The smallest difference between

This result serves both as a sanity check and a confirmation that the non-generalizing models failed to generalize to [s] for the novel word SONIC-ISM$_{\text{VS-suffix}}$ because they indeed lacked the grammar required to do so (Table 4.15).

/sɑnɪk+ɪzm$_{\text{vs-suff}}$/	FAITH	*kɪ$_{\text{vsSuff}}$	*kɪ
sɑnɪkɪzm		1	1
sɑnɪsɪzm	1		

Table 4.15: *Test word: sonic-*ism. Weight inequality 2 missing.*

4.2.6 Lexicon

Recall that the training data consisted of eight unique words. The URs that the traditional phonological analysis would posit for these words is shown in Table 4.16. For the training words *sonic,*

word	*UR*
SONIC	/sɑnɪk/
SONICITY	/sɑnɪk + ɪti$_{\text{vs-suff}}$/
SONICISH	/sɑnɪk + ɪʃ/
KITTY	/kɪti/
SECURE	/sɪkjʊɚ+ɪti$_{\text{vs-suff}}$/
SECURITY	/sɪkjʊɚ
SMALL	/smɑl/
SMALLISH	/smɑlɪʃ/

Table 4.16: *UR of training words under traditional phonological analysis.*

sonicity, sonicish, kitty, security and *smallish*, multiple UR candidates were available for the learner to choose among. The learner's task was to find a distribution over the candidate URs and a corresponding grammar that allowed it to match the observed *word*-SR frequencies as closely as possible. The distribution over candidate URs can be thought of as a window into a model's lexicon. To illustrate, let's imagine a simplified case where the *word sonicity* has two candidate URs: /sɑnɪk + ɪti/ and /sɑnɪsɪti/. If the mono-morphemic UR /sɑnɪsɪti/ takes most of the probability mass, then the UR for *sonicity* is stored as the underived UR /sɑnɪsɪti/ in the lexicon. If however, the multi-morphemic UR /sɑnɪk + ɪti/ takes most of the probability mass, then the UR for *sonicity* is

weights in the correct direction was 9.50.

not stored. Rather it is derived morphologically via the URs for the stem *sonic* and the suffix *-ity*. Hence, for a particular model, the distribution over URs for a given *word* is crucial for understanding whether the model in question treats *word*s that are generally analyzed as morphologically complex (*e.g. sonicity, sonicish*) as wholly memorized forms or whether the model derives such *word*s from component morphemes.

For all *word*s in all 27 ceiling models, the distribution was highly skewed in the direction of at most one UR per training *word*. More specifically, the UR that took the majority of the probability mass for each *word* had a conditional probability, $P(UR|word)$, greater than 95%.

4.2.6.1 Generalizing models

For all generalizing runs, the most probable UR for each training *word* was the one predicted by traditional phonological analysis (Table 4.16). In particular, this UR attained a conditional probability greater than 99% for the training *word*s: *sonic, sonicity, sonicish* and *kitty*.

Traditional phonological analysis posits the UR /sɪkjʊɚ+ɪti$_{\text{vs-suff}}$/ for the training *word security*. 17 of the 21 generalizing models assigned more than 99% conditional probability to this UR while the remaining four assigned probabilities between 97.3-97.4% to it. For these four models, approximately 2.6-2.7% of the probability mass was assigned to the mono-morphemic UR /sɪkjʊɚɪti/.

For the training *word smallish*, traditional phonological analysis posits the UR /smɑl + ɪʃ/. 18 of the 21 generalizing models assigned more than 99% conditional probability to this UR while the remaining three assigned probabilities between 97.3-98.6% to it. For these three models, approximately 1.4-2.7% of the probability mass was again assigned to the mono-morphemic UR /smɑlɪʃ/. These three models are distinct from the four that assigned less than 99% probability to the traditional UR for *security*. Alternatively, some *word*s may remain ambiguous between a derived or a mono-morphemic status. This may be especially so for *word*s with few derived neighbors sharing the same stem or affixes.

In summary, all generalizing models assigned overwhelming probability mass (at least 97%) to the traditional UR. Seven models assigned some probability mass (1.4-2.7%) to the unanalyzed

form for the *word*s *security* and *smallish*. These seven models did not have enough data to push the probability of the multi-morphemic URs for *security* and *smallish* to at least 99%. To do so, the training data would have to include more evidence for the existence of the stems *secure* and *small*, for example by including more multi-morphemic *word*s in the training data like *securely* or *smallness*.

4.2.6.2 Non-generalizing models

We will now turn our attention to the non-generalizing models, where we will see that the UR candidate that gets the highest probability is not always the one posited in traditional phonological analysis. The UR candidates for training *word*s that feature multiple UR choices are shown in Table 4.17. The SR for each *word* is also shown for easy reference.

All six non-generalizing models assigned more than 99% probability to the traditional UR for *sonic* (*i.e.* /sɑnɪk/). This was unsurprising since there was no way in which an underlying /s/ could be realized as [k][8].

The non-generalizing models had the same patterns for the training *word*s *sonicity* and *security*, and also for *sonicish* and *smallish*, so each pair will be discussed in turn. This pairing is rather expected since the *word*s forming each pair have the same "potential" suffix *-ity* or *-ish*.

For *sonicity* and *security*, all non-generalizing models assigned more than 99% probability to the mono-morphemic UR. This corresponds to memorizing full URs (*i.e.* /sɑnɪsɪti/ for *sonicity*; /sɪkjʊɚɪti/ for *security*), rather than using the URs of component morphemes for these potentially complex words. Notice that the UR for *sonicity* has underlying /s/, so no /k/→[s] process is required in these models.

The UR candidate with the highest probability for *sonicish* and *smallish* is indicated in Table 4.18. Each of these UR candidates was assigned a probability greater than 99%. Variants of the UR candidates with underlying /s/ for *sonicish* were omitted because they never achieved the highest probability in any model. Unlike the potentially-complex *-ity word*s above, the models

[8]Recall that there were no Markedness constraints against [s]; hence, no reason to transform underlying /s/ into any other segment.

WORD	SR	*UR Candidate*
SONIC	[sɑnɪk]	/sɑnɪk/ /sɑnɪs/
SONICITY	[sɑnɪsɪɾi]	/sɑnɪkɪɾi/ /sɑnɪk+ɪɾi$_{\text{vs-suff}}$/ /sɑnɪk+ɪɾi/ /sɑnɪsɪɾi/ /sɑnɪs+ɪɾi/ /sɑnɪs+ɪɾi$_{\text{vs-suff}}$/
SONICISH	[sɑnikɪʃ]	/sɑnɪkɪʃ/ /sɑnɪk+ɪʃ/ /sɑnɪk+ɪʃ$_{\text{vs-suff}}$/ /sɑnɪsɪʃ/ /sɑnɪs+ɪʃ/ /sɑnɪs+ɪʃ$_{\text{vs-suff}}$/
KITTY	[kɪɾi]	/kɪɾi/ /k+ɪɾi/ /k+ɪɾi$_{\text{vs-suff}}$/
SECURITY	[sɪkjʊɚɪɾi]	/sɪkjʊɚɪɾi/ /sɪkjʊɚ+ɪɾi/ /sɪkjʊɚ+ɪɾi$_{\text{vs-suff}}$/
SMALLISH	[smɑlɪʃ]	/smɑlɪʃ/ /smɑl+ɪʃ/ /smɑl+ɪʃ$_{\text{vs-suff}}$/

Table 4.17: *UR candidates for velar softening (only words with multiple candidates shown).*

Model	/sɑnɪkɪʃ/	/sɑnɪk+ɪʃ/	/sɑnɪk+ɪʃ$_{\text{vs-suff}}$/	/smɑlɪʃ/	/smɑl+ɪʃ/	/smɑl+ɪʃ$_{\text{vs-suff}}$/
1			✓			✓
8			✓			✓
12			✓			✓
15		✓			✓	
17			✓			✓
27	✓			✓		

Table 4.18: *UR candidate with highest probability for sonicish & smallish (non-generalizing models).*

preferred URs with a morpheme boundary separating the *-ish* suffix from the stem. Five of six non-generalizing models chose a complex UR while a single model went with the mono-morphemic

UR. Recall that none of the generalizing models learned 'Weight inequality 2'. In other words, these models did not learn to change underlying /k/ to an [s] even if this resulted in a [kɪ] or [kaɪ] sequence with one of the segments forming the sequence arising from a VS-triggering suffix. 'Weight inequality 2' was the only one in which VS-triggering suffixes are important, so when a model fails to learn this weight inequality, it essentially makes no distinction between VS-triggering suffixes and regular suffixes. That is to say, VS-triggers might as well not exist in such models. Hence, the non-generalizing models have essentially learned only two URs: /sɑnɪk-ɪʃ/ and /sɑnɪkɪʃ/, with the multi-morphemic /sɑnɪk-ɪʃ/ being more common.

Previously, we saw these models had all learned that the UR for the stem *sonic* had a stem-final /k/. Since underlying /k/ surfaces as [k] for *sonicish*, there is no conflict with using the stem's UR for *sonicish*. This is the likely explanation for the difference in distribution between URs with or without the morpheme boundary for *sonicity* and *sonicish*. There is a general preference for building the UR by utilizing the URs of component morphemes. This is what we see for *sonicish*. However, when the URs of component morphemes cannot be used, the learner instead memorizes a version of the UR without the morpheme boundary. This is the case for *sonicity*, where the non-generalizing models lack the grammar to change /k/ to [s], so they all require the UR /sɑnɪsɪti/ for *sonicity*.

For *kitty*, the non-generalizing models likewise preferred URs with a morpheme boundary (Table 4.19). The checkmark indicates that the most probable UR attained a probability greater

Model	/kɪti/	/k+ɪti/	/k+ɪti$_{\text{vs-suff}}$/
1			✓
8		✓	
12	✓		
15			✓
17			✓
27	*4%*	*96%*	

Table 4.19: *UR candidate with highest probability for kitty (non-generalizing models).*

than 99%. For Model 27, actual percentages are shown since no UR attained greater than 99% probability. With *kitty*, we again observe that models prefer URs with morpheme boundaries. The UR results for *kitty* differ between generalizing and non-generalizing models. The generalizing

models all chose /kɪti/ while the majority of non-generalizing models chose URs with a morpheme boundary. This contrast probably arose because *kitty* is over-parameterized for non-generalizing models in a way that is not so for generalizing ones. All generalizing models shared the general pattern where a morpheme boundary was posited only when there was evidence for both the stem and the suffix. For example, they all learned that *-ity* was a suffix, and chose URs that properly pulled out the *-ity* suffix for all *word*s that indeed had this suffix (*i.e. sonicity* and *security*). The multi-morphemic status of these two *word*s was also supported by evidence that *sonic* and *secure* were stems. In contrast, there was no evidence that the /k/ in *kitty* came from an independent stem. Hence, the generalizing models treated *kitty* differently from *sonicity* and *security*; positing no morpheme boundary in the former's UR but having URs with morpheme boundaries for the latter. In contrast, non-generalizing models did not share this pattern. In fact, they could not. Non-generalizing models had to posit the mono-morphemic /sɑnɪsɪti/ for *sonicity* despite (1) clear evidence that *sonic* was a stem, and (2) some indication that *-ity* was a potential suffix because it appeared in multiple training *word*s. In other words, the non-generalizing models could not use the presence of a morpheme boundary to differentiate between *word*s whose potential stem and suffix were both independently motivated and those *word*s did not have independent evidence for both the potential stem and the potential suffix. Since the presence of a morpheme boundary wasn't used to partition any data in non-generalizing models, it was essentially an unused parameter[9]. It is not too surprising then, that the non-generalizing models have a preference for a morpheme boundary in *kitty*, since this parameter does not do any work in the model.

Indeed, when a particular parameter does no work, but the trained models still prefer one parameter setting over another, we can think of the preferred parameter setting as a property of the model architecture that emerges when the parameter is freed from the burden of partitioning data. As before, we see an emerging preference to posit a morpheme boundary. The evidence for this preference is much clearer with *kitty* than with *sonicish*. Notice that the non-generalizing models do not posit any suffix for *sonicity* or *security*, so the *-ity* suffix does not exist anywhere else in the

[9]In the actual models, the morpheme boundary is indicated by multiple parameters. It would be more precise to refer to the set of parameters that operationalize the "presence or absence of morpheme boundaries". Nevertheless, I refer to the "presence or absence of morpheme boundaries" as a single parameter for the sake of exposition.

language. Neither is there any evidence for the /k/ in *kitty* arising from an independent stem[10]. Despite no evidence for either stem or suffix existing elsewhere in the language, these models still posit a morpheme boundary in *kitty*, thus indicating a clear preference for multi-morphemic URs over mono-morphemic ones.

4.3 Interpretation of learning outcomes for language acquisition

Previously, I had stated that I toke the global maxima to represent the outcome of learning. In this section, I will expound on this and discuss how this can be used to model inter-speaker variation.

When there is a single global maximum in the solution space, all trained models that aren't waylaid by local maxima will converge at the global maximum. This global maximum represents one set of parameter weights, and therefore corresponds to exactly one (morpho-)phonological analysis (*i.e.* one lexicon-grammar combination). If however, the solution space includes multiple global maxima, there is a possibility that each global maximum corresponds to a distinct (morpho-)phonological analysis. To make this concrete, let's imagine that there are three global maxima in the solution space. The number of distinct (morpho-)phonological analyses can be anywhere between one and three. If all three global maxima result in the same lexicon and grammar, then they all represent the same (morpho-)phonological analysis, so the number of distinct analyses is one. If however, the three global maxima all produce different lexicons or grammars, then there are three distinct analyses. I hypothesize that the distribution of distinct (morpho-)phonological analyses amongst global maxima is a model of inter-speaker variation. Such inter-speaker variation can be probed via wug tests, which is a type of generalization task.

In order to elucidate the shape of the solution space, I trained multiple models for each experiment. Each trained model was initialized from a random point in the solution space. For all the phenomena that we have encountered so far (and also those that we will encounter later), the solution space had multiple global maxima. We know that this is the case because multiple trained models tied for the highest-attainable "ceiling" likelihood for velar softening and for Experiment

[10]Contrast this with *sonicish*, whose stem exists as an independent training *word*.

1 of voicing assimilation. For Experiment 2 of Voicing Assimilation, the ceiling likelihood was not attainable, but a closer look the trained models showed that the trained models that tied for equal-highest likelihood were indeed global maxima.

The existence of multiple global maxima is not unexpected. Recall that the model produced a distribution over WORD-UR-SR triples while the training data consisted of the frequencies of WORD-SR pairs. In order to make the two compatible, the marginal distribution, $Pr(WORD, SR)$, was obtained by summing over the probabilities of WORD-UR-SR triples that shared the same UR (Equation 3.9). The solution space was the likelihood function of this marginal distribution. There is no guarantee that the likelihood function of any given marginal distribution is convex. That is to say, multiple local maxima might exist. Multiple of these local maxima might also tie for the global maximum.

For the simulations of English voicing assimilation in Experiment 1 and Experiment 2 of Chapter 3, all global maxima models behaved identically in the wug-style generalization task in §3.4.3. Further analysis of the learned lexicon and grammar confirmed that these trained models had indeed learned the same lexicon and grammar. I thus concluded that these models all converged on the same (morpho-)phonological analysis. This single analysis was available at multiple points in the solution space. Following my hypothesis, all English speakers are predicted to behave identically on wug tests, producing voiceless [s] for the plural of stems that end with voiceless segments and voiced [z] otherwise.

For velar softening, 27 trained models arrived at a global maximum. In §4.2.6 and §4.2.5, I analyzed the lexicons and grammars learned by these 27 models and found that there were six distinct lexicon-grammar combinations. That is to say, these 27 global maxima were distributed amongst six distinct combinations of the lexicon and the grammar (*i.e.* Six distinct (morpho-)phonological analyses). If we allow ourselves to collapse the lexicon-grammar combinations that arise from the over-parameterization discussed in §4.2.6.2, we get three distinct (morpho-)phonological analyses[11]. The distribution of trained models amongst these three analyses is shown in Table 4.20. We can see from Table 4.20 that most of the global maxima correspond to Analysis A. This produces a falsifiable hypothesis, with the prediction that that 77.8% of speakers should have learned the (morpho-)phonological analysis represented by Analysis A, 18.5% should have

	count	*percent*
Analysis A	21	77.8
Analysis B	5	18.5
Analysis C	1	3.7

Table 4.20: *Distribution of global maxima models over distinct (morpho-)phonological analyses.*

learned the (morpho-)phonological analysis represented by Analysis B, and 3.7% should have learned the (morpho-)phonological analysis represented by Analysis C.

To test the hypothesis that the present model is a good model of inter-speaker variation, we would need to verify that speakers had indeed learned the lexicons and grammars in the predicted proportions. Unfortunately, we are unable to directly observe the lexicon and grammars inside speakers' heads. Nevertheless, probes can be designed to distinguish the Analyses from each other. Before moving on to the probes, let's familiarize ourselves with the lexicons and grammars of the three Analyses (Table 4.21). Analysis A differs from Analysis B and Analysis C because the

Analysis	*A*	*B*	*C*
sonic	✓	✓	✓
secure	✓	✓	✓
small	✓	✓	✓
-ity	✓		
-ish	✓	✓	
sonicity		✓	✓
security		✓	✓
sonicish			✓
smallish			✓
weight-inequality for /k/→[s]	✓		

Table 4.21: *Lexicon-grammar combinations represented by Analyses A-C.*

former has the lexicon and grammar to support velar softening while the latter do not. In Analysis A, the *-ity* suffix is tagged as exceptionally triggering velar softening. In contrast, Analysis B and

[11]This collapses the analyses that have the following URs when the Markedness constraint *kɪ$_{\text{vsSuff}}$ is ranked below FAITH such that it makes no sense to have specially indexed VS-triggering suffixes:/sɑnɪk+ɪʃ/ and /sɑnɪk+ɪʃ$_{\text{vs-suff}}$/ for *sonicish*. /k+ɪti/ and /k+ɪti$_{\text{vs-suff}}$/ for *kitty*. It also collapses the analyses with or without a morpheme boundary for *kitty*.

Analysis C appear not to have learned the *-ity* suffix at all. For instance, *sonicity* is composed of the the stem *sonic* and suffix *-ity* in Analysis A, but stored as the separate lexical item *sonicity* in Analysis B and Analysis C. In addition, the weight inequality that drives /k/→[s] in the presence of VS-triggering suffixes was present in the grammar of Analysis A but absent in the grammars of Analysis B and Analysis C. Thus, to distinguish between Analysis A and Analyses B & C, we would need a complex nonce word that has the *-ity* suffix. Speakers that have acquired Analysis A would realize the stem-final /k/ as [s], while those that learned Analysis B or C would realize the stem-final consonant faithfully as [k]. Ideally, we'd have enough probes to distinguish the three Analyses. However, no probe is able to distinguish between Analysis B and Analysis C, which differ only upon the existence of the *-ish* suffix. This suffix does not trigger any suffix-specific phonological processes, which are needed to diagnose whether the suffix is used in a potentially-complex wug word.

Let's summarize the predictions made for velar softening. A probe is available to distinguish Analysis A from Analyses B and C. This probe should be a complex nonce word composed of a stem-final /k/ directly followed by the *-ity* suffix. 77.8% of speakers are predicted to respond by changing the stem-final /k/ to [s], while the remaining 22.2% should realize the stem-final /k/ faithfully. No probe is available to distinguish Analysis B from Analysis C.

The results of Pierrehumbert's study provide supporting evidence for the hypothesis that the present model is a good model of inter-speaker variation in velar softening. Recall that Pierrehumbert performed a nonce word study to test the productivity of velar softening. For /k/-final stems, she found that 80% of participants produced an [s] upon *-ity* suffixation, while the remaining 20% of participants never applied velar softening. This result is very close to the prediction of the distribution of speakers with Analysis A versus those that have Analyses B & C: namely that 77.8% would soften stem-final /k/ to [s], and the remaining 22.2% would not.

To summarize, we have evidence that the present model is a good model of inter-speaker variation in velar softening. Let's recap the procedure used. In velar softening, adult speakers have been observed to have learned the same surface patterns for existing words. However, under the surface, these speakers may indeed have learned different (morpho-)phonological analyses to generate the same surface patterns for existing words. Each distinct (morpho-)phonological analysis consists of

a unique lexicon-grammar combination. By analyzing the lexicon-grammar combinations we can produce a falsifiable hypothesis that predicts the proportion of speakers that should acquire each lexicon-grammar combination. With careful analysis of the different lexicon-grammar combinations, nonce word probes can be designed to distinguish amongst the various combinations. We have evidence that our models of inter-speaker variation are sound when the distribution of human behaviors on these probes mirrors that of the distributions predicted by our models.

Now that we have understood my proposed model of inter-speaker variation, I will address several remaining loose ends. First, I have treated only the trained models that have reached the global maximum as the outcome of learning. For example, I have only analyzed the lexicon and grammars of global maxima models while ignoring the trained models that did not reach the global maximum. The motivation behind this is that speakers generally acquire the surface pattern[12], and the global maximum represents the closest that a trained model can get to replicating the surface pattern.

Second, I have made the assumption that speakers are able to find a global maximum. Nevertheless, I do not believe that human learners utilize the very same learning algorithm that I used to find a global maximum. In this dissertation, I utilized a learner that applied the Expectation-Maximization algorithm with a multitude of randomly initialized starting points to collect local maxima. I then visualized the distribution of local maxima and treated the cluster with the highest log likelihoods as the set of global maxima. As the complexity of the phenomena increased, we saw a decrease in the proportion of trained models reaching the global maximum. If humans do indeed utilize the learning algorithm that I used, they may run into the issue of needing a prohibitively high number of random initializations in order to get even one trained model that reaches a global maximum. Hence, while I stand by my proposed model of inter-speaker variation, I am agnostic as to the learning algorithm used to find the global maxima.

Can I truly retain my model of inter-speaker variation while substituting in another the learning algorithm? Well, it is important to remember that I am proposing a model that produces a

[12]For the phenomena modeled in this dissertation (*e.g.* English voicing assimilation, English velar softening), speakers are known to acquire the surface pattern.

distribution over WORD-UR-SR triples. To proceed from the output of the proposed model to my hypothesis regarding inter-speaker wug test variation, the following steps were taken:

1. Collect a random sample of the local maxima.
    - Each EM run finds a local maximum.
2. Treat the subset of local maxima that tie for equal-highest likelihood as global maxima.
    - Each global maximum represents a particular qualitative analysis.
    - Analyses are treated as qualitatively different if they produce different generalization results.
3. **Hypothesis**: the distribution of generalization results produced by these qualitatively different trained models should mirror the distribution of generalization results observed across different human learners in generalization task – the wug test.

Notice that my hypothesis concerns only the maxima of the solution space. At no point is any reference made to the learning algorithm itself. In other words, it is the landscape of the solution space that matters, rather than the way the algorithm moves around the landscape to land at a local maximum. The multiple random initializations are merely a method to collect a random sample of objects (here: global maxima) that I'm interested in.

I make no claim as to how human learners explore the solution space. What is important is that human learners are able to arrive at a global maximum. Whether each learner find a global maximum by initializing at multiple random points and finding the nearest hill, or via incorporating search heuristics that enable a more efficient way of surveying the space makes no difference for my hypothesis.

In short: human learners do not need to run multiple random initializations nor run the EM algorithm for my hypothesis to hold. The multiple random initializations were a tool I used to randomly sample the global maxima since the set of global maxima could not be found analytically.

In general, over-parameterization is an issue in machine learning because over-parameterized models tend to fit the training data too well, and consequently fail to generalize to test data. How-

ever, when modeling inter-speaker variation, we have seen that over-parameterization can be a feature rather than a bug. Over-parameterization leads to the same surface pattern being generated by multiple distinct lexicon-grammar combinations. These distinct lexicon-grammar combinations generalize in different ways to test data. More importantly, when taken as a whole, we saw that the distribution of models generalizing in particular ways matched the distribution of human behavior in generalization tasks. That is to say, while over-parameterization is inelegant and faces a host of issues, the human learning of the lexicon and the grammar may indeed be a messy process, which may be best modeled by a collection of over-parameterized models. With that being said, there can be genuine cases of over-parameterization that are context-dependent. In one such example, the lexically-specified Markedness constraint (*e.g.* *$\text{k}_{\text{I}_{\text{vs-suff}}}$, which only applied to VS-suffixes) was dominated by the Faithfulness constraint, thus rendering the effects of the Markedness constraint invisible. In such cases, having a corresponding parameter that encodes the lexically-indexed option for a lexical item (*e.g.* whether a suffix is a VS-suffix or not) makes no sense, and the inclusion of such parameters to model the lexicon is a genuine case of over-parameterization. Such over-parameterization is context-dependent because whether or not the lexicon is over-parameterized is dependent upon the grammar learned.

Let us now turn our attention to a different point. While the over-arching model of inter-speaker variation predicted a distinction between Analysis B and Analysis C for velar softening, no probe could be used to distinguish between speakers that had one of these Analyses or the other. The difference between these two Analyses was the presence or absence of the *-ish* suffix in the lexicon, with 83.3% of these models positing the presence of the suffix[13], and one model missing it. In this case, the lone model probably missed the suffix because of the small number of unique training *word*s in which the *-ish* suffix could potentially be segmented out as a separate *morpheme*. If the training data had increased evidence for *-ish* being a suffix, the proportion of models that fail to posit the suffix is likely to decrease. Nevertheless, it is fascinating that there may exist some distinct lexicon-grammar combinations that are indistinguishable because no nonce word probe can be designed to distinguish between them. If so, there may be more internal inter-speaker variation (*i.e.* variation in the lexicon-grammar combination in speakers' heads) than can be outwardly

observed.

4.4 Conclusion

In Chapter 3, I had utilized my model to simultaneously learn URs and grammar. In this chapter, I extended my proposed model to learn morphology concurrently with phonology. For velar softening, the learning of morphology included learning whether to segment a potentially complex *word* into component morphemes, and learning whether a suffix should be lexically-specified as one that triggers velar softening. My model found that multiple distinct lexicon-grammar combinations could all produce the surface pattern of velar softening. I interpreted this to mean that my model predicted inter-speaker variation.

The following case is of particular interest: When the segmentation into morphemes fails, the corresponding grammar needed for the phonological alternation is also not acquired; instead, the mono-morphemic form (*e.g.* /ɪlɛktɪɪsɪti/, /sɑnɪsɪti/) is stored in the lexicon. My model predicted that such cases should occur at a low yet non-negligible rate. The Pierrehumbert study provided supporting evidence that a small minority of humans may indeed fall into this category. For these individuals, not acquiring the required morphology and grammar for the lexically-specified [k]∼[s] alternation results in them losing the generalization to novel stems. This in turn could contribute to the diachronic instability of velar softening as not all speakers would apply velar softening to new stems in the language.

[13] 96% of global maxima models posited the presence of the suffix, if Analysis A models are to be included as well.

CHAPTER 5

Richness of the Base

According to Prince and Smolensky (2004), the role of a constraint-based grammar is to assign an output to each input[1]. In the case of absolute ill-formedness, the grammar (*i.e.* the constraint interactions that govern the UR-SR[2] mapping) must ensure that no input ever leads to ill-formed outputs. For example, in a language that bans all voiced obstruents, any SR with a voiced obstruent would be considered ill-formed. For this language, a grammar that meets Prince and Smolensky's requirement would ensure that all URs (even implausible URs with underlying voiced obstruents) never surface with any voiced obstruents.

As we move from a categorical model like Optimality Theory to a model that has probabilistic UR-SR mappings, the grammar's role is now to ensure that none of the inputs ever map to ill-formed outputs with anything other than a vanishingly small probability. In other words, the grammar should be fail-safe; it should be able to map all URs (even implausible ones like /tag/ in a language without voiced obstruents) to SRs with appropriate probability values (*e.g.* [tak]: 99%, [tag]: 1%).

The concept that the grammar should be able to produce the right outputs[3] even when it is given implausible inputs is known as the Richness of the Base (Prince and Smolensky, 2004). I will call a grammar that meets this requirement a Rich Base Grammar (RBG).

---

[1]Prince and Smolensky (2004) were writing about Optimality Theory, where the grammar consisted of ranked constraints picking a sole output for each input.. Nevertheless, the grammar's role in mapping inputs to outputs still holds for probabilistic constraint-based grammars.

[2]Or more generally, the input-output mapping.

[3]*i.e.* Picking the correct SR in the case of Optimality Theory. Producing the correct probability distribution over SRs in the case of probabilitic constraint based grammatical models.

To illustrate the difference between a Rich Base Grammar and a non-Rich Base one, I use a language that bans all voiced obstruents. The two constraints governing the UR-SR mapping are *D, which bans voiced obstruents, and a general FAITH constraint, which militates against changing the segment. *Grammar*$_1$, in which the markedness constraint *D has a much higher weight than FAITH, is a Rich Base Grammar. In contrast, *Grammar*$_2$, in which the weights are flipped, is not a Rich Base Grammar (Table 5.1).

	Grammar$_1$	*Grammar*$_2$
Param weights:		
*D	5	0
FAITH	0	5

Table 5.1: *Final devoicing grammars.*

I assume that the UR-SR mapping is done via the MaxEnt model that was introduced in §3.2. For both the voiceless UR /tak/ and the voiced UR /tag/, *Grammar*$_1$ produces the correct outcome by assigning the overwhelming majority of the probability mass to the voiceless SR [tak], leaving only a vanishingly small probability mass for the illegal [tag] (Table 5.2). This grammar is thus compatible with both URs /tak/ and /tag/.

/tak/	*Prob*	$\mathcal{H}$	*D $w = 5$	FAITH $w = 0$
→ tak	.99	0		
tag	.01	5	1	1

/tag/	*Prob*	$\mathcal{H}$	*D $w = 5$	FAITH $w = 0$
→ tak	.99	0		1
tag	.01	5	1	

Table 5.2: *Grammar*$_1$ *is compatible with both /tak/ and /tag/.*

In contrast, *Grammar*$_2$ is able to produce the correct outcome only when the UR is the voiceless /tak/. For the UR /tag/, it incorrectly assigns 99.3% of the probability mass to the illegal [tag], which has a word-final voiced stop (Table 5.3). *Grammar*$_2$ is thus compatible with only the UR /tak/.

/tak/	*Prob*	$\mathcal{H}$	*D $w = 0$	FAITH $w = 5$
→ tak	.99	0		
tag	.01	5	1	1

/tag/	*Prob*	$\mathcal{H}$	*D $w = 0$	FAITH $w = 5$
tak	.01	5		1
← tag	.99	0	1	

Table 5.3: *Grammar$_2$ is compatible with* /tak/ *but not* /tag/.

The UR that *Grammar$_2$* is compatible with forms a proper subset of the URs that are compatible with *Grammar$_1$* (Table 5.4). In other words, *Grammar$_1$* is compatible with a richer set of URs than

	Grammar$_1$	*Grammar$_2$*
Param weights:		
*D	5	0
FAITH	0	5
UR-SR mappings:		
/tak/ → [tak]	99.3%	99.3%
/tag/ → [tak]	99.3%	0.7%
Compatible URs:		
/tak/	yes	yes
/tag/	yes	no
Rich base grammar?	*yes*	*no*

Table 5.4: *Summary of final devoicing grammars.*

Grammar$_2$ is. Hence, *Grammar$_1$* is the Rich Base Grammar because it can handle the Rich Base UR /tag/. *Grammar$_2$* is considered the non-Rich Base Grammar because when it is faced with the Rich Base UR /tag/, it is unable to produce a correct probability distribution over the candidate SRs.

In traditional phonological analyses, the implausible URs are included by the analyst in order for the Rich Base Grammar to be learned. Such URs are deemed "implausible" because there seems to be no reason for the child to posit a UR like /tag/ when they have only ever encountered [tak]. Nevertheless, evidence from loan word adaptation studies (Kang, 2011) show that speakers

often repair the illegal sounds and sound sequences of loan words to a sequence that is legal in their language. This suggests that speakers have learned the necessary Rich Base Grammar. For example, if a speaker of the no-voiced-obstruent language encountered the loan word <peg>, he would only repair it to [pɛk] if he had learned the Rich Base Grammar. If he instead had learned the non-Rich Base Grammar, he would realize the final consonant faithfully and produce [pɛg].

In order to force the Rich Base Grammar to be learned, some analysts have adopted particular strategies. In models that only model the UR-SR mapping, the implausible UR (*e.g.* /tag/) has to be included in the analysis despite the uncertain status of URs such as /tag/ ever serving as the actual UR. In a multiple UR analysis, Pater *et al.* (2012) incorporated a prior into the learner that favored higher weights for markedness constraints relative to faithfulness constraints.

My model learns URs and the grammar simultaneously. This presents an opportunity to ask whether there is a preference for the Rich Base Grammar to be learned when a particular UR is not assumed in advance.

In what follows, I present two case studies in which the training data was compatible with both a Rich Base Grammar and a non-Rich Base one. I ran 200 randomly initialized simulations for each case, and treated the trained models that tied for the equal-highest log likelihood as the outcome of learning. These trained models can be thought of as a random sample of the easily accessible local minima. In fact, in these two case studies, it turned out that the trained models that tied for the equal-highest log likelihood all reached the theoretical ceiling log likelihood. In other words, these trained models represent a random sample of the easily accessible global maxima. I found that a vast majority of these global maxima had a Rich Base Grammar, and that the acquisition of the Rich Base Grammar was not dependent on also acquiring a "helpful" Rich Base UR that would push the learner toward learning the Rich Base Grammar.

The rest of this chapter is organized as follows. The first case study, a categorically left-stressed language, is presented in §5.2. The second case study, a categorically right-stressed language, is presented in §5.3. §5.4 concludes.

5.1 Stress languages: Lexicon, constraints, UR and SR candidates

The two languages that I subjected the learner to were inspired by Tesar *et al.* (2003)'s PAKA World dataset.

5.1.1 The lexicon

In Tesar's dataset, the lexicon consisted of two roots: PA and BA, and two suffixes: -KA and -GA. The URs of PA and -KA were always unstressed. In contrast, it was up to the learner to figure out whether BA and -GA bore stress underlyingly in any given language. Table 5.5 provides a summary of the morphemes.

Underlying stress	*Root*	*Suffix*
Never	PA /pa/	-KA /-ka/
For learner to figure out	BA /ˈba, ba/?	-GA /-ˈga, -ga/?

Table 5.5: *Underlying stress of morphemes.*

5.1.2 The constraints

Accordingly, the four relevant UR constraints for this dataset were: (BA, /ˈba/), (BA, /ba/), (-GA, /-ˈga/), and (-GA, /-ga/). For completeness, the two UR constraints for the invariant PA and -KA morphemes were also included: (PA, /pa/), (-KA, /ka/).

Following Tesar *et al.* (2003), there were four phonological constraints that governed the UR-SR mapping (Table 5.6).

5.1.3 Potential sources of hidden structure: UR candidates

The four morphemes in Table 5.5 could be combined to form four training WORDs (left column of Table 5.7). The available underlying forms for these morphemes could be combined to form the potential URs in the right column of Table 5.7).

Constraint	*Abbreviation*	*Description*
MAINLEFT	ML	Stress the leftmost syllable.
MAINRIGHT	MR	Stress the rightmost syllable.
$\text{MAX}_{\text{general}}$-STRESS	F	If a syllable is stressed in the UR, retain its stress in the SR.
MAX_{root}-STRESS	FR	If a root syllable is stressed in the UR, retain its stress in the SR.

Table 5.6: *Phonological constraints for the stress languages.*

WORD	*UR*
PAKA	/pa -ka/
PAGA	/pa -ˈga/ /pa -ga/
BAKA	/ˈba -ka/ / ba -ka/
BAGA	/ˈba -ˈga/ / ba -ˈga/ /ˈba -ga/ / ba -ga/

Table 5.7: *UR candidates for the four training WORDs.*

5.1.4 SR candidates

For each UR candidate, three SR candidates were available: a left-stressed candidate (*e.g.* [ˈpaka]), a right-stressed candidate (*e.g.* [paˈka]), and a stressless candidate (*e.g.* [paka]).

5.1.5 The languages

The morphemes and constraints presented in this section could be combined to produce six logically possible languages. Of these six languages, two languages were of interest to me because these languages could be generated both by a Rich Base Grammar and a non-Rich Base Grammar. These languages were the left-stressed language (§5.2) and the right-stressed language (§5.3).

5.2 Left-stressed language

5.2.1 Inputs to the learner: WORD-SR pair frequencies

The left-stressed language had predictable left stress. As we have seen in §5.1.3, the four morphemes in Table 5.5 could be combined to form four WORDs. This produced the four WORD-SR training items shown in Table 5.8. Each WORD-SR pair was observed once. These WORD-SR

WORD		*SR*
PA	-KA	[ˈpaka]
PA	-GA	[ˈpaga]
BA	-KA	[ˈbaka]
BA	-GA	[ˈbaga]

Table 5.8: *Training WORD-SR pairs (left-stressed language).*

pair frequencies were used to train the models.

5.2.2 The Rich Base Grammar (left-stressed language)

In order for a constraint weight setting to qualify as a Rich Base Grammar, it must at minimum[4] meet the following weight inequality: MAINLEFT > MAINRIGHT + MAX_{gen}. That is, the weight of MAINLEFT should be much higher than the combined weights of MAINRIGHT and MAX_{gen}. The Rich Base Grammar is compatible with both underlyingly stressless /-ga/ and stressed /-ˈga/[5].

Tableau 5.9 and Tableau 5.10 illustrate how a Rich Base Grammar realizes the left-stressed SR [ˈpaga] when the suffix is /-ˈga/ and /-ga/ respectively. In these illustrative tableaux, the phonological constraints have the following weights: MAINLEFT: 8, MAINRIGHT: 1, MAX_{gen}: 2, MAX_{root}: 0. This particular weight setting is one of an infinite number of weight settings that qualify as a

[4]While meeting this weight inequality is a necessary condition of qualifying as a Rich Base Grammar, we will see in §5.2.7 that this is not a sufficient condition in and of itself. Looking ahead, we will see that both the direction and *magnitude* of the weight inequality matter if the Rich Base Grammar is to generalize left-stress to novel words.

[5]Each of the two potential URs for BA (*i.e.* /ba, ˈba/) are compatible with both the Rich Base Grammar and the Non-Rich Base Grammar. I will not be discussing this morpheme for the left-stressed language since it makes no interesting distinction between the Rich Base Grammar and the non-Rich Base Grammar of this language.

Rich Base Grammar because it has the correct weight inequality (Eq 5.1):

$$\begin{aligned} \textsc{MainLeft} &> \textsc{MainRight} + \textsc{Max}_{\text{gen}} \\ 8 &> 1 + 2 \\ 8 &> 3. \end{aligned} \tag{5.1}$$

In Tableau 5.9, the UR is /pa-ˈga/, which is right-stressed on the suffixal syllable in the underlying form. The left-stressed SR candidate [ˈpaga] violates MAINRIGHT because it doesn't stress the rightmost syllable and violates $\textsc{Max}_{\text{gen}}$ because it has deleted an underlying stress from /-ˈga/. This candidate gets a harmony score of 3 from violating MAINRIGHT ($w = 1$) and $\textsc{Max}_{\text{gen}}$ ($w = 2$) once each. The right-stressed SR candidate [paˈga] retains the underlying stress. In doing so, it violates the most important constraint MAINLEFT ($w = 8$) once, thus receiving a harmony score of 8. Applying the softmax function to the negative of these harmony scores results in the left-stressed SR [ˈpaga] surfacing 99% of the time.

/pa -ˈga/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 8$	MAINRIGHT $w = 1$	$\textsc{Max}_{\text{gen}}$ $w = 2$	$\textsc{Max}_{\text{root}}$ $w = 0$
→ ˈpaga	.99	3		1	1	
paˈga	.01	8	1			

Table 5.9: *The Rich Base Grammar is compatible with stressed /-ˈga/ (left-stressed language).*

In Tableau 5.10, the UR is now stressless /pa-ga/. Since there is no underlying stress, no SR candidate violates MAX. The left-stressed SR [ˈpaga] violates the less important MAINLEFT ($w = 1$) once while the right-stressed SR [paˈga] violates the more important MAINRIGHT ($w = 8$) once. Hence, left-stressed [ˈpaga] gets a harmony score of 1, surfacing 99.9% of the time while right-stressed [paˈga] gets a harmony score of 8, surfacing only .1% of the time.

/pa -ga/	*Prob*	$\mathcal{H}$	LEFT $w = 8$	RIGHT $w = 1$	$\textsc{Max}_{\text{gen}}$ $w = 2$	$\textsc{Max}_{\text{root}}$ $w = 0$
→ ˈpaga	.999	1		1		
paˈga	.001	8	1			

Table 5.10: *The Rich Base Grammar is compatible with stressless /-ga/ (left-stressed language).*

5.2.3 The non-Rich Base Grammar (left-stressed language)

In order to qualify as a non-Rich Base Grammar, a constraint weight setting must meet the following weight inequality: MAINLEFT $>$ MAINRIGHT, but must fail to meet the more stringent weight inequality required by a Rich Base Grammar (*i.e.* MAINLEFT $>$ MAINRIGHT + MAX_{gen}). Unlike the Rich Base Grammar, the non-Rich Base Grammar is compatible only with underlyingly stressless /-ga/ but not underlyingly stressed /-ˈga/.

Tableau 5.11 illustrates why a non-Rich Base Grammar is unable to produce a correct distribution over SR candidates when the suffix -GA is underlying stressed (*i.e.* /-ˈga/). Instead, the correct distribution over SR candidates can only be produced when the suffix -GA does not have underlying stress (*i.e.* /ga/ in Tableau 5.12).

In these illustrative tableaux, the phonological constraints have the following weights: MAINLEFT: 6, MAINRIGHT: 1, MAX_{gen}: 10, MAX_{root}: 0. This grammar is one of an infinite number of non-Rich Base Grammars because it meets the weight inequality required for non-Rich Base Grammars (Eq 5.2):

$$\begin{aligned} \text{MAINLEFT} &> \text{MAINRIGHT} \\ 6 &> 1, \end{aligned} \tag{5.2}$$

but fails to meet the more stringent weight inequality required of Rich Base Grammars (Eq 5.3):

$$\begin{aligned} \text{MAINLEFT} &> \text{MAINRIGHT} + \text{MAX}_{\text{gen}} \\ 6 &\not> 1 + 10 \\ 6 &\not> 11. \end{aligned} \tag{5.3}$$

The non-Rich Base Grammar produces a wrong distribution over SR candidates when the UR of the suffix -GA is stressed /-ˈga/. In Tableau 5.11, the UR is /pa-ˈga/, which is underlyingly stressed on the suffixal syllable. The left-stressed SR candidate [ˈpaga] drops the underlying stress from a syllable, thus picking up a violation of the most important constraint, MAX_{gen} ($w = 10$). This left-stressed candidate also picks up one violation of MAINRIGHT ($w = 1$), thus receiving a harmony score of 11. The right-stressed SR candidate [paˈga] retains the underlying stress on the

suffixal syllable, and incurs no violations of MAX constraints. In doing so, it violates the MAINLEFT ($w = 6$) constraint once, and receives a harmony score of 6. These harmony scores translate to a probability of 99% for the right-stressed [pa'ga] and only 1% for the left-stressed ['paga]. This produces a wrong distribution over SR candidates because in a left-stressed language, the overwhelming majority of the probability mass should be assigned to the left-stressed SR candidate ['paga].

/pa -'ga/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 6$	MAINRIGHT $w = 1$	MAX_{gen} $w = 10$	MAX_{root} $w = 0$
'paga	.01	11		1	1	
← pa'ga	.99	6	1			

Table 5.11: *The non-Rich Base Grammar is incompatible with stressed /-'ga/ (left-stressed language).*

In contrast, the non-Rich Base Grammar is able to produce a correct distribution over SR candidates when the UR of the suffix -GA is stressless /-ga/. In Tableau 5.12, the UR is /pa-ga/. Since the UR is stressless, no SR candidates violate the most important constraint MAX_{gen}. Of the markedness constraint violations, the left-stressed SR candidate ['paga] violates the less important MAINRIGHT ($w = 1$) while the right-stressed SR candidate [pa'ga] violates the more important constraint MAINLEFT ($w = 6$). Consequently, the left-stressed ['paga] is assigned the majority (99%) of the probability mass, and the right-stressed [pa'ga] receives the remaining 1%.

/pa -ga/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 6$	MAINRIGHT $w = 1$	MAX_{gen} $w = 10$	MAX_{root} $w = 0$
→ 'paga	.99	1		1		
pa'ga	.01	6	1			

Table 5.12: *The non-Rich Base Grammar is compatible only with stressless /-ga/ (left-stressed language).*

5.2.4 Summary of Grammars and URs (left-stressed language)

To summarize, in a left-stressed language, a correct probability distribution assigns the majority of the probability mass to the left-stressed SR candidate. The Rich Base Grammar is able to produce

a correct probability distribution regardless of whether the UR of the suffix -GA is stressed /-ˈga/ or stressless /-ga/. In contrast, the non-Rich Base Grammar can produce a correct probability distribution only when the UR of the suffix -GA is stressless /-ga/. Hence, I will call stressed /-ˈga/ the Rich Base UR of the suffix -GA because only this UR can distinguish between the Rich Base and non-Rich Base Grammars. As for the root morpheme BA, it turns out that both the Rich Base Grammar and the non-Rich Base Grammar are compatible[6] with both stressed /ˈba/ and stressless /ba/. A summary of the URs that are compatible with the Rich Base and non-Rich Base Grammars is presented in Table 5.13.

		Rich Base Grammar	*Non-Rich Base Grammar*
Lexicon:			
BA	/ba/	yes	yes
	/ˈba/	yes	yes
-GA	/-ga/	yes	yes
	/-ˈga/	yes	no
Grammar:		MAINLEFT > MAINRIGHT + MAX_{gen}	MAINLEFT > MAINRIGHT

'yes': compatible with the grammar in the column. 'no': incompatible.

Table 5.13: *Lexicon and grammar for the Rich Base and non-Rich Base Grammars (left-stressed language).*

The minimum required weight inequalities for the Rich Base and non-Rich Base Grammars are also summarized in Table 5.13. To qualify as a Rich Base Grammar, a set of constraint weights must meet the weight inequality: MAINLEFT > MAINRIGHT + MAX_{gen}. To qualify as a non-Rich Base Grammar, a set of constraint weights must meet the weight inequality set out for non-Rich Base Grammars (*i.e.* MAINLEFT > MAINRIGHT) and fail to meet the stricter weight inequality for Rich Base Grammars[7].

[6] *i.e.* Produces a correct probability distribution.

[7] Due to the restriction that constraint weights cannot be negative, all Rich Base Grammars also meet the weight inequality required by the non-Rich Base Grammar. So if a set of constraint weights meets the requirements of both the non-Rich Base Grammar and the stricter Rich Base Grammar, this set of constraint weights will be classified as a Rich Base Grammar.

5.2.5 Results

The results section is organized as follows. In §5.2.5.1, I report that 145 of the 200 trained models tied for the best match to the training data (*i.e.* WORD-SR frequencies), and were in fact able to match the surface pattern of the training data perfectly. These 145 trained models were then subjected to further tasks in order to ascertain the hidden structures (here: lexicon) and grammars that they had learned.

An inspection of the constraint weights showed that all 145 trained models met the weight inequality associated with the Rich Base Grammar (*i.e.* MAINLEFT > MAINRIGHT + MAX_{gen}) §5.2.5.2.

In §5.2.6, I elaborate why merely meeting the direction of this weight inequality is not good enough for a constraint weight setting to truly function as Rich Base Grammar for a **categorically** left-stressed language. In short, when a model attains the Rich Base Grammar weight inequality, we know that it will always give the left-stressed SR the best (*i.e.* lowest) harmony score out of all the SR candidates. However, what we really hope for with a categorically left-stressed language is for the left-stressed SR to get an **overwhelming majority of the probability mass**, as opposed to just being the SR candidate that gets the highest probability mass out of all the SR candidates. Within the model architecture, the harmony scores are converted to probabilities via a softmax function, which takes the difference between harmony scores into account. Small differences in harmony scores between SR candidates result in a relatively flat probability distribution while large differences result in skewed probability distributions that when large enough, do look categorical. In other words, the **magnitude** of a model's Rich Base Grammar weight inequality must be large enough for the model to produce a categorically left-stressed language.

How large is large enough? Rather than stipulating an arbitrary value for magnitude of the Rich Base Grammar weight inequality, I assess whether trained models are able to generalize categorical left-stress to novel words. The wug word test (§5.2.7.1), presents an easier task in which a trained model is able to use the "easier" non-Rich Base UR (if it had so acquired one) to make up for any deficiency in the magnitude of its Rich Base Grammar weight inequality. In the loan word test (§5.2.7.4), successful generalization requires the trained models to rely on the magnitude of the

Rich Base Grammar weight inequality alone.

The proportion of trained models that had learned the Rich Base UR /-ˈga/ versus the non-Rich Base UR /-ga/ is reported in §5.2.7.2.

5.2.5.1 Matching WORD-SR frequencies of training data

The log likelihood of a trained model for a given data set is a measure of how well the model fits the data set. Log likelihoods range from $-\inf$ to 0, with a higher log likelihood indicating a better model fit to data.

I ran 200 randomly initialized simulations. A histogram of final log likelihoods (Figure 5.1) indicated that the majority of the trained models had log likelihoods in the highest bin, which ranged from -5.62 to -5.42.

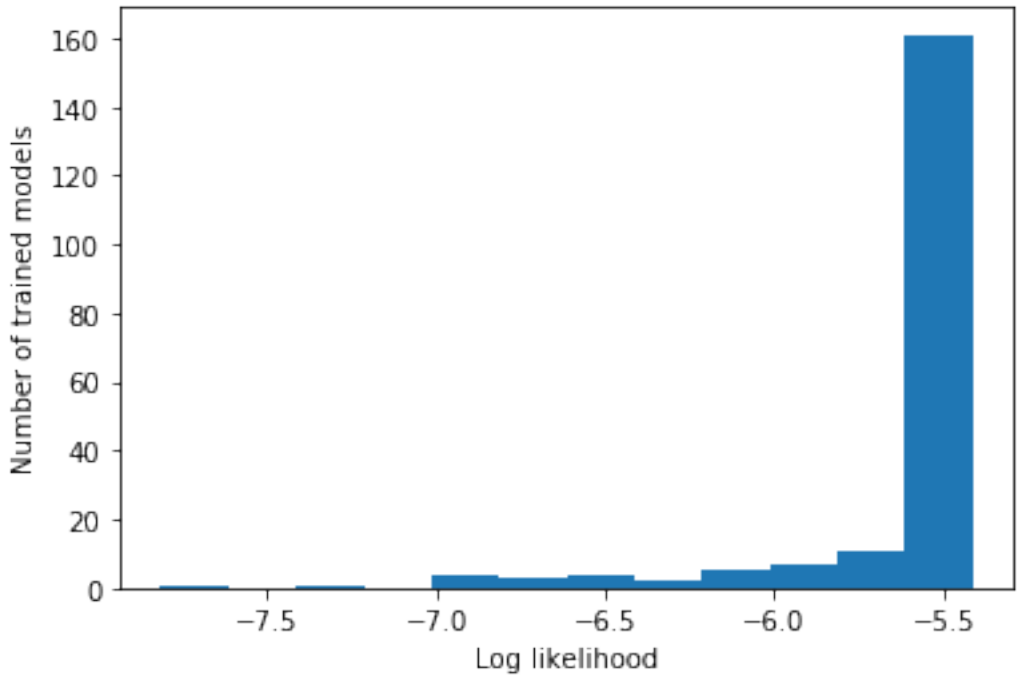

Figure 5.1: *Distribution of log likelihoods of trained models (left-stressed).*

A finer-grained visualization (Figure 5.2) showed that there was a cluster of highest final log likelihoods between -5.550 and -5.545, so I considered trained models whose final log likelihoods fell above -5.550 as tying for the equal-highest final log likelihood. 145 trained models

met this criterion.

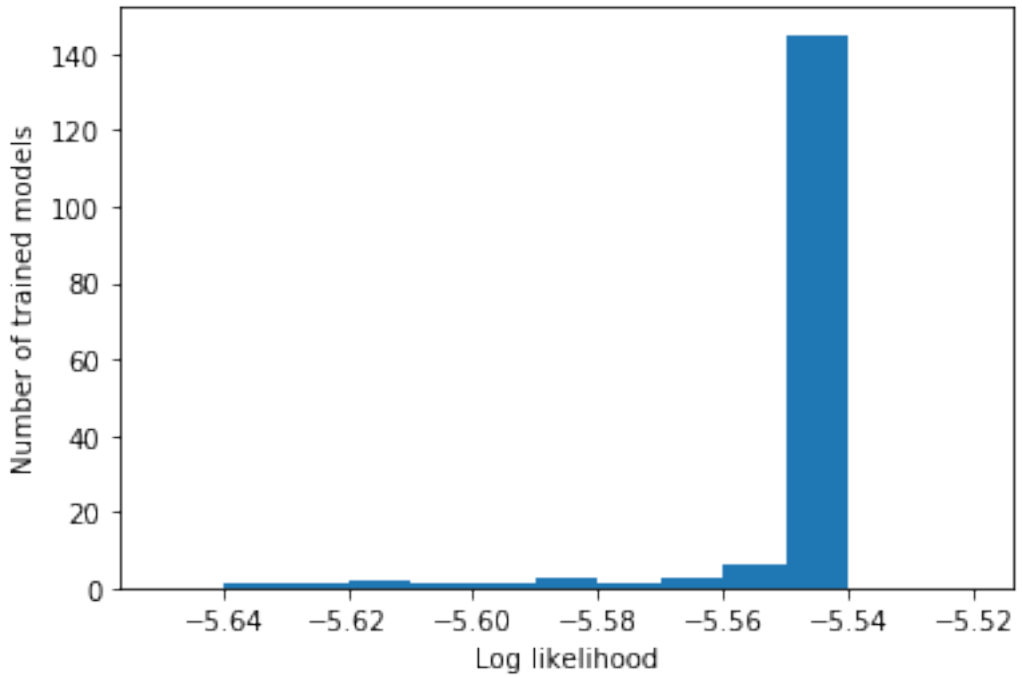

Figure 5.2: *Distribution of log likelihoods of trained models, which were higher than –5.65 (left-stressed).*

Each data set has a theoretical ceiling log likelihood, which can be calculated (Eq 5.4). For the data set $\mathcal{D}$ that consists of training examples x,

$$\text{ceiling log likelihood} = \sum_{x \in \mathcal{D}} ln(P(x)) \tag{5.4}$$

Applying Eq 5.4 to the training data gives:

$$\begin{aligned} \text{ceiling log likelihood} &= \sum_{x \in \mathcal{D}} ln(P(x)) \\ &= \begin{matrix} ln(P(\text{(PAKA, ['paka])))} + ln(P(\text{(PAGA, ['paga])))} \\ + ln(P(\text{(BAKA, ['baka])))} + ln(P(\text{(BAGA, ['baga])))} \end{matrix} \\ &= ln(1/4) + ln(1/4) + ln(1/4) + ln(1/4) \\ &= -5.545. \end{aligned} \tag{5.5}$$

These 145 trained models essentially reached the ceiling log likelihood. Of these 145 trained models, the most deviant one had a log likelihood that was .0002 away from the theoretical ceiling.

The learner was able to find 145 unique[8] parameter settings that matched the training data perfectly. This means that these 145 trained models will perfectly generate the surface pattern (*i.e.* the WORD-SR frequencies) of the training data for the left-stressed language.

5.2.5.2 Proportion of Rich Base Grammars

The surface pattern for the left-stressed language could be generated via a Rich Base Grammar (§5.2.2) or a non-Rich Base one (§5.2.3). Since my model allowed the learner to train a model without assuming URs in advance, we could ask what proportion of these 145 perfect trained models had grammars that corresponded to the Rich Base Grammar when URs were not stipulated in advance.

To recap, the weight inequality required to be a Rich Base Grammar was: MAINLEFT $>$ MAINRIGHT + MAX_{gen}. It turned out that all 145 trained models that learned the left-stressed language perfectly met the weight inequality required of the Rich Base Grammar. In fact, all 200 trained models met the weight inequality required of the Rich Base Grammar too. That is, all global (and indeed local) maxima found correspond to a Rich Base Grammar.

5.2.6 Categoricity

The toy language that was submitted to the learner was categorically left-stressed. That is to say, for a given WORD, the left-stressed SR had a relative frequency of 100%. The model, however, produced a probabilistic distribution over WORD-SR pairs because harmony scores were converted to probabilities via a softmax[9] function. The misalignment between a categorical phenomenon and a probabilistic model produces an additional challenge, which is discussed below.

In general, if the difference between the left-hand side and the right-hand side of the weight

[8] In fact, all 200 local maxima found by the learner were unique.

[9] If the harmony scores were instead converted to the output of the model via a max function, then meeting the direction of the weight inequality would be enough to guarantee categoricity. Model architectures which apply a max function to the harmony scores include the ones in Legendre *et al.* (1990), Legendre *et al.* (2006), Potts *et al.* (2010), Pater (2016), *a.o.*

inequality is large enough, then an essentially categorical outcome ensues. Tableau 5.14 illustrates that a difference of 5 results in one SR[10] (here: [ˈpaga]) receiving the overwhelming (99%) of the probability mass. However, if the difference between the left-hand side and the right-hand side

/pa -ˈga/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 8$	MAINRIGHT $w = 1$	MAX_{gen} $w = 2$	MAX_{root} $w = 0$
→ ˈpaga	.99	3		1	1	
paˈga	.01	8	1			

Table 5.14: *An essentially categorical outcome (left-stressed language).*

of the weight inequality is not large enough, then the outcome clearly cannot be considered to be categorical. Tableau 5.15 shows that a smaller difference of 1 results in a less skewed distribution over SRs.

/pa -ˈga/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 6$	MAINRIGHT $w = 2$	MAX_{gen} $w = 3$	MAX_{root} $w = 0$
→ ˈpaga	.73	5		1	1	
paˈga	.27	6	1			

Table 5.15: *A non-categorical outcome (left-stressed language).*

In both Tableau 5.14 and Tableau 5.15, the "correct" SR wins because it receives the highest conditional probability out of all the SR candidates. Nevertheless, the essentially categorical outcome in Tableau 5.14 is a much better model of the categorically left-stressed language; Tableau 5.15 is an unsatisfactory model.

Hence, merely meeting the direction of the weight inequality required by the Rich Base Grammar is not good enough. The difference between the left-hand and right-hand sides of the inequality much be large enough to produce (essentially) categorical outcomes.

[10]The model produces a probability distribution over WORD-SR pairs. Within the model are two constituent MaxEnt models, one of which produces a probability distribution over SRs for each UR. It is this constituent MaxEnt model that is discussed here since the weight inequality concerns phonological constraints, which govern the UR-SR mapping.

5.2.7 Generalization to test sets

In this section, I introduce two test words whose purpose is to confirm that the difference between the left-hand and right-hand sides of the inequality is large enough to generalize **categorical** left-stress to novel words. I set the criterion for a categorical outcome at 99%. That is to say, a trained model passes the generalization test for a novel test word if it assigns at least 99% of the predicted probability mass to the left-stressed SR candidate.

The first test word simulates a wug test, wherein the test word consists of one trained morpheme and one new morpheme (§5.2.7.1). The second test word simulates a loan word, in which all morphemes are new (§5.2.7.4).

In order to design a rigorous test, the test words were chosen to reflect the most difficult case – that of the implausible UR – whenever possible[11]. In a left-stressed language, the implausible UR is one that bears the following shape: $/\sigma\acute{\sigma}/$ – in other words, bearing stress on the rightmost syllable but not on the leftmost syllable.

5.2.7.1 Wug test

For the wug test, the test word SO-GA consisted of the new (*i.e.* untrained) root SO, and the old suffix -GA. Since the trained models had never encountered the root before and did not know anything about it, I told these models that SO was underlyingly stressless $/\text{so}/$. As for the -GA suffix, it was up to the trained models to fill in what they had each learned during training about whether -GA was stressed or stressless.

Each model's task was to predict the conditional probability of the left-stressed SR candidate ['soga] for the given wug word SO-GA. I considered a trained model to have successfully generalized the categorical left-stressed language to a novel $\text{ROOT}_{\text{novel}}$-$\text{SUFF}_{\text{trained}}$ word if the conditional probability $P(\text{['soga]}|\text{SO-GA})$ was at least 99%.

Recall that each model outputs a joint probability distribution over WORD-UR-SR triples.

[11] In the wug test, it is not possible to dictate that the second syllable always bears stress. This is because the second syllable comes from a trained morpheme, and a trained model would utilize the UR of the morpheme that it had learned during training.

An example of a possible output is shown in Table 5.16. The conditional probability $P(\text{SR} =$

WORD-UR-SR	Predicted probability
P(SO-GA, /so-ˈga/, [ˈsoga])	.550
P(SO-GA, /so-ˈga/, [soˈga])	.004
P(SO-GA, /so-ˈga/, [soga])	.001
P(SO-GA, /so-ga/, [ˈsoga])	.440
P(SO-GA, /so-ga/, [soˈga])	.004
P(SO-GA, /so-ga/, [soga])	.001

Table 5.16: *A potential predicted joint probability distribution over WORD-UR-SR triples.*

$s|\text{WORD} = w)$ can be calculated from the joint probability over triples (P(WORD, UR, SR)) as follows:

$$P(\text{SR} = s|\text{WORD} = w) = \frac{\sum_{u'} P(w, u', s)}{\sum_{u'} \sum_{s'} P(w, u', s')}. \tag{5.6}$$

Applying Eq 5.13 to the joint probability distribution in Table 5.16 shows that the model that generates this joint probability distribution meets the criterion for successfully generalizing to wug words because the left-stressed SR [ˈpaga] gets 99% of the probability mass for the wug word SO-GA:

$$\begin{aligned} P([\text{ˈsoga}]|\text{SO-GA}) &= \frac{P(\text{SO-GA, /so-ˈga/, [ˈsoga]}) + P(\text{SO-GA, /so-ga/, [ˈsoga]})}{\begin{array}{c} P(\text{SO-GA, /so-ˈga/, [ˈsoga]}) + P(\text{SO-GA, /so-ˈga/, [soˈga]}) \\ + P(\text{SO-GA, /so-ˈga/, [soga]}) + P(\text{SO-GA, /so-ga/, [ˈsoga]}) \\ + P(\text{SO-GA, /so-ga/, [soˈga]}) + P(\text{SO-GA, /so-ga/, [soga]}) \end{array}} \\ &= \frac{.55 + .44}{.55 + .004 + .001 + .44 + .004 + .001} \\ &= .99\ . \end{aligned} \tag{5.7}$$

It turned out that all 145 trained models that reached the ceiling log likelihood were able to meet the criterion for generalizing categorical left-stress to a wug word by attaining $P([\text{ˈsoga}]|\text{SO-GA}) \geq 99\%$. In fact, all 200 trained models generalized to the left-stressed SR [ˈsoga] for the wug word SO-GA at rates greater than 99%.

5.2.7.2 Lexicon

There were two hidden "unobserved" sources from which [ˈsoga] could have arisen: /so-ˈga, so-ga/. In fact, the UR that a particular trained model had used for the wug word SO-GA could be fully /so-ˈga/, fully /so-ga/, or a mixture of the two. The model that produced the joint probabilities in Table 5.16 was an example of a model that used a mixed-UR: 55.5% /so-ˈga/, 44.5% /so-ga/.

Since my learner did not require URs to be specified in advance, it would be interesting to discover the proportion of models that learned to use /so-ˈga/, /so-ga/, or a mixture of the two. For each trained model, I considered the UR of SO-GA to be fully /so-ˈga/ when the conditional probability of /so-ˈga/ given the wug word SO-GA, $P(/\text{so-ˈga}/|\text{SO-GA})$, was at least 99%. The same 99% threshold applied likewise for the UR to be considered fully /so-ga/. If neither of these criteria were met, then I considered the UR to be "mixed". The relevant conditional probabilities were calculated as follows:

$$P(\text{UR} = u|\text{WORD} = w) = \frac{\sum_{s'} P(w, u, s')}{\sum_{u'} \sum_{s'} P(w, u', s')}. \tag{5.8}$$

For the 145 trained models that matched the surface pattern of the training data perfectly, 76 (52.4%) used the implausible Rich Base UR /so-ˈga/ while 69 (47.6%) used the non-Rich Base UR /so-ga/. None of these perfect trained models used a mixture of /so-ˈga/ and /so-ga/[12].

5.2.7.3 Interim summary

Let us take stock of what we have seen so far. All 145 trained models that matched the surface pattern of the training data perfectly also had weight inequalities that pointed in the direction that was required of Rich Base Grammars. This indicated that these 145 models would at worst produce a probabilistic left-stressed language. Nevertheless, the language to be learned was categorical while the trained models produced probabilistic outcomes. The wug task was introduced to check

[12]For the 55 trained models that failed to match the surface pattern of the training data perfectly, 18 (32.7%) used the Rich Base UR /so-ˈga/, 16 (29.1%) used the non-Rich Base UR /so-ga/, and 21 (38.2%) used a mixture of the two URs.

that the magnitude of the weight inequality was large enough to produce essentially categorical outcomes. All 145 trained models were found to successfully generalize to the left-stressed SR ['soga] of the wug word SO-GA at a rate greater than 99%, which I had set as the threshold for categoricity.

The wug word consisted of a new root SO and the trained suffix -GA. In other words, the wug task simultaneously depended on the trained models' knowledge about both (1) the grammar (*i.e.* whether the required weight inequality was large enough and in the right direction) and (2) whether the UR of the suffix -GA was the Rich Base UR /-ˈga/ or the non-Rich Base /-ga/. I found that 47.6% of these 145 trained models had learned the non-Rich Base UR /-ga/ for the -GA suffix. In other words, while all 145 trained models succeeded in generalizing categorical left-stress to the wug word SO-GA, it was unclear whether this work was done entirely by acquiring a Rich Base Grammar or whether the same work was done using a combination of a non-Rich Base Grammar with the "easier" non-Rich Base UR /-ga/.

In order to isolate the result of the generalization task to the grammar alone, I introduce the loan word generalization task (§5.2.7.4).

5.2.7.4 Loan word test

The loan word NOVE did not contain any old "trained" morphemes. Hence, I could ensure that all trained models used exactly the same UR /noˈve/. The stress pattern of this loan word was designed to bear the stress pattern of the implausible UR: /$\sigma\acute{\sigma}$/.

As before, each model's task was to predict the conditional probability of the left-stressed SR candidate ['nove] for the given loan word NOVE. A model was considered to have successfully generalized the categorical left-stressed language to loan words only when $P(\text{['nove]}|\text{NOVE}) \geq .99$.

All 145 trained models that perfectly matched the surface pattern of the training data were able to generalize categorical left-stress to the loan word NOVE, which had the anomalous right-stressed UR /noˈve/[13]. This result confirms that these 145 trained models each learned a setting of constraint weights that corresponded to the Rich Base Grammar, both in the direction of the weight

inequality (by giving the left-stressed SR the greatest probability mass) and in the magnitude of the weight inequality (by achieving categoricity).

5.2.8 Summary: left-stressed language

I trained 200 randomly initialized models to learn a categorically left-stressed language. It was up to each model to decide whether the suffix -GA[14] was unstressed /-ga/ or stressed /-ˈga/ and also to learn an appropriate grammar. The choice of /-ga/ or /-ˈga/ had different implications when it came to the grammar. With the "easier" unstressed UR /-ga/, two weight inequalities (*i.e.* grammars) were able to produce a distribution over SR candidates (*i.e.* [ˈpaga, paˈga, paga]) that correctly assigned the majority of the probability mass to the left-stressed [ˈpaga]. However, with the "difficult" stressed UR /-ˈga/, only one weight inequality (*i.e.* grammar) was able to produce a correct probability distribution over SRs. Thus, the "difficult" UR /-ˈga/ was termed the Rich Base UR, and the weight inequality that could produce a correct probability distribution over SRs from this UR was called the Rich Base Grammar.

Of the 200 trained models, 145 were able to match the surface pattern of the training data perfectly. An inspection of the learned constraint weights showed that all 145 of these models attained the minimum weight inequality required of a Rich Base Grammar (*i.e.* at minimum, they could model a non-categorical left-stressed language).

Additional wug and loan word tests revealed that these 145 models could generalize both the **left-stressed nature** and the **categoricity** of the training data to novel words. This result was not trivial – if a trained model had learned the "easier" non-Rich Base UR /-ga/ without acquiring a Rich Base weight inequality that was both (1) in the correct direction and (2) large enough, then it could potentially pass the generalization test with the wug word by relying on the UR /soga/, but fail the loan word generalization test in which the UR was /noˈve/.

Rather interestingly, I found that the acquisition of the Rich Base UR was not required in order

[13]In fact, all 200 trained models were able to generalize categorical left-stress to NOVE /noˈve/.

[14]These models also had to learn whether the root BA was /ba/ or /ˈba/. However, this choice was of no significance to the left-stressed language, so it was not discussed.

for the Rich Base Grammar to be learned. While all 145 of these trained models learned a Rich Base Grammar, close to half (52.4%) of them learned the Rich Base UR /-ˈga/ while the remainder (47.6%) learned the non-Rich Base UR /-ga/.

Taken together, these results indicate that for the categorically left-stressed language, there was (1) an overwhelming preference for the Rich Base Grammar to be learned when no Rich Base UR was assumed in advance, and that (2) the Rich Base Grammar was acquired regardless of whether the Rich Base UR was learned.

5.3 Right-stressed language

5.3.1 Inputs to the learner: WORD-SR pair frequencies

The right-stressed language had predictable right stress. The four morphemes in Table 5.5 could be combined to form four WORDs. These were the same four WORDs that were used in the left-stressed language; however, the left-stressed and right-stressed languages differed because the very same WORDs had different surface realizations (*i.e.* SRs). The four WORD-SR training items are shown in Table 5.17. Each WORD-SR pair was observed once. These WORD-SR pair frequencies

WORD		*SR*
PA	-KA	[paˈka]
PA	-GA	[paˈga]
BA	-KA	[baˈka]
BA	-GA	[baˈga]

Table 5.17: *Training WORD-SR pairs (right-stressed language).*

were used to train the models.

5.3.2 The Rich Base Grammar (right-stressed language)

The Rich Base Grammar requires the following weight inequality: MAINRIGHT > MAINLEFT + MAX_{gen} + MAX_{root}. This should be thus interpreted: the weight of MAINRIGHT should be much higher than the combined weights of MAINLEFT, MAX_{gen} and MAX_{root}. This Rich Base Grammar

is compatible with both the underlyingly stressless root /ba/ and stressed root /ˈba/[15].

The Rich Base Grammar is able to realize right-stressed [baˈka] when the root BA has the "easier" non-Rich Base UR /ba/ (Tableau 5.18) and also when it has the "difficult" Rich Base UR /ˈba/ (Tableau 5.19). In these illustrative tableaux, the phonological constraints have the following weights: MAINLEFT: 1, MAINRIGHT: 8, MAX_{gen}: 1, MAX_{root}: 1. This weight setting happens to be one of an infinite number of weight settings that meet the weight inequality requirement to qualify as a Rich Base Grammar for the right-stressed language (Eq 5.9):

$$\begin{aligned} \text{MAINRIGHT} &> \text{MAINLEFT} + \text{MAX}_{\text{gen}} + \text{MAX}_{\text{root}} \\ 8 &> 1 + 1 + 1 \\ 8 &> 3. \end{aligned} \tag{5.9}$$

In Tableau 5.18, the UR is /ˈba-ka/, which is left-stressed on the root syllable in the underlying form. The right-stressed SR candidate [baˈka] violates MAINLEFT because it doesn't stress the leftmost syllable and violates both faithfulness constraints MAX_{root} and MAX_{gen} because it deletes an underlying stress from the root syllable /ˈba/. This candidate gets a harmony score of 3 from violating MAINRIGHT ($w = 1$), MAX_{root} ($w = 1$) and MAX_{gen} ($w = 1$) once each. The left-stressed SR candidate [ˈbaka] retains the underlying stress of the root syllable /ˈba/. This results in a violation of the most important constraint MAINRIGHT ($w = 8$), so left-stressed [ˈbaka] gets a harmony score of 8. Converting these negative of these harmony scores to probabilities via the softmax function results in the right-stressed SR [baˈka] surfacing 99% of the time.

/ˈba -ka/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 1$	MAINRIGHT $w = 8$	MAX_{gen} $w = 1$	MAX_{root} $w = 1$
ˈbaka	.01	8		1		
→ baˈka	.99	3	1		1	1

Table 5.18: *The Rich Base Grammar is compatible with stressed /ˈba/ (right-stressed language).*

In Tableau 5.19, the UR is the stressless /ba-ka/. Since there is no underlying stress, no SR

[15]Each of the two potential URs for -GA (*i.e.* /-ga, -ˈga/) are compatible with both the Rich Base Grammar and the Non-Rich Base Grammar. I will not be discussing this morpheme for the right-stressed language since it makes no interesting distinctions between the Rich Base Grammar and the non-Rich Base Grammar of this language.

candidate violate the MAX constraints. The left-stressed SR ['baka] violates the very much more important MAINRIGHT ($w = 8$) once while the left-stressed SR ['baka] violates the less important MAINRIGHT ($w = 1$) once. Hence, the right-stressed [ba'ka] gets a small harmony score of 1 and surfaces 99.9% of the time while left-stressed ['baka] gets a large harmony score of 8 and surfaces only $.1\%$ of the time.

/ba -ka/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 1$	MAINRIGHT $w = 8$	MAX_{gen} $w = 1$	MAX_{root} $w = 1$
'baka	.001	8		1		
→ ba'ka	.999	1	1			

Table 5.19: *The Rich Base Grammar is compatible with stressless /ba/ (right-stressed language).*

5.3.3 The non-Rich Base Grammar (right-stressed language)

For the right-stressed language, a setting of constraint weights must meet the following weight inequality to qualify as a non-Rich Base Grammar : MAINRIGHT > MAINLEFT. Such a weight setting must also fail to meet the more stringent weight inequality required by a Rich Base Grammar (*i.e.* MAINRIGHT > MAINLEFT + MAX_{gen} + MAX_{root}), otherwise it would instead qualify as a Rich Base Grammar. Unlike the Rich Base Grammar, the non-Rich Base Grammar is compatible only with the underlyingly stressless /ba/ but not the underlyingly stressed /'ba/.

Tableau 5.20 shows why a non-Rich Base Grammar is cannot produce a correct distribution over SR candidates when the root BA is underlying stressed (*i.e.* /'ba/). The correct distribution over SR candidates can only be produced when the root BA does not have underlying stress (*i.e.* /ba/ in Tableau 5.21).

In these illustrative tableaux, the phonological constraints have the following weights: MAINLEFT: 1, MAINRIGHT: 8, MAX_{gen}: 5, MAX_{root}: 5. This particular setting of constraint weights is one of an infinite number of non-Rich Base Grammars because it meets the weight inequality

required of non-Rich Base Grammars (Eq 5.10):

$$\begin{gathered} \textsc{MainRight} > \textsc{MainLeft} \\ 8 > 1, \end{gathered} \tag{5.10}$$

but fails to meet the more stringent weight inequality required of Rich Base Grammars (Eq 5.3):

$$\begin{aligned} \textsc{MainRight} &> \textsc{MainLeft} + \textsc{Max}_{\text{gen}} + \textsc{Max}_{\text{root}} \\ 8 &\ngtr 1 + 5 + 5 \\ 8 &\ngtr 11. \end{aligned} \tag{5.11}$$

The non-Rich Base Grammar produces a wrong distribution over SR candidates for the right-stressed language when the UR of the root BA is stressed /ˈba/. In Tableau 5.20, the UR is /ˈba-ka/, which is underlyingly stressed on the root, which is the leftmost syllable. The right-stressed SR candidate [baˈka] drops the underlying stress from the root syllable, thus picking up a violation each from both $\textsc{Max}_{\text{gen}}$ ($w = 5$) and $\textsc{Max}_{\text{root}}$ ($w = 5$). This right-stressed candidate also picks up one violation of MainLeft ($w = 1$), and so receives a harmony score of 11. The left-stressed SR candidate [ˈbaka] retains the root syllable's underlying stress, so it does not incur any violations of the Max constraints. In order to preserve the stress on the root syllable, it violates the MainRight ($w = 8$) constraint once, and receives a harmony score of 8. These harmony scores convert to a probability of 95% for the left-stressed [ˈbaka] and 5% for the right-stressed [baˈka]. In a right-stressed language, the overwhelming majority of the probability mass should be assigned to the right-stressed SR candidate [baˈka]; hence, this particular setting of constraint weights produces a wrong distribution over SR candidates.

/ˈba -ka/	*Prob*	$\mathcal{H}$	MainLeft $w = 1$	MainRight $w = 8$	$\textsc{Max}_{\text{gen}}$ $w = 5$	$\textsc{Max}_{\text{root}}$ $w = 5$
← ˈbaka	.95	8		1		
baˈka	.05	11	1		1	1

Table 5.20: *The non-Rich Base Grammar is incompatible with stressed /ba/ (right-stressed language).*

In contrast, when when the UR of the root BA is stressless /ba/,the non-Rich Base Grammar is able to produce a correct distribution over SR candidates. In Tableau 5.21, the UR is /ba-ka/. Since the UR is stressless, no SR candidates violate the MAX constraints, which when ganged up, would make deletion of stress from the root syllable the most costly choice. Shifting our attention to the markedness constraints, the right-stressed SR candidate [ba'ka] violates the less costly MAINLEFT ($w = 1$) while the left-stressed SR candidate ['baka] violates the more costly constraint MAINRIGHT ($w = 8$). As a result, the right-stressed [ba'ka] receives the majority (99.9%) of the probability mass, leaving only 0.1% to the left-stressed ['baka].

/'ba -ka/	*Prob*	$\mathcal{H}$	MAINLEFT $w = 1$	MAINRIGHT $w = 8$	MAX_{gen} $w = 5$	MAX_{root} $w = 5$
'baka	.001	8		1		
→ ba'ka	.999	1	1			

Table 5.21: *The non-Rich Base Grammar is compatible only with stressless /-ba/ (left-stressed language).*

5.3.4 Summary of Grammars and URs (right-stressed language)

In a right-stressed language, a correct probability distribution gives the majority of the probability mass to the right-stressed SR candidate. We have seen that the Rich Base Grammar is able to produce a correct probability distribution regardless of whether the UR of the root BA is stressed /-'ba/ or stressless /-ba/. However, the non-Rich Base Grammar produces a correct probability distribution only when the UR of the root BA is stressless /-ba/. Consequently, the stressed UR /'ba/ is treated as the Rich Base UR of the root BA since this is the only UR that can distinguish between the Rich Base and non-Rich Base Grammars. Regarding the suffix -GA, both URs /-'ga/ and /-ga/ are compatible with both the Rich Base and non-Rich Base Grammars. A summary of the Rich Base and non-Rich Base Grammars and the URs that they are compatible with is found in Table 5.22.

The minimum required weight inequalities for the Rich Base and non-Rich Base Grammars are also summarized in Table 5.22. To qualify as a Rich Base Grammar, a set of constraint weights must at minimum meet the more stringent weight inequality: MAINRIGHT $>$ MAIN-

		Rich Base Grammar	*Non-Rich Base Grammar*
Lexicon:			
BA	/ba/	yes	yes
	/ˈba/	yes	no
-GA	/-ga/	yes	yes
	/-ˈga/	yes	yes
Grammar:		MAINRIGHT > MAINLEFT + MAX_{gen} + MAX_{root}	MAINRIGHT > MAINLEFT

'yes': compatible with the grammar in the column. 'no': incompatible.

Table 5.22: *Lexicon and grammar for the Rich Base and non-Rich Base Grammars (right-stressed language).*

LEFT + MAX_{gen} + MAX_{root}. To qualify as a non-Rich Base Grammar, a set of constraint weights must at minimum meet the (less stringent) weight inequality: MAINRIGHT > MAINLEFT. Since the constraint weight settings that meet the weight inequality for the Rich Base Grammar form a proper subset of the constraint weight settings that meet the weight inequality associated with the non-Rich Base one[16], only weight settings that meet the less stringent weight inequality for non-Rich Base Grammars and fail to meet the more stringent weight inequality for Rich Base Grammars will be classified as instances of a non-Rich Base Grammar.

5.3.5 Results

The results section for the categorically right-stressed language is organized as follows. In §5.3.5.1, I report that the 136 (of 200) trained models that tied for the equal-highest log likelihood matched the surface pattern of the training data perfectly. In §5.3.5.2, we see that these 136 models satisfied the direction of the weight inequality for the right-stressed language's Rich Base Grammar (*i.e.* MAINRIGHT > MAINLEFT + MAX_{gen} + MAX_{root}).

As previously established, satisfying the direction of the Rich Base Grammar's weight inequality only tells us that a particular trained model is able to generalize to a probabilistic right-stressed language. In order to generalize to a categorically right-stressed language, the magnitude of the

[16]This is due to the restriction that constraint weights cannot be negative.

weight inequality needs to be large enough. Hence, the two generalizations tasks were introduced. The first of these tasks was wug word test §5.3.6.1, in which a trained model could potentially rely on the "easier" non-Rich Base UR /ba/ to make up for insufficient magnitude. The second task was a loan word test §5.3.6.3, in which the trained models could only depend on a Rich Base Grammar weight inequality of sufficiently large magnitude in order to pass the generalization test.

The proportion of these 136 perfectly trained models that learned the Rich Base UR /ˈba/ versus those that learned the non-Rich Base UR /ba/ is reported in §5.3.6.2.

5.3.5.1 Matching WORD-SR frequencies of training data

I ran 200 randomly initialized simulations. A histogram of the log likelihoods of these 200 trained models (Figure 5.3) indicated that the majority of the trained models had likelihoods in the highest bin, which ranged from -5.75 to -5.44.

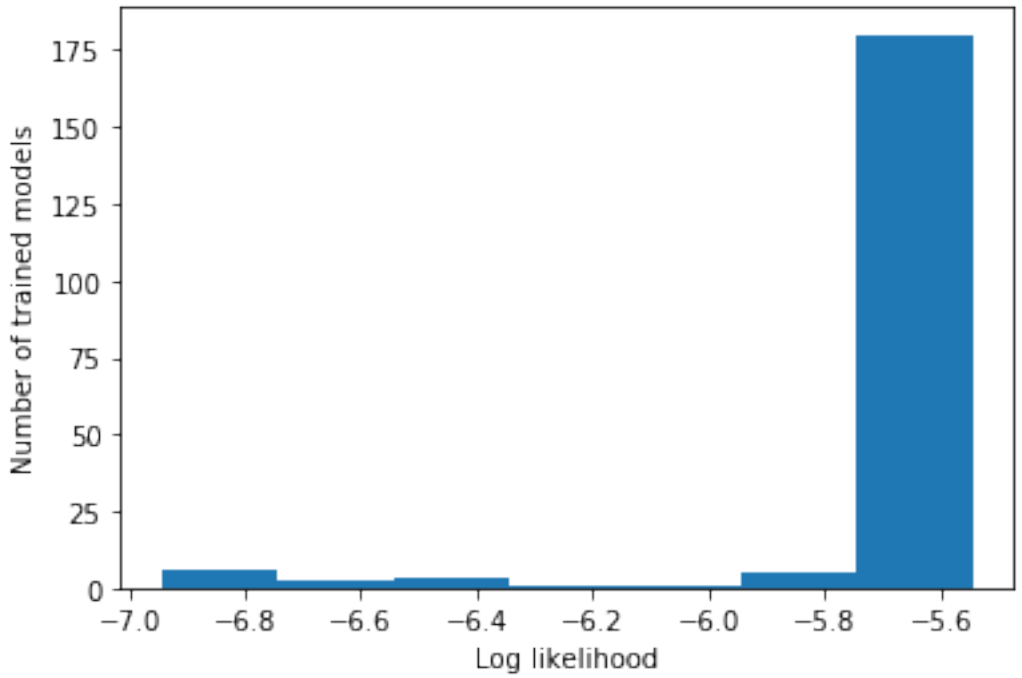

Figure 5.3: *Distribution of log likelihoods of trained models (right-stressed).*

A finer-grained visualization (Figure 5.4) indicated a cluster of highest log likelihoods between -5.550 and -5.545, so I considered trained models whose final log likelihoods fell above -5.550

as tying for the equal-highest final log likelihood. 136 trained models met this criterion.

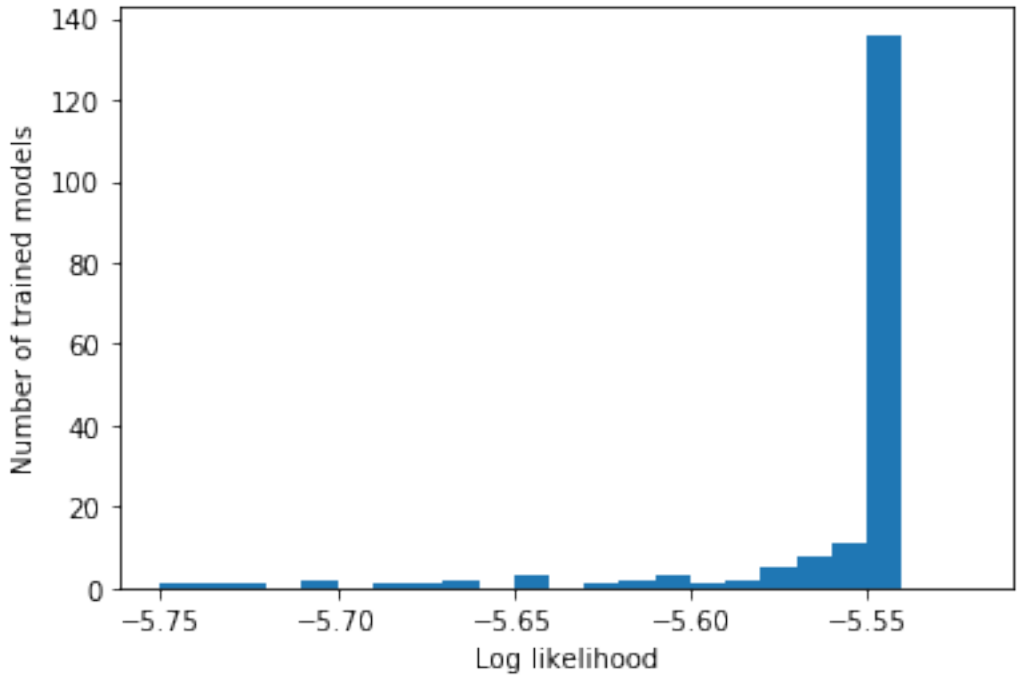

Figure 5.4: *Distribution of log likelihoods of trained models, which were higher than –5.75 (right-stressed).*

Applying Eq 5.4 to the training data for the right-stressed language gives the value of the theoretical ceiling log likelihood:

$$\begin{aligned} \text{ceiling log likelihood} &= \sum_{x \in \mathcal{D}} ln(P(x)) \\ &= \begin{matrix} ln(P(\text{(PAKA, [pa'ka])})) + ln(P(\text{(PAGA, [pa'ga])})) \\ + ln(P(\text{(BAKA, [ba'ka])})) + ln(P(\text{(BAGA, [ba'ga])})) \end{matrix} \\ &= ln(1/4) + ln(1/4) + ln(1/4) + ln(1/4) \\ &= -5.545. \end{aligned} \tag{5.12}$$

All 136 trained models that tied for the equal-highest log likelihood essentially reached the ceiling log likelihood (Eq 5.12). Of these 136 trained models, the most aberrant one had a log likelihood that was a mere .0002 away from the theoretical ceiling.

All 136 of these constraint weight settings were unique[17]. Since these 136 trained models essentially reached the theoretical ceiling log likelihood, they matched the training data perfectly.

In other words, the learner found 136 unique weight settings that could perfectly produce the WORD-SR pair frequencies that consisted the training data for the right-stressed language.

5.3.5.2 Proportion of Rich Base Grammars

The surface pattern for the right-stressed language could be generated via a Rich Base Grammar (§5.3.2) or a non-Rich Base one (§5.3.3). To recap, the weight inequality required to be a Rich Base Grammar was: MAINRIGHT $>$ MAINLEFT + MAX_{gen} + MAX_{root}. All 136 of the trained models that matched the surface pattern of the training data perfectly also had a weight inequality that was in the correct direction. That is, to say the weight of MAINRIGHT was larger than the sum of the weights of MAINLEFT, MAX_{gen} and MAX_{root} for each of these 136 trained models. In fact, all 200 trained models satisfied the direction of the the weight inequality that was required to be a Rich Base Grammar.

When a trained model met the direction of the weight inequality for a Rich Base Grammar, it produced a language in which the right-stressed SR received the highest probability for a given UR. However, merely giving the right-stressed SR the highest probability out of all SRs under consideration for a particular UR was not good enough. The language that these models were trained on was a categorically right-stressed language, rather than a language in which right-stress was the most common but left-stress was possible too. In other words, in addition to the weight inequality being in the correct direction, the magnitude of the inequality also needed to be large enough.

The wug and loan word generalization tasks §5.3.6 were designed to check whether the magnitude of the weight inequality acquired by the trained models was large enough for each model to be classified as having learned the Rich Base Grammar.

[17] In fact, all 200 local maxima found by the learner were unique.

5.3.6 Generalization to test sets

The wug word and loan word generalization tasks were introduced to assess whether the Rich Base Grammar weight inequality, which was acquired by all 136 perfectly trained models, was of sufficient magnitude to generalize categorical right-stress to novel words. The criterion for a categorical outcome was set at 99%. To be concrete, a trained model passes the generalization test for a novel test word if it assigns at least 99% of the predicted probability mass to the right-stressed SR candidate.

Similar to the test items for the left-stressed language, the wug test item consists of one trained morpheme and one new morpheme (§5.3.6.1), while the loan word test item consists of only new morphemes (§5.3.6.3).

In order to design a rigorous test, the test words were chosen to reflect the most difficult case – that of the implausible UR – whenever possible[18]. For the right-stressed language, the implausible UR has the following shape: /$\acute{\sigma}\sigma$/ – in other words, bearing stress on the leftmost syllable but not on the rightmost syllable.

5.3.6.1 Wug test

The wug word was BA-FO, which consisted of the new (*i.e.* untrained) suffix -FO, and the old (*i.e.* trained) root BA. Since the models were trained on BA, it was each trained model's job to fill in what it had learned about this root – that is, whether this root carried underlying stress or not. As for the new suffix -FO, since this morpheme was entirely new, I could dictate the underlying form that the models should use. To make the generalization task more challenging, I decided that -FO should be underlyingly stressless (*i.e.* /-fo/), since this would create wug URs (*i.e.* /ˈba-fo, ba-fo/) in which the rightmost syllable was underlyingly unstressed[19].

As before, each model's task was to predict the conditional probability of the right-stressed SR

[18]In the wug test, it is not possible to dictate that the first syllable always bears stress. This is because the first syllable comes from a trained morpheme, and a trained model would utilize the UR of the morpheme that it had learned during training.

[19]Recall that the language to which extend the novel words was a right-stressed language.

candidate [baˈfo] given the wug word BA-FO. Specifically, I considered a trained model to have successfully generalized the categorical right-stressed language to a novel $\text{ROOT}_{\text{trained}}$-$\text{SUFF}_{\text{novel}}$ word if the conditional probability $P([\text{baˈfo}]|\text{BA-FO})$ was at least 99%.

The relevant WORD-UR-SR triples for BA-FO are shown in Table 5.23. Just as we have seen

WORD-UR-SR
$P(\text{BA-FO}, /\text{ˈba-fo}/, [\text{ˈbafo}])$
$P(\text{BA-FO}, /\text{ˈba-fo}/, [\text{baˈfo}])$
$P(\text{BA-FO}, /\text{ˈba-fo}/, [\text{bafo}])$
$P(\text{BA-FO}, /\text{ba-fo}/, [\text{ˈbafo}])$
$P(\text{BA-FO}, /\text{ba-fo}/, [\text{baˈfo}])$
$P(\text{BA-FO}, /\text{ba-fo}/, [\text{bafo}])$

Table 5.23: *WORD-UR-SR triples for the wug word BA-FO.*

previously, the conditional probability $P(\text{SR} = s|\text{WORD} = w)$ can be calculated from the joint probability over triples ($P(\text{WORD, UR, SR})$) as follows:

$$P(\text{SR} = s|\text{WORD} = w) = \frac{\sum_{u'} P(w, u', s)}{\sum_{u'} \sum_{s'} P(w, u', s')}. \tag{5.13}$$

Eq 5.13 can be thus applied to the right-stressed SR [baˈfo] for the given WORD BA-FO:

$$P([\text{baˈfo}]|\text{BA-FO}) = \frac{P(\text{BA-FO}, /\text{ˈba-fo}/, [\text{baˈfo}]) + P(\text{BA-FO}, /\text{ba-fo}/, [\text{baˈfo}])}{\begin{array}{c} P(\text{BA-FO}, /\text{ˈba-fo}/, [\text{ˈbafo}]) + P(\text{BA-FO}, /\text{ˈba-fo}/, [\text{baˈfo}]) \\ + P(\text{BA-FO}, /\text{ˈba-fo}/, [\text{bafo}]) + P(\text{BA-FO}, /\text{ba-fo}/, [\text{ˈbafo}]) \\ + P(\text{BA-FO}, /\text{ba-fo}/, [\text{baˈfo}]) + P(\text{BA-FO}, /\text{ba-fo}/, [\text{bafo}]) \end{array}}. \tag{5.14}$$

It turned out that all 136 trained models that reached the ceiling log likelihood were able to generalize right-stress to the wug word that had a novel stressed suffix at 99% or greater. In fact, all 200 trained models generalized to the right-stressed SR [baˈfo] for the wug word BA-FO at rates greater than 99%.

5.3.6.2 Lexicon

Because the choice of the UR of the trained root BA was left up to the trained models, there were two hidden "unobserved" sources from which [baˈfo] could have arisen: /ˈba-fo, ba-fo/. Indeed,

the UR that a particular trained model used have been fully /ˈba-fo/, fully /ba-fo/, or a mixture of the two.

I considered the UR to be fully /ˈba-fo/ when the conditional probability of /ˈba-fo/ given the wug word BA-FO, $P(/\text{ˈba-fo}/|\text{BA-FO})$, was at least 99%. The same 99% threshold applied likewise for the UR to be considered fully /ba-fo/. If a trained model did not meet these two thresholds, then it was considered to have a "mixed" UR for BA-FO. To calculate the relevant conditional probabilities the following equation was used:

$$P(\text{UR} = u|\text{WORD} = w) = \frac{\sum_{s'} P(w, u, s')}{\sum_{u'} \sum_{s'} P(w, u', s')}. \tag{5.15}$$

Of the 136 trained models that matched the surface pattern of the training data perfectly, 57 (41.9%) used the implausible Rich Base UR /ˈba-fo/, and 79 (58.1%) used the non-Rich Base UR /ba-fo/. None of these perfect trained models used a mixture of /ˈba-fo/ and /ba-fo/.

Since a good proportion of these trained models (58.1%) used the "easier" non-Rich Base UR, their ability to generalize right-stress to a $\text{ROOT}_{\text{trained}}$-$\text{SUFF}_{\text{novel}}$ novel word cannot be definitively attributed to the grammar alone. In other words, we still do not know whether the grammars acquired by each of these 136 trained models were good enough to extend categorical right-stress to a novel word (*i.e.* a true Rich Base Grammar for a categorical right-stress language that can handle any UR), or whether the wug test result in this section was due to a serendipitous combination of the "easier" non-Rich Base UR /ba-fo/ with a non-Rich Base Grammar.

The loan word generalization task (§5.3.6.3) isolates the successful generalization of categorical right-stress to the grammar alone. Hence, the loan word task distinguishes between trained models that have acquired a true Rich Base Grammar versus models that rely on an "easier" non-Rich Base UR to pass the wug word generalization task.

5.3.6.3 Loan word test

The loan word ZEMO did not contain any old "trained" morphemes. All trained models used exactly the same UR /ˈzemo/. As before, the stress pattern of this loan word was designed to

bear the stress pattern of the implausible UR, which in the case of the right-stressed language was: /σ́σ/. I considered a model to have successfully generalized the categorical right-stressed language to loan words only when the conditional probability of the right-stressed SR candidate [zeˈmo] for the given loan word ZEMO was at least 99% (*i.e.* $P([\text{zeˈmo}]|\text{ZEMO}) \geq .99$).

133 of the 136 trained models that perfectly matched the surface pattern of the training data were able to generalize categorical right-stress to the loan word ZEMO from the implausible left-stressed UR /ˈzemo/[20]. The three remaining models generalized right-stress to /ˈzemo/ at a rate greater than 97.5% but lower than 99%. In other words, 97.8% (133/136) of these perfect trained models had learned a constraint weight setting in which the required weight inequality to qualify as a Rich Base Grammar was both (1) in the correct direction (produced right-stress) and (2) large enough (produced catgoricity when the threshold was set at 99%). A small minority (2.2%) of these models learned a weight setting that was in the correct direction (produced right-stress), but was not large enough to generalize categoricity to the loan word /ˈzemo/, which was underlyingly left-stressed.

Unsurprisingly, a check of these three trained model's lexicon showed that they had learned the "easier" non-Rich Base UR /ba/ for the root BA. That is to say, for each WORD that was part of the training data, these models did not posit any left-stressed URs. Accordingly, it was not necessary for these models to learn a grammar that could realize left-stressed URs as right-stressed SRs. Choosing to go with the "easier" non-Rich Base /ba/ also meant that these three models did not need to encounter any left-stressed URs even with the $\text{ROOT}_{\text{trained}}$-$\text{SUFF}_{\text{novel}}$ wug word BA-FO, for which they employed the UR /ba-fo/. However, the left-stressed loan word ZEMO /ˈzemo/ exposed these three model's flaw – while they did acquire a grammar that allowed them to generalize right-stress to novel words, they could not generalize *categorical* right-stress when the threshold for categoricity was set at a rate of 99%.

Nevertheless, these results still indicated that there was an overall preference for a true Rich Base Grammar (both in terms of the weight inequality pointing in the correct direction for right-

[20]Rather interestingly, all 64 trained models that did not reach the theoretical ceiling log likelihood were able to generalize categorical right-stress to the loan word ZEMO /ˈzemo/.

stress and with a large enough magnitude for categoricity) to be acquired even when paired with a non-helpful "easy" non-Rich Base UR. Of the 136 trained models that reached the theoretical ceiling log likelihood, 79 went with the unhelpful stressless non-Rich Base UR /ba/. Despite not needing to deal with a left-stressed UR in the training data[21], only 3 of these 79 "non-Rich Base UR" models failed to acquire a true Rich Base Grammar that could handle a left-stressed UR. This is rather remarkable because these 79 models were under no pressure to acquire a true Rich Base Grammar, yet 76 of the 79 (96.2%) of them did so.

To appreciate why this is remarkable, this situation could be contrasted with that of the other 57 perfect trained models that decided that the UR of BA should be the stressed Rich Base UR /ˈba/. In the case of these 57 models, it was not surprising that this decision pushed them all towards also acquiring a true Rich Base Grammar. By deciding to go with the left-stressed /ˈba-ka/ as the UR of the training WORD BA-KA, these models had no other way to convert the left-stressed UR to the right-stressed [baˈka] other than by relying on a true Rich Base Grammar.

5.3.7 Summary: right-stressed language

I trained 200 randomly initialized models to learn a categorically right-stressed language. URs were not specified in advance, so each model could decide for itself whether to acquire the unstressed /ba/ or the stressed /ˈba/ for the root BA [22]. These models were also tasked with learning an appropriate grammar that worked in tandem with the URs it acquired to match the training data as closely as possible. The choice of /ba/ or /ˈba/ had different implications when it came to the grammar. With the "easier" unstressed UR /ba/, two weight inequalities (*i.e.* grammars) were able to produce a distribution over SR candidates (*i.e.* [ˈbaka, baˈka, baka]) that correctly assigned the majority of the probability mass to the right-stressed [baˈka]. However, with the "difficult" stressed UR /ˈba/, only one weight inequality (*i.e.* grammar) was able to produce a correct probability distribution over SRs. Thus, the "difficult" UR /ˈba/ was termed the Rich Base UR, and the weight inequality that could produce a correct probability distribution over SRs from this UR was called

the Rich Base Grammar.

Of the 200 trained models, 136 were able to match the surface pattern of the training data perfectly. An inspection of the learned constraint weights showed that all 136 of these models had acquired a Rich Base Grammar weight inequality that pointed in the correct direction. This indicated that in the worst case scenario, these 136 models could model a non-categorical right-stressed language.

The wug test showed that all 136 models generalized categorical right-stress to the novel $\text{ROOT}_{\text{trained}}$-$\text{SUFF}_{\text{novel}}$ word when the threshold for categoricity was set at 99%. In contrast, the loan word test revealed that only 133 of these models generalized categorical right-stress to a novel word. These results indicate that while all 136 perfect trained models acquired constraint weight settings whose Rich Base Grammar weight inequality was in the correct direction, for three of these models, their Rich Base Grammar weight inequality was not of sufficient magnitude to produce categorical right-stress in novel loan words. In other words, these three models relied on a combination of the "easier" non-Rich Base UR /ba/ and the grammar in order to attain categoricity for the training data as well as in the wug test. Once the ability to rely on /ba/ was removed in the loan word test (in which no trained morphemes appeared), these three models' ability to generalize to the right-stressed SR fell below 99%. Nevertheless, 96.2% of the trained models that learned the "easier" non-Rich Base UR /ba/ were able to acquire a constraint weight setting that was a true Rich Base Grammar (both in terms of direction and magnitude), because they could generalize categorical right-stress to novel words without needing to rely on the non-Rich Base UR.

Overall, the results of the generalization tasks for the categorically right-stressed language were mostly similar to that of the categorically left-stressed language. In both of these languages, there was an overwhelming preference for the Rich Base Grammar to be learned when no UR was assumed in advance. In both the categorically left- and right-stressed languages, there was also a very strong preference for the Rich Base Grammar to be learned by models that did not also acquire a Rich Base UR. However, while this pattern was absolute in the case of the categorically

[22]These models also had to learn whether the suffix -GA was /-ga/ or /-ˈga/. However, this choice made no difference for the right-stressed language, so it was not discussed.

left-stressed language, it was no longer absolute for the categorically right-stressed language.

5.4 Conclusion

In this chapter, I presented two case studies (a categorically left-stressed language and a categorically right-stressed language) in which the surface pattern could theoretically be modeled equally well by a Rich Base Grammar or a non-Rich Base one. I ran 200 randomly initialized simulations for each case study, and collected the trained models that tied for being the best matches of the surface pattern. A check of the log likelihoods of these best models showed that they were global maxima. Thus, the collection of these best models represents a random sample of the global maxima.

I found that 100% of the sampled global maxima for the left-stressed language and 97.8% of the sampled global maxima for the right-stressed language corresponded to a Rich Base Grammar. These best models passed a generalization task to novel nonce and loan words, in order to be considered as having acquired a Rich Base Grammar. In addition, I also found that the acquisition of Rich Base Grammar was not contingent on also having acquired the Rich Base UR; 47.6% of the sampled global maxima for the left-stressed language had learned a Rich Base Grammar without acquiring the associated Rich Base UR, and 59.4% of the sampled global maxima for the right-stressed language had learned a Rich Base Grammar without acquiring the associated Rich Base UR.

In sum, my simulations found an overwhelming preference for the Rich Base Grammar to be learned.

CHAPTER 6

Model comparisons

In the preceding Richness of the Base chapter (Ch 5), we encountered an overwhelming preference for the Rich Base Grammar to be learned. This occured even when a model learned an unhelpful easy non-Rich Base UR that would not necessitate a Rich Base Grammar. In this chapter, I dig deeper into the characteristics of my over-arching model to explore why the overwhelming preference to learn the Rich Base Grammar exists.

In §6.1, I provide a stark illustration of how skewed the distribution of trained models toward Rich Base Grammars is. Synthesizing the properties of my over-arching model of joint hidden-structure-and-grammar with the properties of the Rich Base problem itself, I deduce that the learner experiences the pressure to learn the Rich Base Grammar at every single point in the solution curve (§6.2). In §6.3, I discuss the particular properties of my model that promote this omnipresent preference towards Rich Base Grammars.

The main result of §6.2 indicates that every single point (*i.e.* every single weight vector) in the solution curve has a gradient that points in the direction of learning a Rich Base Grammar. It follows that there are no global maxima in the solution curve (at any local or global maximum, the gradients should be zero[1]). Yet, in this dissertation, my general approach has been to collect a random sample of the global maxima, which I treat as a simulated population of language learners!

In §6.4, I reconcile my general approach with the global-maximum-less solution curve. I show that the randomly sampled points that I have collected are in the close vicinity of "the missing global maximum" when the vicinity is assessed on the model fit to data rather than being assessed on the parameter weights. Accordingly, the randomly sampled points in the global-maxima-less

[1]More precisely, each element in the vector of first-order partial derivatives should be zero. Going forward, I use the term 'gradient' as shorthand for the vector of first-order partial derivatives.

solution curve are effectively the same as those that would have been sampled in a solution curve with global maxima when an iterative method of optimization is used (§6.4.1).

In §6.4.3, I show that global-maxima-less solution curves occur when unobserved outcomes are included in the candidate set of an (over-parameterized) MaxEnt (*i.e.* log-linear) model. This configuration of candidates is often encountered in models of categorical phonological phenomena. This section closes with a discussion of appropriate optimization methods in a global-maxima-less solution curve (§6.4.4).

Constraint-based models of phonology that adjust constraint weights (*i.e.* parameter weights) to fit the model to data often suffer from over-parameterization. The Harmonic Grammar-family of models, of which the MaxEnt model is a member, falls into the class of weighted constraint models. In phonological models, a parameter is often included in a model to maintain the constraint conflict worldview of Optimality Theory (OT). For example, both *D and FAITH are included in a MaxEnt model even though FAITH only ever prefers losing candidates.

In §6.5, I present two schematic MaxEnt models: one whose solution curve has an infinite number of global maxima, and another whose solution curve is global-maxima-less. The first solution curve is from an over-parameterized MaxEnt model. The second solution curve is from an over-parameterized MaxEnt model that has unobserved SR candidates. I show that the introduction of a Gaussian prior-type regularization term transforms both of these solutions curves into having a single global maximum (§6.5.3).

The existence of log conditional likelihood functions (*i.e.* solution curves) with an infinite number of global maxima or no global maximum is not consistent with Goldwater and Johnson's assertion that MaxEnt models have only one global maximum. In §6.5.5, I propose a refinement of Goldwater and Johnson's generalization, which leaves room for solution curves that have these two shapes.

I discuss over-parameterization in §6.6. Moving from an OT world of strict rankings to a MaxEnt world of weighted constraints where each constraint is a free parameter is one way in which over-parameterization enters into MaxEnt models (§6.6.1). Regularization may be successful in producing a single global maximum in solution curves that previously lacked one. Nevertheless,

for MaxEnt models with this flavor of over-parameterization, I demonstrate that regularization cannot help us with identifiability (*i.e.* recovering each constraint's one and only true weight from the observed data).

In §6.7, I apply my approach of simulating a population of language learners to the identifiability problem. If we interpret each trained MaxEnt model as a single human speaker, then being unable to recover a constraint's true weight from the observed data is not so bad after all. Each speaker only needs to learn one setting of constraint weights that works for them. A speaker does not need to recover the 'one true value' of each constraint.

Under this perspective, the observed data is the combined output of speakers that each run their own unique MaxEnt model. Since the observed data is not the output of a single source, the 'one true value' of each constraint may not hold much meaning.

§6.8 summarizes the highlights of this chapter.

6.1 Visualizing the overwhelming preference for Rich Base Grammars

We have seen that for both case studies, there was an overwhelming preference for a Rich Base Grammar to be acquired by the trained models, regardless of whether the models also acquired the Rich Base UR. This result is highly improbable. To see why this is so, let us consider the available grammars. For each language, there are three available grammar classes:

(12) Rich Base Grammar

(13) Non-Rich Base Grammar

(14) Neither (RBG nor non-RBG)

A grammar is a vector of phonological parameter weights (*i.e.* it includes only the weights of the phonological constraints and not the weights of the UR constraints). A phonological constraint's weight can only take on a non-negative real value. A grammar falls into a particular class of grammars (*e.g.* Rich Base Grammar, non-Rich Base Grammar, neither) when its parameter weights fulfill certain inequalities. For example, in the left-stressed language, there are three relevant pa-

rameters whose weights determine whether the grammar learned by a trained model falls into a particular grammar class. These three parameters are: MAINLEFT, MAINRIGHT, and $\text{MAX}_{\text{general}}$. When the threshold for categoricity is set to 99% and there are two competing candidates, the corresponding difference in the weight inequality is 4.6 $(= ln(.99))$.

Let the triple <MAINLEFT, MAINRIGHT, and $\text{MAX}_{\text{general}}$> be abbreviated as $< l, r, g >$.

(15) For the left-stresed language, $< l, r, g >$ is in the set of

a. Rich Base Grammars if and only if: $l - r - 4.6 \geq g$.

b. Non-Rich Base Grammars if and only if: $g > l - r - 4.6 \geq 0$.

c. Neither (RBG nor non-RBG) if and only if neither of the criteria above are met.

The scatterplot in Figure 6.1[2] illustrates the skewed distribution of the best trained models amongst the Rich Base Grammar and non-Rich Base Grammar classes. The 145 trained mod-

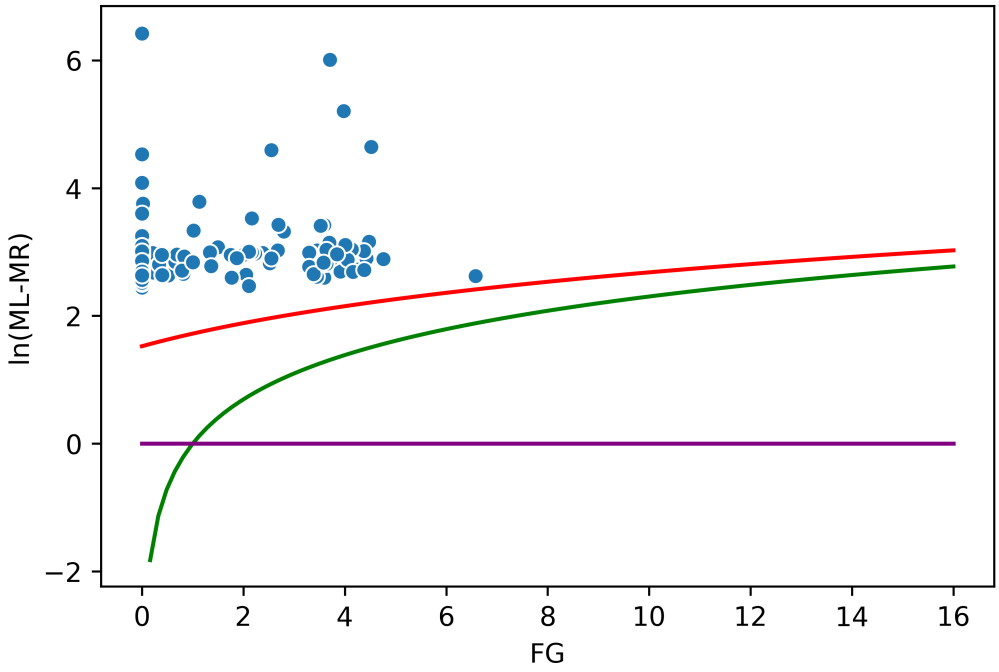

Figure 6.1: *Distribution of trained models that acquired a Rich Base Grammar vs. a non-Rich Base Grammar (left-stressed language). ML:* MAINLEFT. *MR:* MAINRIGHT. *FG:* $\text{MAX}_{general}$.

els that tied for the equal-highest likelihood have attained the ceiling likelihood, so they fit the

[2]I use a logarithmic scale for the vertical axis because the numbers get very big; otherwise, we cannot see much.

training data perfectly and cannot possibly belong to the third grammar class: Neither (RBG nor non-RBG). The lower green curve represents the threshold for a minimally Rich Base Grammar. *i.e.* A grammar that turns underlying right-stressed $/\sigma\acute{\sigma}/$ to surface left-stressed $[\acute{\sigma}\sigma]$ at a rate of at least 50% will appear above the lower green curve. The higher red curve represents the threshold for a categorical Rich Base Grammar. *i.e.* A grammar that turns underlying right-stressed $/\sigma\acute{\sigma}/$ to surface left-stressed $[\acute{\sigma}\sigma]$ at a rate of at least 99% will appear above the upper red curve. The horizontal purple line represents the dividing line between non-Rich Base Grammars and the grammars that are neither Rich Base nor non-Rich Base. Thus, the "neithers" fall below the purple line, the non-Rich Base Grammars between the purple and red lines, and the Rich Base Grammars above the red line.

For the right-stressed language, all four phonological parameters are relevant. Let the 4-tuple $<$MAINLEFT, MAINRIGHT, $\text{MAX}_{\text{general}}$, and $\text{MAX}_{\text{root}}>$ be abbreviated as $< l, r, g, t >$.

(16) For the right-stresed language, $< l, r, g, t >$ is in the set of

a. Rich Base Grammars if and only if: $r - l - 4.6 \geq g + t$.

b. Non-Rich Base Grammars if and only if: $g + t > r - l - 4.6 \geq 0$.

c. Neither (RBG nor non-RBG) if and only if neither of the criteria above are met.

The distribution of the 136 trained models that tied for the best likelihood amongst the Rich Base Grammar and non-Rich Base Grammar classes is shown in Figure 6.2. As before, the "neithers" fall below the purple line, the non-Rich Base Grammars between the purple and red lines, and the Rich Base Grammars above the red line.

The two scatterplots in Figure 6.1 and Figure 6.2 are a stark visualization that the ceiling-likelihood trained models are heavily skewed towards being Rich Base Grammars. For the left-stressed language, all trained models at the effective ceiling have learned categorical Rich Base Grammars. For the right-stressed language, only three (out of 136) models miss the criterion to be classified as a categorical[3] Rich Base Grammar (they are just below the red curve).

[3]These three models just barely miss the criterion to be categorical Rich Base Grammars. The worst one generalizes at a rate slightly above 97.5%. *cf.* The criterion I had set to be classified as a Rich Base Grammar was generalizing at a rate of 99% or higher.

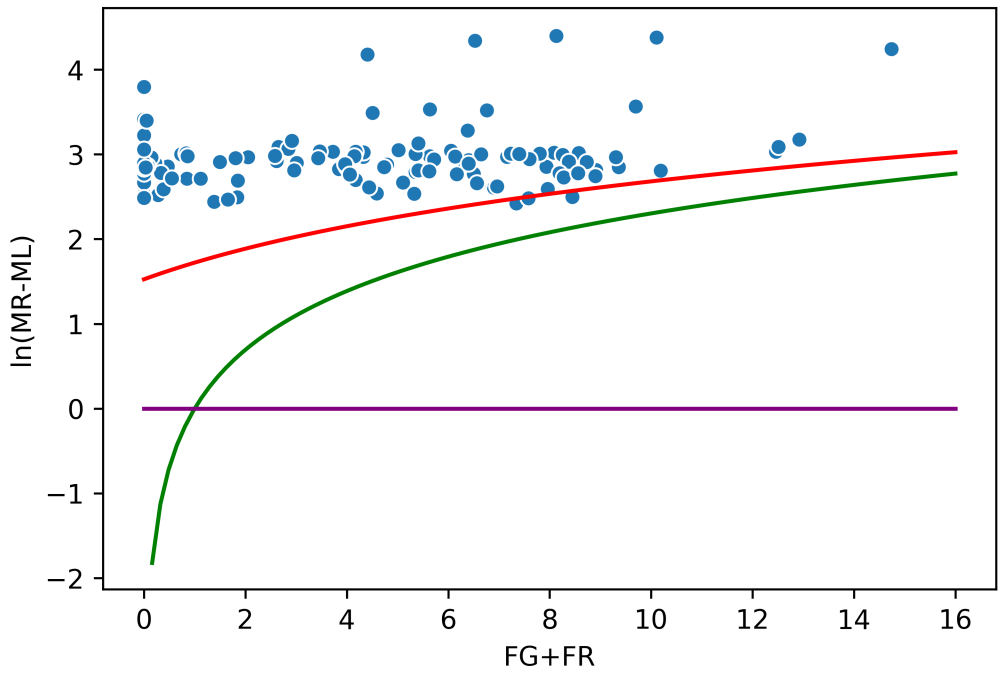

Figure 6.2: *Distribution of trained models that acquired a Rich Base Grammar vs. a non-Rich Base Grammar (right-stressed language). ML:* MAINLEFT. *MR:* MAINRIGHT. *FG:* $\text{MAX}_{general}$. *FR:* MAX_{root}.

6.2 The pressure to learn a Rich Base Grammar is present throughout the solution space

In the preceding section, I showed that there was an overwhelming preference for Rich Base Grammars. This section explores why this overwhelming preference for Rich Base Grammars exist from two distinct angles. From the machine learning perspective, I deduce that there are more points in the solution space whose gradients point towards Rich Base Grammar global maxima than there are pointing toward non-Rich Base Grammar global maxima. Another way to reason about the overwhelming preference for Rich Base Grammars is to approach it from the details of the phonological phenomenon. From this bottom-up perspective, I deduce that at every point in the solution space, the learner is always incentivized towards a Rich Base Grammar. To translate this to machine learning terms, every single point in the solution space has gradients that point towards a Rich Base Grammar. Finally, I show that whenever a log-linear model is used in phonological modeling (whether (1) singularly as in the majority of phonological modeling or (2) as a sub-model within

a larger model as in my proposed over-arching model), the global maxima can often become undefined, which corresponds to the ceiling likelihood being an asymptote. Such cases occur when there is a mismatch between the training data and the output (*i.e.* candidate) classes when the model is over-parameterized. This has implications for the choice of optimization techniques (*i.e.* using numerical approaches over an analytical solver).

Let us first take the machine learning perspective. Recall that the learner uses the Expectation-Maximization algorithm to find a local maximum by hill-climbing from a randomly initialized point within the solution space. The solution space is the log likelihood function of the marginal distribution (Eq (3.9)) of the model architecture defined in §3.2. Hill-climbing is guided by the gradients of the solution space at the current point. Thus, from the perspective of machine learning, the preference for converging at maxima corresponding to a Rich Base Grammar indicates the following: within the solution space, there are more points with gradients pointing towards maxima corresponding to a Rich Base Grammar than there are points with gradients pointing towards maxima that correspond to the non-Rich Base Grammar.

Let us now take the bottom-up perspective, starting with the details of phonological problem itself. In the Rich Base problem, there are two competing grammars – the Rich Base Grammar and the non-Rich Base Grammar. The Rich Base Grammar is able to model a proper superset of the data that the non-Rich Base Grammar can model. In other words, the Rich Base Grammar is much more useful than the non-Rich Base one because the former is able to handle all the URs that the latter can handle and then some. In order to have a concrete running example, let us return to the language that opened our Rich Base study in Chapter 5: the language in which all voiced obstruents (*e.g.* [b, d, ɡ, v, z, ʒ, dʒ]) are banned. For this language, the Rich Base Grammar devoices voiced obstruents while the non-Rich Base Grammar basically does nothing. In effect, the non-Rich Base Grammar results in illegal voiced obstruents surfacing. The Rich Base UR /tag/ is challenging to deal with because the final voiced /ɡ/ is banned in the language, and must somehow be handled before surfacing. The non-Rich Base UR /tak/ is easy to deal with because all of its sounds are legal, and it can surface faithfully. The Rich Base Grammar is more useful because it produces the correct surface form [tak] for both the "challenging" Rich Base /tag/ (via devoicing the final /ɡ/) and the "easy" non-Rich Base /tak/. This contrasts with the non-Rich Base Grammar, which

produces the incorrect surface form *[tag] for /tag/ since this grammar basically does nothing to the final /g/.

We now have a clear picture of the asymmetric utility of these two classes of grammars with respect to the "challenging" UR /tag/ and the "easy" UR /tak/. Next, I will plug these two classes of grammars and the two URs into my over-arching model. Let us imagine that we have randomly initiated the learning process at the following distribution over URs: 80% /tak/ and 20% /tag/. Models that have a Rich Base Grammar can handle 100%[4] of the data because they are able to change the illegal /g/ in /tag/ to a legal [k] in [tak]. In contrast, models that have a non-Rich Base Grammar can only handle 80% of the data: they are stuck with the illegal /g/, which leads them to incorrectly produce surface [tag] 20% of the time. The latter model has a poorer fit to data; so it is incentivized to change something in order to handle the "challenging" Rich Base UR /tag/. Since my learner can simultaneously adjust both the parameters that produce the distribution over URs and the grammar, the best strategy is to both decrease the probability of the "challenging" UR /tag/ and learn a Rich Base Grammar that can deal with the any remaining /tag/'s.

Let us imagine that the latter model has now made one update to its parameters, and now has the following distribution over URs: 95% /tak/ and 5% /tag/. It has also updated its phonological parameters, to get closer to that of a Rich Base Grammar. Unfortunately, it started off really far away from a Rich Base Grammar, so its current grammar still qualifies as a non-Rich Base Grammar. Even with this improved distribution over URs, the learner is still incentivized to learn a Rich Base Grammar because a Rich Base Grammar will allow it produce the correct SR 100%[5] of the time while the non-Rich Base Grammar still leaves it with a 5% error rate. Let us imagine an eve[6]n better distribution over URs: 99% /tak/ and 1% /tag/. Even with this much better

[4]This is a minor simplification. In my over-arching model, the grammar (made up of phonological parameters) is modeled by a MaxEnt-style log linear model. Such models never give 100% or 0% probabilities to any outcome, so instead of 100%, a value like 99.998% is more realistic.

[5]This is likewise a minor simplification since MaxEnt-style log linear models never assign 100% probability to any outcome. A more realistic value would be something like 99.997%.

distribution, the learner is still incentivized to learn a Rich Base Grammar (0%[7] error rate[8]) while the non-Rich Base Grammar has an error rate of 1%. The greater utility of the Rich Base Grammar over the non-Rich Base Grammar results in the following principle: **No matter the distribution over URs, the learner is always incentivized to learn a Rich Base Grammar**. This arises from the fact that no matter what the distribution over URs is, both URs always have at least some probability mass. This incentive is expressed as a gradient. Thus, in machine learning terms, **every single point in the solution space has a gradient that points towards learning a Rich Base Grammar**.

In slightly friendlier terms, we can imagine the learner traveling around the solution space. No matter where the learner walks to, it keeps hearing the same message "learn a Rich Base Grammar!". Sometimes the message is louder (*e.g.* the UR distribution 80% /tak/ and 20% /tag/ means a 20% improvement in error rate). Sometimes the message is softer (*e.g.* the UR distribution 97% /tak/ and 3% /tag/ means a less impactful 3% improvement in error rate). But the important takeaway is that the message is always present throughout the solution space.

6.2.1 Interim summary

Let us take stock of the discussion so far. I set out to explain the empirical finding that the overwhelming majority of global maxima that were found by the learner corresponded to having learned a Rich Base Grammar. I argued that the reason for this overwhelming preference for Rich Base Grammars follows from the learner always being incentivized to learn a Rich Base Grammar no matter what the distribution over URs currently is. Mathematically, this corresponds to every point in the solution space having a gradient that points towards learning a Rich Base Grammar. The above claim follows directly from the fact that Rich Base Grammars can handle a

[7]Again, a simplification. A more realistic value would be 0.002%.

[8]For the actual optimization process, model fitting was accomplished via maximizing the log likelihood. For the sake of exposition, I use the error rate as a proxy measure of model (mis-)fit.

[8]The distribution is 'better' because the UR distribution is closer to that of the observed surface data, thus reducing the burden that is placed on the grammar to produce the correct surface data.

proper superset of URs that non-Rich Base Grammars can. In summary, in my proposed model of concurrent hidden structure and grammar learning, there exists an ever-present pressure to learn a Rich Base Grammar that naturally emerges from the basic facts of the Rich Base problem itself.

6.3 Properties of my model that promote the omnipresent pressure towards a Rich Base Grammar

I have shown that my over-arching model of concurrent hidden structure and grammar learning leverages the asymmetric utility of the Rich Base Grammar against the non-Rich Base Grammar in order to push the learner towards learning the Rich Base Grammar. Next, I discuss three properties of my over-arching model that affect its ability to make use of the asymmetric utility of the two classes of grammars in order to learn the Rich Base Grammar (17).

(17) Properties of my model that enable it to leverage the asymmetric utility of Rich Base Grammars (over non-Rich Base Grammars) in order to incentivize the acquisition of a Rich Base Grammar throughout the solution space:

a. The architecture of my full model allows both the Rich Base UR and the non-Rich Base UR to exert their influence on the grammar at the same time

b. The particular mathematical model I use to produce the probability distribution over competing URs is one that gives some non-zero probability mass to each UR outcome

c. No regularization terms are used

Before elaborating on the first property, let us recall the architecture of my full model. My full model is composed of a single UR sub-model and a single grammar sub-model. In order to get the "challenging" UR to be under consideration **alongside** the "easy" UR, I have a single UR sub-model that produces a single probability distribution over both UR outcomes. The single UR sub-model interacts with the single grammar sub-model. This results in both URs exerting their influence on the grammar at the same time. During the course of learning, **even as the probability of the "challenging" UR decreases, it still exerts its influence on acquiring a Rich Base Grammar**. In my over-arching model, the WORD-UR sub-model produces a joint probability dis-

tribution over WORD-UR pairs, rather than directly producing a probability distribution over URs. For example, <DOG, /tak/>: 80%, <DOG, /tag/>: 20%. Nevertheless, this joint distribution over WORD-UR pairs still places URs in competition alongside each other, so this WORD-UR distribution is sufficient.

Second, having a MaxEnt (*i.e.* log-linear) model produce the distribution over URs[9] leads to a fascinating consequence. Log-linear models never assign 0% probability to any outcome, so the probability of the "challenging" UR can never reach 0%. For example, a disfavored UR like the "challenging" /tag/ can have its probability approach 0%, but never actually get to 0%. This means that the "challenging" UR will never fall out of the competition against the "easy" UR. The "challenging" **Rich Base UR, while mostly "forgotten", will always be exerting some influence (no matter how tiny) to push the grammar toward a Rich Base Grammar**.

Third, I have not used any regularization terms for the models that I have trained in the Rich Base chapter (Ch 5). This means that the sole goal of the learner is to fit the model to the data. The learner adjusts the parameter weights solely with this goal in mind. The inclusion of a regularization term would result in the learner now having two goals. In addition to adjusting the parameter weights to fit the model to the data, the learner would also have to adjust the very same parameter weights to satisfy the regularization term. I have already established that the pressure to learn a Rich Base Grammar exists at every point in the solution space. The lack of a regularization term means that there is only ever a benefit (and no cost) to learning a Rich Base Grammar. The introduction of a regularization term means that when parameter weights are shifted toward a Rich Base Grammar, doing so might now incur a cost on the regularization end. For example, a Rich Base Grammar that generalizes the left-stressed language to novel data at 99% or greater requires the value of '$\textsc{MainLeft} - \textsc{MainRight} + 4.6$' to be greater than or equal to the value of '$\textsc{Max}_{\text{general}}$'. If a regularization term is introduced to keep all phonological parameter weights low (*e.g.* by penalizing weights that deviate from 0, with greater deviations incurring a larger penalty), then the above weight inequality would be more difficult to achieve. Thus, the inclusion of such a regularization term is predicted to decrease the proportion of trained models that have learned

[9]Or specifically in my over-arching model, over WORD-UR pairs.

Rich Base Grammars.

6.4 The undefined global maxima

I have argued that the incentive towards a Rich Base Grammar[10] exists at every point in the solution space. This implies that there is always a non-zero gradient at every point in the solution space. The above statement appears to be in conflict with the general approach that I have employed in this dissertation, which involves collecting a random sample of global maxima. At a global (or even local) maximum, the gradient should be zero! If there are no points in the solution space at which the gradient is zero, then what have I been collecting? How can these two seemingly contradictory statements be reconciled? First, I establish that no true global maximum can exist because the actual ceiling likelihood is impossible to achieve. Instead, the "global maxima" that I have been collecting can be more accurately described as points in the neighborhood of the would-be "global maxima".

In the two case studies we have looked at in the Rich Base chapter (Ch 5), the actual ceiling likelihood is an asymptote. An asymptote is a line that a curve approaches when at least one of the curve's parameters approaches infinity. The solution space is the set of all possible settings of the parameters (*i.e.* the set of all n-dimensional vectors, where n is the combined number of UR and phonological constraints). This set is the domain of the objective function, so these vectors are inputs to the objective function. Up to this point, the objective function is composed of only one component: the likelihood[11], which is a measure of model fit to data. The objective function

[10]Under a MaxEnt phonological analysis of categorical phenomena, there is often no such thing as having arrived at a grammar. This is unlike the world of strict ranking. In MaxEnt, each SR candidate always receives some non-zero probability. A Rich Base Grammar is one that is able to turn underlying /g/ to surface [k]. In other words, it is impossible to find a 'perfect' Rich Base Grammar: $Pr([\text{tak}]|/\text{tag}/) = 1$. Instead, there are only grammars that get very close to producing this result: $Pr([\text{tak}]|/\text{tag}/) = .95$, $Pr([\text{tak}]|/\text{tag}/) = .97$, $Pr([\text{tak}]|/\text{tag}/) = .99$, $Pr([\text{tak}]|/\text{tag}/) = .995$, *etc.* The analyst has to decide where to draw the line in the sand for what classifies as a Rich Base Grammar. For me, I set the threshold at 99%. Trained models whose grammars turn at least 99% of the underlying /g/'s into surface [k]'s are classified as having learned a Rich Base Grammar. Even after passing the 99% threshold, a particular trained model still experiences the incentive towards improving its Rich Base Grammar, for example by learning weights that would turn more than 99% of the /g/'s into [k]'s.

[11]During implementation, the objective function consisted of the log likelihood rather than the likelihood. Maximizing both expressions is functionally equivalent. So, for the sake of readability, I use the term 'likelihood' here.

takes the n-dimensional vector of parameter weights as input; the objective function outputs the likelihood. The curve represents the objective function. In other words, the curve represents the model fit to data for a given vector of n parameter weights. When the ceiling likelihood[12] is an asymptote, it means that as the value of a parameter (*e.g.* MAINLEFT, *D[13]) increases and approaches positive infinity, the likelihood of the model gets ever closer to the ceiling likelihood without actually reaching it. In other words, the fit of the model gets ever closer to that of the observed data without actually replicating the data 100%.

The asymptotic nature of the ceiling likelihood appears when the following three conditions are present (18):

(18) The ceiling likelihood is an asymptote when:

a. The probability distribution over a variable, v, is produced by a MaxEnt model (log-linear model).

b. The set of candidate outcomes of v is a proper superset of the set of observed outcomes of v.

c. The model is over-parameterized[14].

The UR or the SR is an example of a variable. If we take the SR as a variable, then the candidate outcomes are the individual SRs under consideration (*e.g.* [tak, tag]). If the set of observed outcomes is a proper subset of the set of candidate outcomes (*e.g.* only [tak] is observed), then the ceiling likelihood becomes an asymptote (*i.e.* it becomes impossible to match the observed training data exactly).

To see why this is so, recall that the MaxEnt (log-linear) model can never assign 0% probability to any outcome. Within my joint hidden-structure and grammar model, this means that the "chal-

[12]Here, I return to using 'likelihood' in place of 'objective function' since they are equivalent for everything we have seen so far and are equivalent for the ensuing cases. The points that I will be making recall the model fit to data, which is directly linked with the 'likelihood', and is at least one degree removed from the concept of the 'objective function'. The objective function will reappear in §6.5, once the introduction of regularization breaks the equivalence.

[13]*D is a phonological parameter that bans voiced obstruents.

[14]I discuss over-parameterization in §6.5.3 and §6.6.

lenging" UR /tag/ will get some non-zero probability. For the sake of exemplification, let's say it /tag/ has 0.5% probability. Since the underlying /g/ is banned from surfacing in the language, the grammar has to devoice it to a [k]. The *D constraint bans voiced obstruents (like [g]) while the FAITH constraint works to retain the underlying consonant. The following weight inequality thus indicates a devoicing grammar: *D $\gg$ FAITH. The larger the difference between the weights of *D and FAITH, the more underlying /g/'s turn to surface [k]'s. For example, a difference of 4.6 produces $.99$ surface [tak]'s for each underlying /tag/ ($Pr([\text{tak}]|/\text{tag}/) = 99.0\%$). A difference of 6.9 produces $.999$ surface [tak]'s for each underlying /tag/ ($Pr([\text{tak}]|/\text{tag}/) = 99.9\%$). In our example, the SR [tak] was observed 100% of the time. However, the nature of the MaxEnt model means that while the predicted proabability of [tak] can approach 100%, it can never actually reach the actual observed value. Since the ceiling likelihood indicates a perfect model fit to the observed data, the inability of the model to fit the data exactly means that the ceiling likelihood is out of reach.

For the sake of argument, let us imagine a global maximum at the ceiling likelihood. At this global maximum, there is a perfect model fit to the data, so there is no incentive to change any parameter weights. Thus, the gradients at this global maximum would be 0. However, I have already shown that no such global maximum can exist in my joint hidden structure and grammar models of the Rich Base languages we have looked at. Global maxima are undefined in my model. Hence, it is possible for every single weight vector in the solution space to have a gradient that points it towards learning a Rich Base Grammar.

6.4.1 Parameter estimates in the vicinity of a global maxima are good enough estimates

The approach that I have employed in Chapter 4 (Velar Softening) and Chapter 5 (Richness of the Base) relies on collecting a random sample of global maxima and then studying the properties of the sample. If the global maxima are undefined, then what exactly are the objects that I have randomly sampled? In what follows, I argue that the points on the solution curve that I have collected are in the neighborhood of a global maximum, and that there is no effective difference between the objects that I have randomly sampled and the "global maxima" that an iterative method

of optimization would have collected in cases where the global maxima exist. First, I explain how the iterative method of optimization finds estimates for the parameter weights when the function it is maximizing has defined global maxima. Thereafter, I show that the factors that drive and halt the optimization process for well-behaved models with global maxima function in the same way for my global-maximum-less model. Accordingly, there is no effective difference between the objects collected via optimization from these two types of models.

In general, iterative methods[15] use the gradient at the current point (and sometimes also the history of gradients at previous points) to steer their way to the top of a nearby hill. As we get closer to the peak, the slopes get shallower and shallower; this corresponds to gradients getting closer to 0. At the peak, the gradient is 0. As the algorithm proceeds closer to the peak, the shallower slopes correspond to smaller improvements in the model fit to data. In practice, the algorithm halts when the improvement of the model fit to data falls below a certain threshold. In my learner, I had set the threshold to $< .0001$ improvement to the log-likelihood. Halting the learning algorithm when the improvement of the model fit to data falls below a certain threshold (rather than when the gradient at the current point is 0) means that we stop the learning process when we are in the close vicinity of a maximum. We could have landed right on the maximum itself (if we were lucky!) but more than likely we are at some point very close to the maximum. Usually, we treat this close-enough-but-not-quite-the-maximum point as effectively being at the maximum since the true maximum is very close by. Under an iterative method, getting to the actual maximum is impractical. Once we are very close to the maximum, the next update might overshoot the maximum. The following update might then try to correct course by heading back in the opposite direction, yet overshooting the maximum again. Thus, the learner can bounce back and forth many times without ever landing on the maximum. The technique of halting the learning process once the improvement of the model fit to data falls below a certain threshold allows us to get a close enough estimate of a maximum without wasting time bouncing around its vicinity.

[15] I have used the L-BFGS-B algorithm, which is a quasi-Newtonian method that employs an iterative method to update parameter estimates. This algorithm makes use of the gradient and the second derivative to steer itself through the solution curve. The 'L' in L-BFGS-B stands for limited-memory; the algorithm utilizes the history of the previous m parameter estimates and the previous m gradients to guide the next step that it takes up the hill. m is usually a small number less than 10.

Under my model, the global maximum is undefined. Nevertheless, the other characteristics of my model are identical to the well-behaved model described above. As we get closer to the would-be "global maximum", the slopes likewise get shallower and shallower, which corresponds to the gradients getting closer to 0, and the improvements in the model fit to data become smaller and smaller. Eventually, the algorithm halts when the slope is so shallow that the improvement in the model fit to data falls below the threshold. While the point in the solution curve at which learning stops is not the global maximum, it is in the vicinity of the "missing global maximum" when vicinity is assessed on the model fit to data. This point serves as a close enough estimate of the "missing global maximum". More importantly, the "close enough estimates" learned for this global-maximum-less model are functionally the same as the "close enough estimates" learned for the model that has a global maximum.

6.4.2 Additional support for the good enough estimates

In the Rich Base chapter (Ch 5), I have utilized the iterative method to find "close enough estimates" of the global maxima. I have treated these "close enough estimates" as good enough approximations of the parameter weights at the would-be global maxima. There are two additional reasons pertaining specifically to the cases in the Rich Base chapter that support the usage of these "close enough estimates" as good enough approximations.

First, the collection of trained models that I have treated as essentially fitting the data perfectly, are indeed very close to the would-be global maxima. The theoretical ceiling log-likelihood, which indicates a perfect model fit to the data is: $4ln(\frac{1}{4}) = -5.54518$. For both the left- and right-stressed languages, the most aberrant of the trained models that I still considered to be essentially at ceiling were a mere 0.0002 away from the theoretical ceiling log-likelihood. In other words, these trained models are all in very close proximity to a would-be global maximum when we focus on the fit of the trained models to the data.

Second, the purpose of collecting these trained models (that are essentially at ceiling) is to study their high-level properties as a group. I am not interested in the specific parameter weights that are learned by each trained model. Rather, I am interested specifically in the proportion of trained

models that have learned a Rich Base Grammar. In order to be classified as a Rich Base Grammar, a trained model needs to apply the phonological process to novel loan words at a rate of at least 99%. For example, given the novel loan word /pɛg/, a trained model needs to produce surface [pɛk] 99% of the time for me to classify it as having learned a Rich Base Grammar. A $99\% : 1\%$ distribution is a rather skewed distribution. When the phonological parameter weights change, the distribution over SRs changes along with it. However, the relation between the two is not linear. For a nearly uniform distribution like $49\% : 51\%$, a small change in the phonological constraint weights can result in a big shift in the distribution. For skewed distributions like $99\% : 1\%$, the same change in the phonological constraint weights results in a very small shift in the distribution. I have set the $99\% : 1\%$ distribution as the dividing line between Rich Base Grammars and non-Rich Base Grammars. Given how skewed the $99\% : 1\%$ distribution is, in order for a trained model to cross the dividing line, its phonological parameters weights would have to move a considerable distance from their current estimates. Hence, the "close enough estimates" can function as a good enough approximation of the parameter weights at the (would-be)[16] global maxima.

6.4.3 Categorical phonology and undefined global maxima

MaxEnt (*i.e.* log-linear) models of categorical phonological phenomena are very likely to have solutions spaces in which the global maxima are undefined. This results from the interplay between MaxEnt models and the mismatch between observed and candidate outcomes ((18), repeated in (19)):

(19) The ceiling likelihood is an asymptote when:

a. The probability distribution over a variable, v, is produced by a MaxEnt model (log-linear model).

b. The set of candidate outcomes of v is a proper superset of the set of observed outcomes of v.

[16] 'Would-be' is in parentheses because this argument holds regardless of whether the global maxima exist or whether they are undefined.

c. The model is over-parameterized[17].

To have concrete examples, let's consider the Categorical data set and the Variation data set in (6.1). Let the two SRs [tak, tag] be the competing outcomes of the same UR /tag/.

Data set	[tak]	[tag]
Categorical	10	0
Variation	7	3

Table 6.1: *Surface form frequencies for the Categorical and Variation data sets.*

When modeling the Categorical data set, phonologists usually include both [tak] and [tag] SR outcomes in their model despite the SR outcome [tag] being observed 0 times. The set of candidate outcomes has two members, {[tak], [tag]}, while the set of observed outcomes has only one member, {[tak]}. Hence, the candidate outcomes form a proper superset of the observed outcomes, thus meeting the condition in (19b). MaxEnt models can never assign 0% probability to any outcome, while a perfect fit to the observed data must assign 0% probability to the [tag] SR outcome. The best that a MaxEnt model can do is to inch ever-closer to the observed '100% [tak], 0% [tag]' relative frequencies without ever matching these frequencies exactly. This behavior is indicative of an asymptote at the theoretical ceiling likelihood. Such an asymptote precludes a global maximum.

Let us now consider the Variation data set. The members of both the set of candidate outcomes and the set of observed outcomes are identical: {[tak], [tag]}. Hence, the condition in (19b) is not met because no proper superset relation holds between these two sets. Assuming that appropriate phonological parameters are used (*i.e.* there is at least one parameter that distinguishes between the two candidate outcomes), it is possible for a trained model to match the observed '70% [tak], 30% [tag]' relative frequencies exactly. The corresponding point in the solution curve would be a global maximum that has reached the theoretical ceiling likelihood.

To visualize the behaviors described above, we need some concrete phonological parameters to work with (Table 6.2): In Figure 6.3, the blue curve shows the probability of surface [tak] given

[17] I discuss over-parameterization in §6.5.3 and §6.6.

/tag/	*D	FAITH
tak		1
tag	1	

Table 6.2: *Phonological parameters for the two SR outcomes.*

the UR /tag/ as a function of the difference in parameter weight between *D and FAITH. The blue

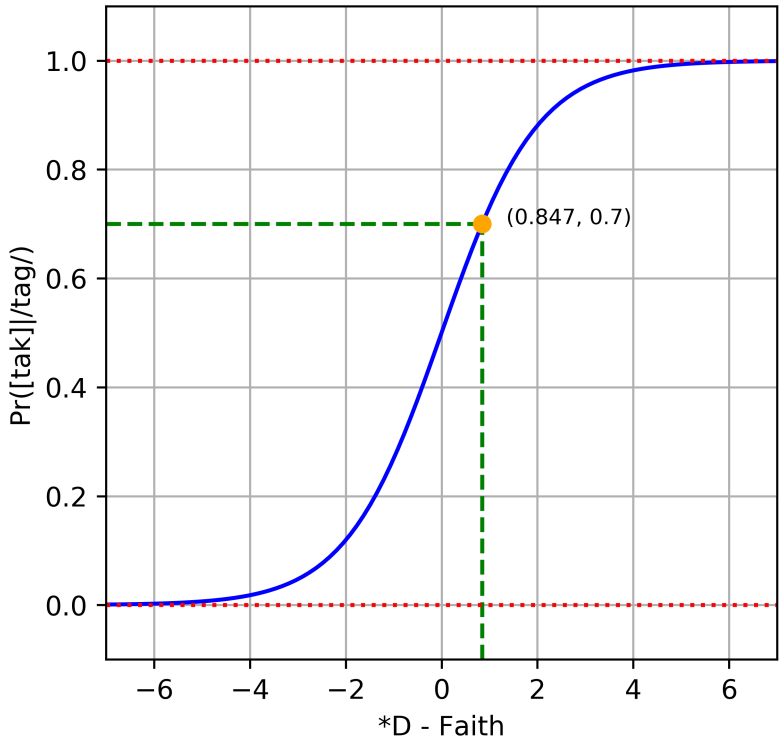

Figure 6.3: *Probability of surface [tak] given the UR /tag/ as a function of *D − FAITH.*

curve thus represents all possible grammars, and shows the distribution over [tak] and [tag] that correspond to each possible grammar.

Recall that the Variation data set has the following observed relative frequencies: 70% [tak] and 30% [tag]. This distribution over SRs is represented by the horizontal green line. The orange dot identifies the point at which this green line intersects the blue curve. Reading off the graph, a weight difference of 0.847 results in learning the Variation data set perfectly. In other words, *D needs to be exactly 0.847 heavier than FAITH to convert exactly 70% of /tag/ to [tak].

The Categorical data set has the following observed relative frequencies: 100% [tak] and 0% [tag]. This distribution over SRs is represented by the top red line. At no point does the blue curve

ever intersect the top red line. Zooming in, we can see that the blue curve gets ever closer to the red line without ever touching it (Figure 6.4). The red line is an asymptote because while the blue line

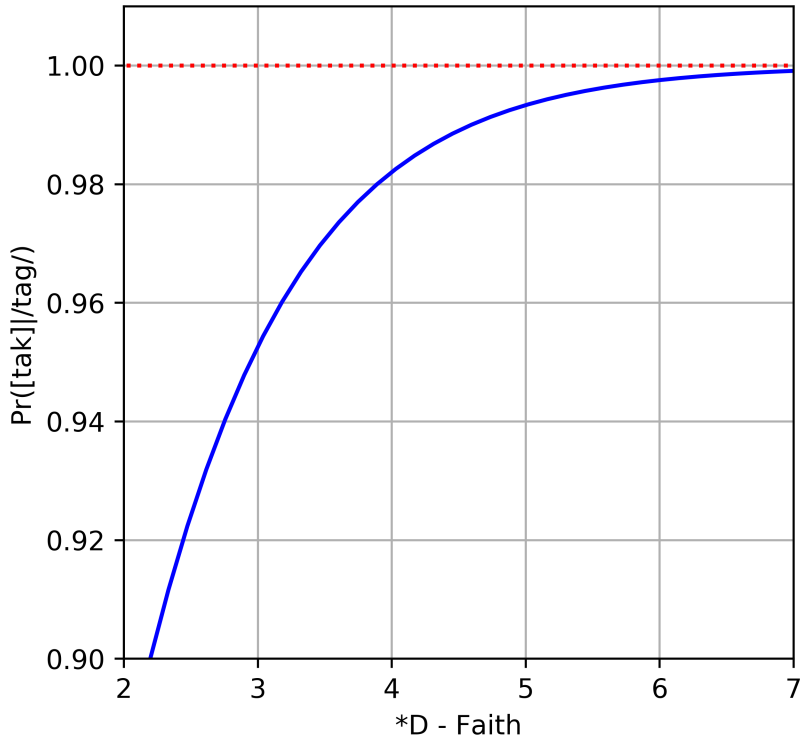

Figure 6.4: *The red line is an asymptote – the blue curve approaches but never reaches the red line.*

approaches it, the blue line never actually reaches it. Phonologically, this means that as the weight of *D grows ever larger and further from FAITH, a higher percentage of /tag/ gets converted to [tak], so the distribution over SRs gets ever closer to that of the Categorical data set. However, no weights for *D and FAITH will ever be able convert 100% of /tag/ to [tak]. Hence, no global maximum exists for the Categorical data set because there are no weight setting for *D and FAITH which can produce the Categorical data set exactly.

In Figure 6.5, a modified version of the solution curve is shown in blue. Instead of showing the model fit to the training data as a function of the individual parameters, it is shown as a function of the difference in the parameter weights: *D $-$ FAITH. The log-likelihood, which is a measure of model fit to data, is on the y-axis. Higher values of log-likelihood indicate a better model fit. At a difference of 0.847, the Variation data set reaches peak model fit. As we move away from 0.847

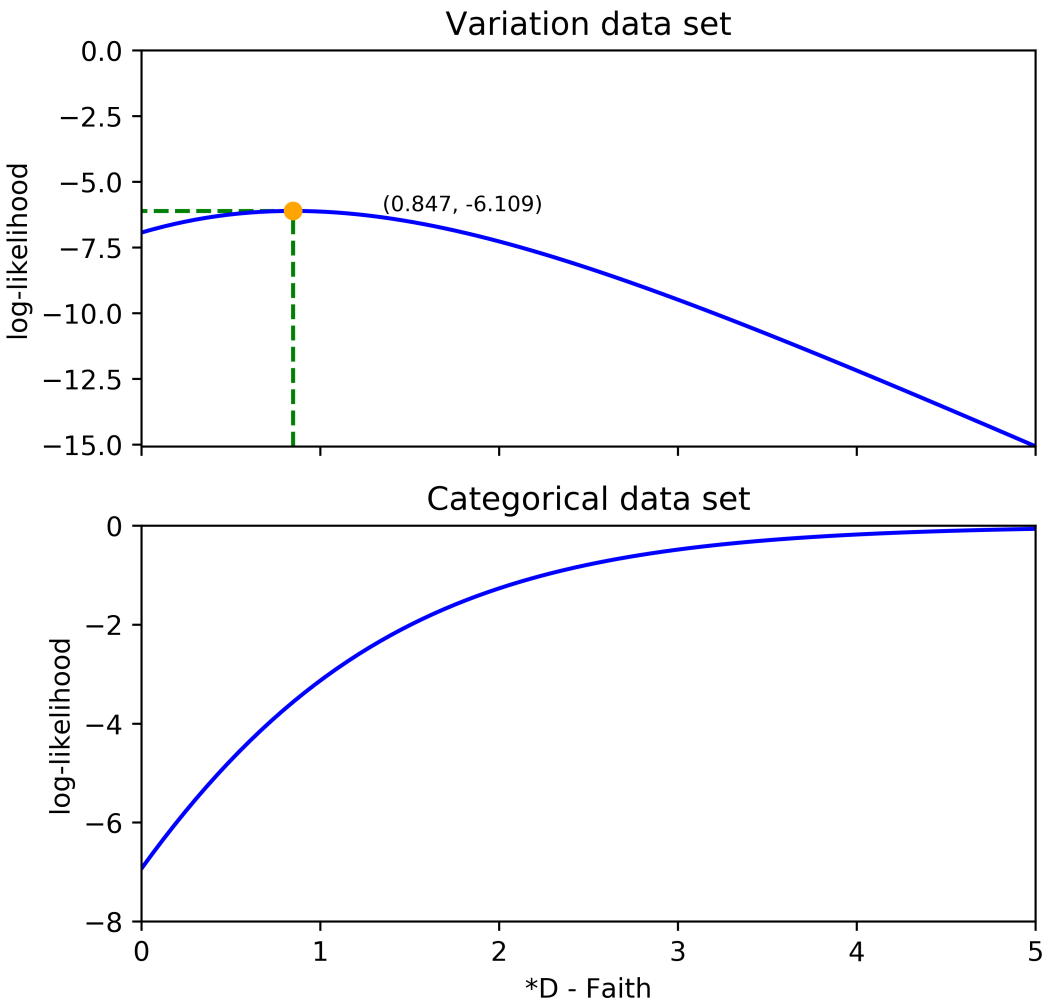

Figure 6.5: *The Variation data set has a global maximum. The Categorical data set has no global maximum.*

whether by increasing or decreasing difference between *D and FAITH, the model fit to the data gets worse. For the Categorical data set, however, there is no point at which the peak model fit is achieved.

The main takeaway is that when categorical phonological data is modeled with (an over-parameterized) MaxEnt model, the inclusion of unobserved outcomes in the set of candidate outcomes leads to a solution curve that has an undefined global maximum. In place of the global maximum, we get an asymptote. When the set of candidate outcomes includes only the observed outcomes, the asymptotic issue disappears, and the solution curve may once again have global

maxima[18].

6.4.4 Optimization methods for undefined global maxima in categorical phonology

In §6.4.1, I have argued that the points in the solution curve that are found via an iterative method are effectively the same regardless of whether the global maximum is defined or undefined. This is because the iterative method is informed by the gradient(s) at the current (and the previous) point(s).

An alternative to the iterative method is the analytical method. The analytical method finds the global maximum directly by setting the gradient to 0 and solving the ensuing equations. This direct method of identifying the global maximum is incompatible with my models because there are no points in the solution curve of my models at which the partial derivatives of the solution curve with respect to each parameter are simultaneously 0.

When fitting a phonological model to data, we want to find the global maximum, which represents the best-fitting model. Various optimization methods can be utilized to help us find the global maximum. The goal of the various optimization methods is to find a maximum (*i.e. the end-point*). The iterative method makes use of current (and sometimes past) information to guide its path to the end-point. The analytical method uses information at the end-point to arrive directly at the end-point. When the solution curve is known to not contain a global maximum, the end-point itself does not exist, so the analytical method should not be used. If the analytical method produces a result when global maxima are known to be undefined, then the point that it landed on is likely to be a local maximum, a saddle point, or a minimum. In other words, these are points in the solution curve that have a gradient of 0 but are not the global maximum. In contrast, the iterative method can produce good enough estimates of parameter weights. These estimates define a distribution that is very close to the one that would have existed at the global maximum.

[18] In the particular example used in this section, we actually get an infinite set of global maxima rather than a single global maximum. Visually, we have a ridge in the solution curve rather than a single peak. The ridge is likely the shape of many phonological MaxEnt models (when regularization is absent). I expand on this point in §6.5.2.

6.5 Regularization and the single global maximum

According to Goldwater and Johnson (2003), for MaxEnt models, "it is possible to show that the log conditional likelihood is concave, so there is only one global maximum (Berger et al., 1996)". Yet, in the preceding sections of this chapter, I have shown that (over-parameterized) MaxEnt models of categorical phonological data can have asymptotic curves (*i.e.* no global maximum) rather than having exactly one global maximum.

In this section, we will look at two schematic unregularized MaxEnt models whose solution curves do not have a lone global maximum. The first of these models has a solution curve with a ridge (*i.e.* an infinite number of global maxima). The second one has an asymptotic solution curve (*i.e.* no global maximum). We will then see that the use of a Gaussian prior as regularization restores both of these toy cases to having exactly one global maximum in their solution curves.

I then propose a refinement to Goldwater and Johnson's generalization that ties concavity with 'the relationship between local and global maxima'. Namely: since the log conditional likelihood of a MaxEnt model is concave, any local maximum is also a global maximum. This refined statement leaves room for the solution curves that have infinite number of global maxima as well as for those that have no global maximum.

6.5.1 Regularized Maximum Entropy models

In Goldwater and Johnson (2003), the objective function that they optimize consists of two components:

$$J(\theta) = \ell(\theta|D) - \sum_{i=1}^{m} \frac{(\theta_i - \mu_i)^2}{2\sigma_i^2}, \tag{6.1}$$

where θ_i represents the i^{th} parameter's weight and D is the training data. The first component, $\ell(\theta|D)$, is the log conditional likelihood, a measure of model fit. Increases in this first term correspond to improvements in fit to the training data. When this first term has been maximized, we get the best possible model fit to the training data. The second term goes by many names: regularization term, smoothing term, prior, bias term. Its purpose is to penalize deviation from an ideal weight. The value of the ideal weight for each constraint, μ_i, is usually set by the analyst. When

the second term is maximized, the learned weights do not differ from the previously set ideal value. When both of these terms are present in the objective function, the learner has to balance between two goals: fitting the model to the training data versus keeping the parameter weights at their previously-determined ideal value. The relative importance of these two goals is determined by the value of σ_i. σ_i is a value that the analyst can set for each parameter to indicate the importance of sticking closely to its previously-set ideal weight. A low σ_i indicates that it is very important to keep the i^{th} parameter's weight very close to its previously-determined ideal value, and vice versa. When the σ_i value is very low for all parameters, the learner will stick to the previously-set weights at the expense of fitting the data. In the opposite case, where σ_i is very high for all parameters, the learner will fit the data as well as it possibly can while giving little heed to the previously-set ideal weights. It is up to the analyst to choose appropriate values of σ_i so that the learner is able to satisfactorily fit the model to the data while allowing the previously-set ideal weights to have some say in the learned weights. For Goldwater and Johnson, the first term is implemented as the natural logarithm of the pseudo[19] likelihood. The second term is implemented as a Gaussian prior.

In machine learning, a regularization term is often included in the objective function in order to avoiding over-fitting the training data. The peril of over-fitting to the training data is the lessened ability of the trained model to generalize to novel data. For example, an over-fitted model may pick up on some accidental patterns that occur only in the training data. These accidental patterns are really only an artefact of how the complete data set was randomly split into the training and the generalization data sets, and are not present when the complete data set is looked at as a whole. Hence, learning these spurious accidental patterns results in doing a poorer job with the generalization data set.

6.5.2 Reasons for excluding the regularization term

As discussed in the preceding section, one of the reasons to include a regularization term is to avoid over-fitting to the training data. For phonologists, the same term can be used to model

[19]Goldwater and Johnson use the term 'pseudo' because the probabilities that phonologists use are conditional probabilities (*i.e.* $Pr(\text{SR}|\text{UR})$) rather than straight probabilities (*e.g.* $Pr(\text{SR})$ or $Pr(\text{UR, SR})$).

substantive biases (Wilson (2006), White (2017), Mayer (2021)) and to model a bias towards retaining the values of previously-learned weights over the time-course of language acquisition (O'Hara (2017)).

In the models that I presented in this dissertation, I did not utilize any such regularization or bias terms. Following the practice of including regularization within the model, I did originally experiment with incorporating a regularization term. However, I ultimately decided against it because trained models are easier to interpret if regularization is not in there exerting its influence on weights.

Additionally, for the phenomena that I modeled, there was no phonology-specific reason to use a bias term.

Finally, in the case studies that I modeled in this dissertation, over-fitting to the training data was not a concern for three reasons.

First, the training data that I exposed the learner to were carefully curated to include only representative alternations. Crucially, the training task was not to model phonotactics and the training data were not large corpora. Had the training task been to model phonotactics from a large corpus, a regularization term would be necessary to protect against over-fitting. In such a scenario, a regularization term that made high weights costlier would guard against fitting accidental patterns in the data.

The second reason that over-fitting was not an issue was because the negative impact of models that are over-fitted to their training data was not a concern for me. Over-fitting to the training data is generally seen as undesirable because fitting the training data too well often compromises a trained model's ability to generalize to novel data. In this dissertation, however, I am interested in the generalization capabilities of a random sample of trained models rather than the generalization capability of a single trained model.

In my approach, I take that the surface pattern (*i.e.* WORD-SR frequencies) can be perfectly learned because humans are able to produce the correct surface forms for words that already exist in their language. Since the intermediate morphological structure and URs are unobserved, multiple paths are available for the WORD-SR mapping to proceed along. Crucially, the ability of a

trained model to generalize to novel data turns on the unobserved intermediate structures that it has learned. The ability to generalize does not critically depend on the model fit to the training data.

For example, of all the trained models that matched the Velar Softening training data perfectly, only 77.8% generalized the /k/ $\rightarrow$ [s] process to novel data. These 77.8% of models were found to have learned one type of hidden structure, which the remaining 22.2% did not learn.

This has the nice interpretation that while all children learn the observed surface forms in their language, the unobserved intermediate structures that they learn may differ – some children may learn unobserved intermediate structures consistent with generalizing the phonological pattern to novel data, yet others may learn unobserved intermediate structures that do not allow them to generalize the phonological pattern to novel data. In the former case, building the intermediate UR from /ɪlɛktɪɪk/ and /-ɪti/ means that the child must learn the /k/ $\rightarrow$ [s] process. In the latter case, storing the intermediate UR /ɪlɛktɪɪsɪti/ in the lexicon leaves the child with no need to learn the /k/ $\rightarrow$ [s] process. The former child is predicted to generalize the /k/ $\rightarrow$ [s] process to novel words while the latter will not. Yet, these two different paths via different intermediary URs produce the same correct surface form [ɪlɛktɪɪsɪɾi].

Finally, the goal of modeling was not to get a very high generalization rate to novel data. Instead, the goal was to achieve a generalization rate close to that observed in humans. My models' 77.8% generalization rate was very close to the human rate of 80%.

6.5.3 Introducing regularization changes the shape of the solution curve

In §6.4.3, I discussed the properties of MaxEnt models without regularization. In particular, regularization-less (over-parameterized) MaxEnt models of categorical phonology will have an asymptote in place of a global maximum. For MaxEnt models of phonological variation, whether the solution curve of the regularization-less MaxEnt model has global maxima or not depends on the analyst's choice of candidate outcomes. If all the candidate outcomes are observed at least once, then there will be at least one global maximum in the solution curve.

Rather interestingly, the regularization-less MaxEnt model of the Variation data set in §6.4.3 has an infinite set of global maxima, rather than a single global maximum. Visually, we have a ridge

in the solution space rather than a single peak. The ridge is likely the shape of many phonological MaxEnt models (when regularization is absent). This is because phonological models tend to be over-parameterized. In this particular case, we have two parameters, *D and FAITH, when mathematically, one parameter would have sufficed. The relevant parameter is '*D − FAITH'. In other words, it is the difference in the weights of these two parameters rather than their individual weights that matter. A weight difference of 0.847 produces the perfect fit to the data. Hence, every weight setting of *D and FAITH that produces a difference of 0.847 is a global maximum.

Figure 6.6 provides an illustration of the infinite number of global maxima. The blue line

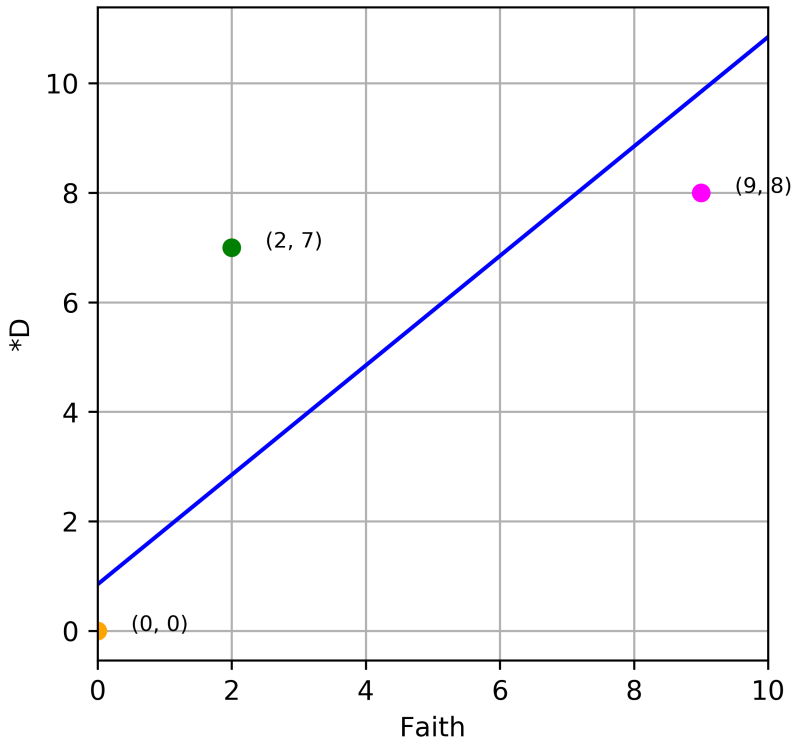

Figure 6.6: *The Variation data set has a global maximum. The Categorical data set has no global maximum.*

represents every setting of weights where *D is exactly 0.847 heavier than FAITH. The blue line is the top of the ridge, and is uniformly tall along its length. The regions to the left and right of the blue line slope downwards from the ridge. The blue line continues forever to the right[20]. This illustrates that there are an infinite number of weight settings where *D is exactly 0.847 heavier than FAITH. Consequently, that there are an infinite number of global maxima because each point

on the blue line is an equally good solution to fitting the training data exactly.

If a regularization or bias term is used to nudge weights to a previously-set ideal, then a peak (*i.e.* global maximum) emerges. For example, a regularization term may be introduced to keep all parameter weights as close to 0 as possible. In Figure 6.6, this regularization term is represented as the orange dot[21]. Its coordinates, $(0, 0)$, indicate that the previously-determined ideal weight for both FAITH and *D is 0. The blue line still represents the weights needed to fit the training data exactly. The new peak will appear somewhere between the blue line and the orange dot, with its exact location depending on the relative strength of the regularization term with respect to the importance of fitting the model to the data. A strong regularization term would have the new peak emerge closer to the orange dot while a weak regularization term would leave the new peak closer to the blue line.

In Figure 6.7, we have an illustration of how the solution space tranforms from a ridge (consisting of an infinite number of global maxima) to having just one peak once a regularization term is introduced. The highest points of the ridge are projected above the red diagonal lines in the bottom surface of the left subplot. Upon regularization, the red lines transform into an ellipsis[22]. The regularization term[23] that I used corresponds to the orange dot in Figure 6.6. As expected, the peak (approximately $(0, 0.606)$) occurs between the blue line and the previously-set ideal weights, $(0, 0)$.

---

[20]Phonological constraint weights cannot take negative values, so the blue line can only continue towards the right. Continuing towards the left would result in negative-valued weights.

[21]The green and pink dots represent other possible bias terms that a phonologist might use. For example, a phonologist might believe that FAITH should have some low but non-0 weight and that there is an innate bias for markedness constraints like *D to have higher weights than faithfulness constraints. This phonologist might choose to set the ideal weights for FAITH at 2 and for *D at 7. The green dot, $(2, 7)$, represents this bias term. Another phonologist might be interested in the time-course of learning (*cf.* O'Hara (2017)), and wish to keep constraint weights close to the weights that were learned in a previous stage of language acquisition. For example, if the weights learned at a prior stage of language acquisition were 9 for FAITH and 8 for *D, then the pink dot, whose coordinates are $(9, 8)$, represents this prior term.

[22]The ellipsis is cut off due to the lower bound on phonological parameter weights because negative weights are not allowed.

[23]The ideal weight, μ, for both *D and FAITH was set to 0. The value of σ for both of these parameters was set to 1.

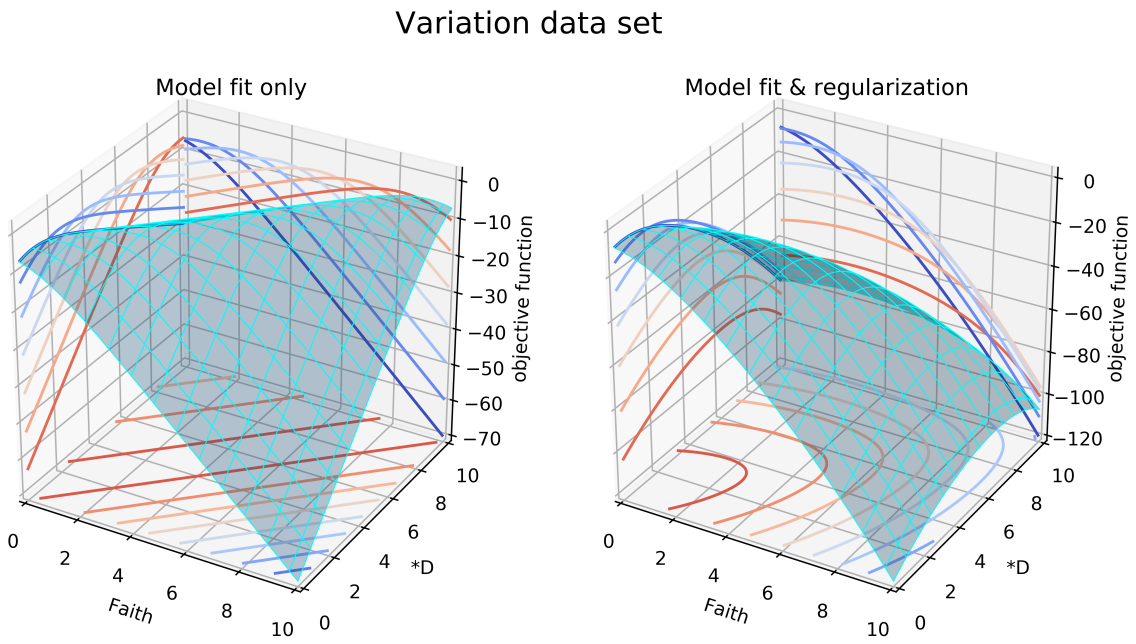

Figure 6.7: *Regularization gives the solution space a single global maximum (Variation data set).*

We have previously seen that regularization-less over-parameterized[24] MaxEnt models of categorical phonology will have an asymptote in place of a global maximum. This is illutrated in the left subplot of Figure 6.8. Here, we see an asymptote – the objective function (which is equivalent to the log-likelihood in regularization-less models) creeps every closer to the elusive value of 0 as the weight of *D increases (left wall of the left subplot in Figure 6.8). When a regularization term[25] is introduced, a peak emerges. The global maximum occurs when the weight of FAITH is approximately 0 and the weight of *D is approximately 1.62.

Just to be extra certain that we are indeed looking at a global maximum, I extend the lower bounds so that the peak can be seen in full (Figure 6.9). The global maximum now occurs when the weight of FAITH is approximately -1.03 and the weight of *D is approximately 1.03. There are two things to note. First, the shape of the solution space has not changed at all. Rather, the bounds that are set on the possible weights for the phonological parameters (*i.e.* only non-negative

[24]In fact, the asymptotic toy example has the same over-parameterization issue as the toy model for the Variation data set. The only difference between these two cases is that the asymptotic one also has unobserved SR outcomes.

[25]As before, the ideal weight, μ, for both *D and FAITH was set to 0. The value of σ for both of these parameters was set to 1.

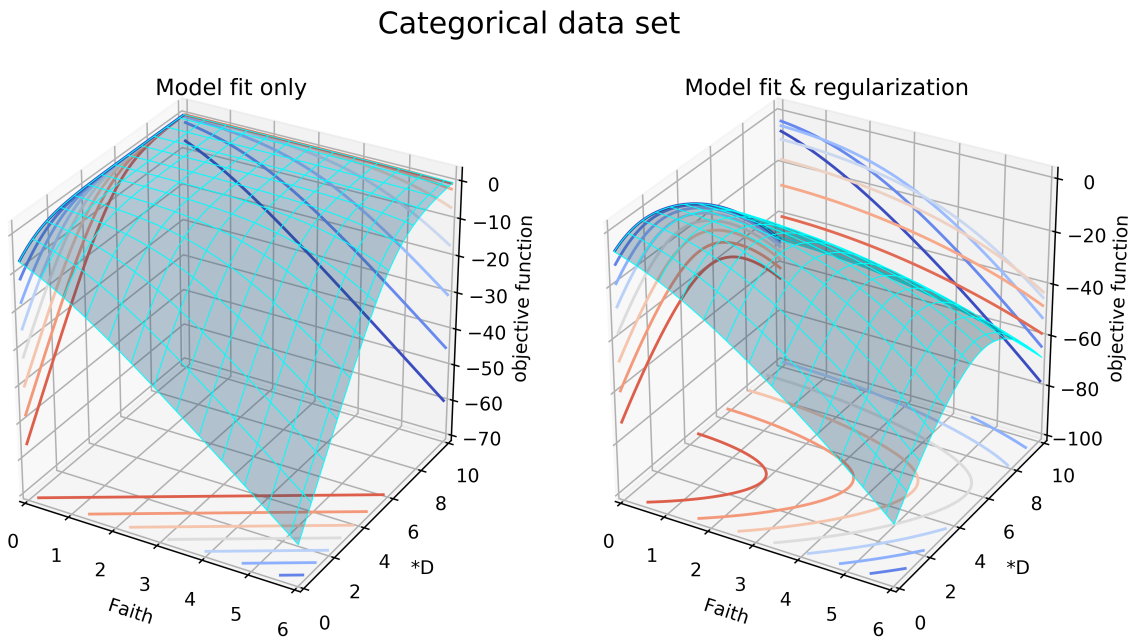

Figure 6.8: *Regularization gives the solution space a single global maximum (Categorical data set).*

values are permissible) means that this set of better weights are inaccessible. Second, I have used a very common strategy for setting the ideal weight, $\mu = 0$, and the cost of deviating from the ideal weight, $\sigma = 1$. In phonology, it is very common to set the value of μ for all parameters to 0. It is also very common to have the same σ value for all parameters. By extending the bounds into the negative-value territory, we now appreciate that it is preferable to equally split the burden of deviating from the ideal weight across multiple parameters. For example, both parameter weights have the same magnitude (1.03), but their signs are in the opposite direction. In phonology, we often observe that the weight of the disfavored parameter is 0 while the favored parameter has some positive weight. This is observed even when the μ's and σ's are the same for all phonological parameters. This unequally-split burden is a result of the lower-bound of permissible parameter weights being 0, rather than any other type of pressure to unevenly split the burden (of deviating from the ideal weight) between parameters.

How does a regularization term cause the solution curve for the over-parameterized categorical model to go from having no global maximum to having a single global maximum? In the former case, no global maximum exists because the model fit can always be improved by increasing the difference in weight between *D and FAITH. While FAITH gets stuck once it reaches the

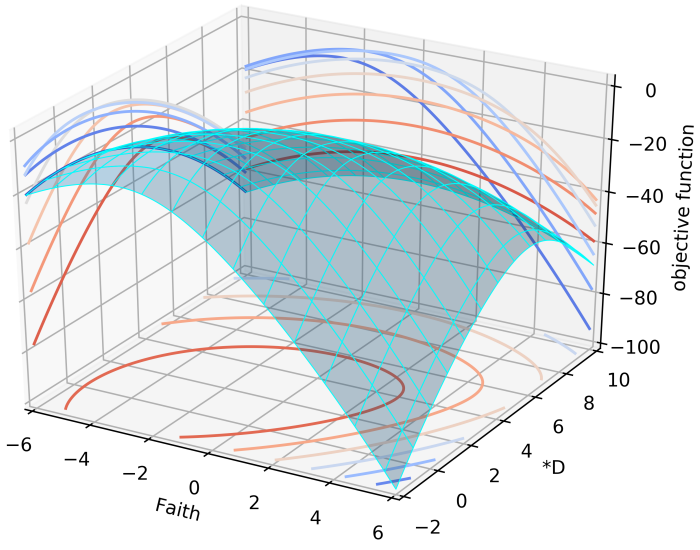

Figure 6.9: *Regularization gives the solution space a single global maximum (Categorical data set).*

lower-bound (*i.e.* 0), *D can increase its value indefinitely. Once a regularization term is in the mix, weights that differ from the previously-determined ideal weight, μ, are penalized, with larger deviations being penalized more harshly. It is thus no longer free to increase the weight of *D indefinitely. For example, as the weight of *D increases from 0, the model fit to data keeps improving, albeit at a decreasing rate. However, once the weight of *D starts deviating too much from its previously-determined ideal weight, μ, it becomes more costly to increase the weight of *D further. The peak exists because it is beneficial to increase the weight of *D up to a certain point, after which further increases in weight become more and more detrimental.

6.5.4 Interim summary

This section opened with a mystery. Goldwater and Johnson assert that MaxEnt models have a single global maximum. Despite Goldwater and Johnson's claim, I find that the solution curve of my Rich Base models do not have a single global maximum. There are two ways to not have a single global maximum – having multiple global maxima or having no global maxima. In my case, the solution curve of my models are asymptotic, which leaves the global maximum undefined (*i.e.* a solution curve without a global maximum).

One key difference between my MaxEnt model and Goldwater and Johnson's is that the former uses plain and simple MaxEnt while the latter uses a regularized MaxEnt model. I set out to explore whether regularization was the factor that guarantees a solution curve with a single global maximum. I looked at two schematic unregularized MaxEnt models whose solution curves did not have a lone global maximum. The first one had a ridge (*i.e.* an infinite number of global maxima). The second one had an asymptotic solution curve (*i.e.* no global maximum). The regularization term I used was the Gaussian prior (second term in Eq 6.1) with μ for all parameters set to 0 and σ for all parameters set to 1. In these two cases, regularization was indeed able to transform a solution curve from one without a lone global maximum into one that had a single global maximum. These two cases demonstrated that regularization was indeed able to transform a solution curve from one without a lone global maximum into one that had a single global maximum.

6.5.5 Refining Goldwater & Johnson (2003)'s generalization

The Goldwater and Johnson quote is repeated for convenience: "This discussion ignores the possibility of multiple local maxima. In fact it is possible to show that the log conditional likelihood is concave, so there is only one global maximum (Berger et al., 1996)". There are two arguments Goldwater and Johnson make with this statement. The first argument asserts that the log conditional likelihood is concave. The second argument asserts that concavity leads to a single global maximum. The two cases we have looked at have concave log conditional likelihoods. This implies that the second argument cannot be accurate since these two cases do not have one global maximum (they either have none or an infinite number of global maxima). I propose a refinement to Goldwater and Johnson's statement. In particular, I tie concavity to the relationship between local and global maxima. I then tie regularization to its ability to reduce the number of global maxima (ideally to one).

Both the solution curve with the ridge and the solution curve with the asymptote meet the definition of concavity. In addition, the one with the asymptote meets the definition of strict concavity. In other words, these are log conditional likelihood functions that are concave; thus, aligning with the Goldwater and Johnson's first argument. But, these concave log conditional likelihood func-

tions do not have a one global maximum in the solution curve; thus, serving as counter-examples to Goldwater and Johnson's second argument.

In order to reconcile this contradiction, I propose a refinement that links the concept of 'concavity' with 'the relationship between the local maxima and the global maxima'. My refined generalization: **Since the log conditional likelihood of a MaxEnt model is concave, any local maximum is also a global maximum**.

My refined generalization makes no claim about the number of global maxima. This leaves open the possibility of there being multiple global maxima (*e.g.* the solution curve with the ridge) or no global maxima (*e.g.* the asymptotic solution curve). Thus, my refined generalization is consistent with the two cases that we have previously seen, whose log conditional likelihoods were concave, yet did not have one global maximum.

Rather than being concerned with the number of global maxima, my refined generalization claims that the set of global maxima and the set of local maxima are one and the same. The implication of this is that once a maximum has been found, we can be certain that it is a global maximum. The great advantage of this is that there are no "tricky" points in the solution curve that can trap the learner and keep it from finding the global maximum.

There are a couple of caveats to be aware of (especially in the practice of optimization). First, while we can be certain that any maximum found is a global maximum, we cannot be certain that the maximum we are currently looking at is the only global maximum. This first scenario corresponds to the "ridge" solution curve. Second, an iterative method of optimization usually returns a solution that is in the vicinity of a stationary point (very seldom landing right on the stationary point itself). Thus, when using an iterative method, we can only be certain that the found solution is close to the global maximum. (Since the function we are talking about is the log conditional likelihood, the found solution is extremely close to producing the best fit to the data set.) We, however, cannot conclude that the presence of a found solution means that there is at least one global maximum. For example, in the "asymptotic" solution curve, an iterative method will be able to find solutions that produce very good fits to the data. This despite there being no global maximum in the "asymptotic" solution curve.

There is one loose end to tie up regarding the role that regularization plays. Notice that Goldwater and Johnson link the term 'log conditional likelihood' rather than the term 'objective function' (which for them includes the regularization term) (Equation 6.1) with the concept of concavity. They then link concavity to having one global maximum. This implies that Goldwater and Johnson do not attribute having one global maximum to regularization (since the regularization term is a component of only the objective function but not a component of the log conditional likelihood).

Nevertheless, we have seen with Figure 6.7 and Figure 6.8 that the introduction of regularization can result in a solution curve having one global maximum, when the solution curve previously did not have one global maximum. Indeed, according to Wilson (2022), "[r]egularization is a general method for reducing the set of possible estimates (ideally to one) for a nonidentifiable model".

Thus, Goldwater and Johnson may yet be correct in their assertion that their formulation of the regularized Maximum Entropy model has only one global maximum. Nevertheless, they mistakenly attribute this property to concavity. Instead, the reduction of possible estimates to fewer peaks (ideally to a single peak) is properly attributed to regularization.

6.6 Over-parameterization

In the preceding section (§6.5), we explored the relationship between regularization and the number of global maxima. We saw that regularization can be used to reduce the number of global maxima (ideally to one).

In this section, I explore the roots of over-parameterization in MaxEnt as we move from the world of strict rankings in OT to the MaxEnt world of weighted constraints (§6.6.1). I then consider over-parameterization from two angles. The first angle compares the number of parameters to the number of candidates (§6.6.2). The second angle looks at over-parameterization from the concept of identifiability (§6.6.3). In §6.6.5, we see that regularization cannot undo this flavor of over-parameterization to give us an identifiable model. In fact, rather than bringing us closer to the 'one true value' of each parameter, regularization further obscures the true values of the parameters.

Finally a note on terminology: since I will be discussing classical Optimality Theory (OT), I

will be using the terms (phonological) constraint and phonological parameter interchangeably.

6.6.1 From OT to Maximum Entropy: The roots of over-parameterization

Optimality Theory was developed to formalize the idea that the grammar (*i.e.* the UR-to-SR mapping) is "a system of conflicting forces" (Kager, 1999). These conflicting forces can generally be classified into two major classes – the MARKEDNESS and FAITHFULNESS constraints. Generally speaking, markedness constraints encode pressures that are external to the grammatical system. For example, the markedness constraint *D bans voiced obstruents. This markedness constraint is grounded in articulation since it is difficult to have air flowing through the larynx in a steady enough stream to keep the vocal folds vibrating (*i.e.* voicing) when there is a severe constriction in the downstream oral cavity[26] (*i.e.* an obstruent). Markedness is opposed by faithfulness, the latter of which works to preserve lexical contrasts. For example, voicing is contrastive in English, as evidence by the minimal pair: *rip* [ɹɪp] and *rib* [ɹɪb]. In English, the faithfulness constraint IDENT-IO (voice) works to preserve the underlying voicing contrast at the surface.

Each conflicting force is encoded by a constraint. Each meaningful constraint conflict is represented by a CONSTRAINT RANKING. The most important thing to understand about Optimality Theory is that Optimality Theory is a theory of constraint interactions. It is not a theory about constraints.

In Bulgarian, word-final voiced obstruents are devoiced. For example, the word *grad* 'city' ends in a voiceless obstruent [grat] (*cf.* *gradove* [gradove] 'cities'). The markedness constraint *D# bans word-final voiced obstruents while the faithfulness constraint IDENT-IO (voice) works to keep each sound's surface voicing value the same as that of its underlying correspondent's. The constraint ranking that describes word-final devoicing in Bulgarian is: *D# $\gg$ IDENT-IO (voice). This ranking states that it is more important to avoid a voiced word-final obstruent than it is to retain a sound's underlying voicing value. Hence, the bad word-final voiced obstruent /d/ is repaired to an acceptable [t] by changing its voicing value (Table 6.3).

[26]Nasal consonants also have a severe constriction (in fact, a complete occlusion) in the oral cavity. But they have an open side branch (in the form of the open pharyngeal port that allows air to flow through the nasal cavity instead), so there is no articulatory pressure against voiced nasal consonants.

/grad/	*D#	IDENT-IO (VOICE)
☞ a. grat		*
b. grad	*!	

Table 6.3: *OT tableau for word-final devoicing in Bulgarian.*

While classical Optimality Theory shines in the modeling of categorical data, it cannot handle data sets with free variation (Goldwater and Johnson, 2003). Hence, Goldwater and Johnson propose their Maximum Entropy model. Of the various similarities and differences between OT and MaxEnt, two are of particular importance to our discussion. First, MaxEnt retains the constraint conflict character of OT. Second, MaxEnt replaces OT's constraint rankings with constraint weights.

To give MaxEnt its chance to shine with free variation data, I create the pseudo-Bulgarian data (Table 6.4). The MaxEnt tableau for pseudo-Bulgarian (Table 6.5 is identical to that of actual

Data set	*Data type*	[grat]	[grad]
Bulgarian	*Categorical*	10	0
pseudo-Bulgarian	*Free variation*	7	3

Table 6.4: *Word-final devoicing data sets.*

Bulgarian (Table 6.3), save the markings related to constraint rankings. It turns out that a (non-

/grad/	*D#	IDENT-IO (VOICE)
a. grat		*
b. grad	*	

Table 6.5: *MaxEnt tableau for word-final devoicing in pseudo-Bulgarian.*

regularized) MaxEnt model produces the best fit to the pseudo-Bulgarian free variation data when *D# is exactly $0.847 (= ln(\frac{7}{3}))$ heavier than IDENT-IO (voice)[27].

We can immediately see that the MaxEnt statement "*D# is exactly 0.847 heavier than IDENT-IO (voice)" is the analog of the OT constraint ranking "*D# $\gg$ IDENT-IO (voice)".

[27] In fact, the generalization here is that if we observe x occurrences of [grat] and y occurrences of [grad], then we know that the weight difference between *D# and IDENT-IO (voice) is $ln\frac{x}{y}$.

Yet, there is a sense that the MaxEnt model is over-parameterized while the OT one is not! In fact, the MaxEnt model's solution curve has a ridge (left plot of Figure 6.7, repeated in Figure 6.10). Each point along the top of the infinitely long ridge corresponds to pairs of constraint

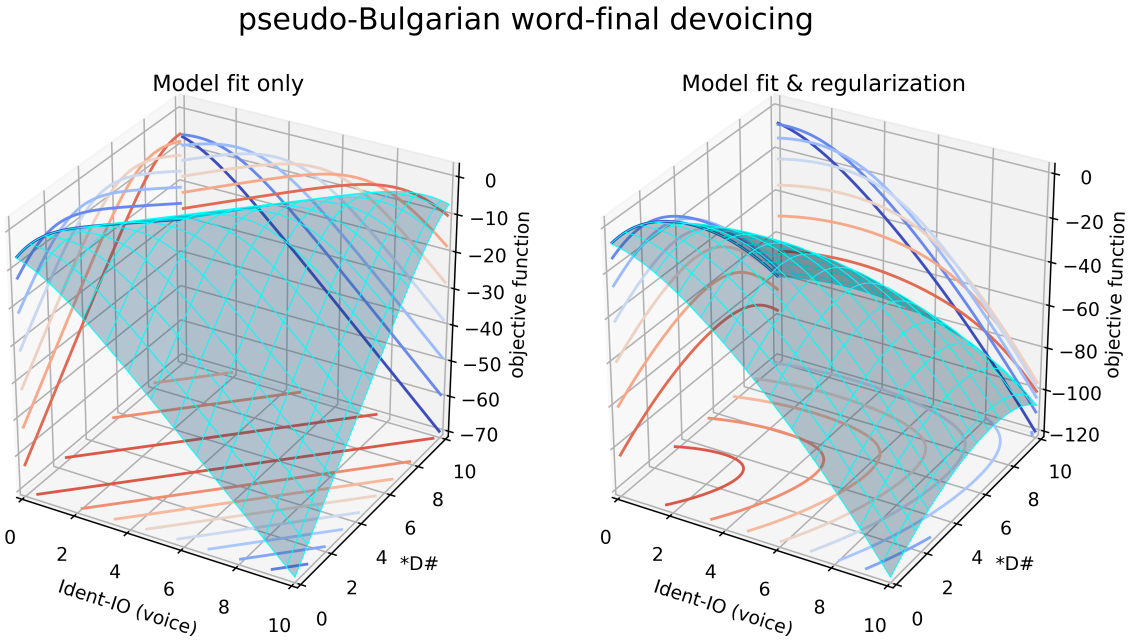

Figure 6.10: *MaxEnt model of pseudo-Bulgarian has a ridge.*

weights in which *D# is exactly 0.847 heavier than IDENT-IO (voice).

We can explore this sense of over-parameterization from two angles. First, we can compare the number of free parameters to the number of candidates. In general, the number of free parameters can be at most $n - 1$ for n candidates in order for the model to not be over-parameterized. Second, over-parameterized model often lack model identifiability.

6.6.2 Too many parameters to too few candidates

To see why the MaxEnt model is over-parameterized, we can compare the number of candidates with the number of free parameters.

As to the number of relevant data points, what really matters is not the actual number of observations itself. Rather, it is the number of distinct violation profiles of the SR candidates that matter. For both the pseudo-Bulgarian and Bulgarian data sets, the number of distinct violation

profiles is 2. In the pseudo-Bulgarian data, seven observations have one type of violation profile and the other three have the other type of violation profile. In other words, it is the candidate that is the relevant data point. So, the number of relevant data points for both pseudo-Bulgarian (MaxEnt) and Bulgarian (OT) is actually 2!

For the MaxEnt model, the relevant free parameter is the phonological constraint itself. In other words, each phonological constraint counts as one parameter. Since we have two phonological constraints in the model, the number of free parameters is 2.

In general, a model can have at most $n - 1$ free parameters for n data points in order to not be over-parameterized. The MaxEnt model has 2 free parameters for 2 relevant data points (*i.e.* candidates), so it is over-parameterized.

6.6.2.1 A short detour on parameterization in OT

As an aside, we can consider what the relevant free parameter in OT is.

One possible take would be to treat a pair of constraints as a free parameter. The free parameter can be set by ranking the first member of the pair above the second member. Or it can be set in the opposite direction. Since there are only two phonological constraints in the model for Bulgarian word-final devoicing, there is only one pair of constraints. So, the number of free parameters in the OT model is 1.

Under this view of parameterization in OT, the OT model has 1 free parameter for 2 relevant observations, so it is not over-parameterized.

One interesting thing to notice is that the number of constraint pairs in OT technically grows faster than the number of constraints in MaxEnt (Table 6.6). So will an OT model suffer from greater ill-effects of over-parameterization as model complexity increases?

First, the number of free parameters in OT is generally smaller than the number of constraint pairs. An OT grammar is really a partial order on the set of phonological constraints. The relevant binary relation is the ranking relation. A partial order has to fulfill the criterion of transitivity: If Constraint A $\gg$ Constraint B and Constraint B $\gg$ Constraint C, then Constraint A $\gg$ Constraint C. In other words, due to transitivity, it is not the case that the ranking relation between each member

Number of constraints	*Number of constraint pairs*
1	$C_2^1 = na$
2	$C_2^2 = 1$
3	$C_2^3 = 3$
4	$C_2^4 = 6$
5	$C_2^5 = 10$
⋮	⋮

Table 6.6: *The number of constraint pairs grows quickly.*

of a constraint pair can be freely set. So, while we know that the number of free parameters in OT is at most C_2^n, where n is the number of constraints, the number of free parameters is lower than this value in most instances.

Second, over-parameterization of the following kind is not considered by most phonologists to be a problem in OT. Let us imagine an OT model in which Constraint A $\gg$ Constraint C and Constraint B $\gg$ Constraint C produces the correct SRs. No ranking relation holds between Constraint A and Constraint B. This model is over-parameterized because there is one remaining free parameter (*i.e.* the ranking relation between Constraint A and Constraint B), that will not affect the predicted SR outcomes in any way. We could arbitrarily set this free parameter to Constraint A $\gg$ Constraint B or vice versa: the outcome over SRs would be the same.

6.6.3 Identifiability

According to Wilson (2022), a model is identifiable if there is a one-to-one relation between the vector of constraint weights, θ, and the probability distribution, $p_(x;\theta)$. With an identifiable model, it is possible to recover the true values of the model's parameters.

A MaxEnt grammar is a vector of constraint weights. The MaxEnt model of pseudo-Bulgarian word-final devoicing (Table 6.4) has two constraints, so the vector of constraint weights consists of two elements whose weights can be freely adjusted.

The MaxEnt model of pseudo-Bulgarian word-final devoicing (Table 6.4) is not an identifiable model. It is not possible to recover the true values of *D# and IDENT-IO (voice). Instead, what is recoverable is the true value of the difference between *D# and IDENT-IO (voice) (the value

is: $ln(\frac{7}{3}) = 0.847$). Rather than a one-to-one relation, we have a many-to-one relation because multiple vectors of constraint weights map to the same probability distribution. For example, every weight vector at the top of the ridge maps to the same probability distribution (*i.e.* 70% [grat], 30% [grad]).

6.6.3.1 A short detour on identifiability in OT

Identifiability is a concept that was developed for statistical models. While classical OT is not a statistical model, we can try to extend the concept of identifiability to OT as best we can.

An OT grammar is a partial order on constraints. This is more restrictive than saying that an OT grammar is a set of constraint rankings. A partial order means that the ranking relation is asymmetric. If Constraint A outranks Constraint B, then it must not also be the case that Constraint B outranks Constraint A. In other words, the model of Bulgarian word-final devoicing (Table 6.3) has only one boolean-valued[28] parameter. We can either have *D# $\gg$ IDENT-IO (voice) or IDENT-IO (voice) $\gg$ *D#, not both. One parameter setting (*i.e.* constraint ranking) precludes the other.

These two opposing parameter settings produce distinct distributions. *D# $\gg$ IDENT-IO (voice) gives us 10 [grat] and 0 [grad]. IDENT-IO (voice) $\gg$ *D# gives us 0 [grat] and 10 [grad]. In this sense, the OT model of Bulgarian word-final devoicing is identifiable. The true parameter setting that produces the data can be recovered. It is: *D# $\gg$ IDENT-IO (voice).

6.6.4 Interim summary

In OT, a parameter consists of a pair of constraints. In MaxEnt, each constraint counts as a separate parameter. Therefore, the conversion of a single OT parameter into its MaxEnt equivalent is one source of over-parameterization. A symptom of this particular flavor of over-parameterization is when it is the difference between a pair of constraint weights rather than a constraint's individual weight that truly matters.

This flavor of over-parameterization appears in MaxEnt because (1) MaxEnt has the worldview

[28] Or potentially ternary-valued if we consider remaining agnostic as a distinct third parameter setting.

in which the grammar is a system of conflicting forces, and (2) each of these conflicting forces (*i.e.* phonological constraints) is allowed to be a free parameter, whose weight can be adjusted freely by the learner. The current style of MaxEnt modeling is 'factorized' down to the level of each conflicting force. For example, each articulatory factor like the ban against voiced obstruents (*D) and the ban against consonant clusters (*COMPLEX) is a parameter whose weight can be freely adjusted.

A possible alternative to modeling at the level of each conflicting force is to model at the level of the 'conflict'. For example, the OT constraint ranking *D# $\gg$ IDENT-IO (voice) is one such conflict. Modeling at the level of the conflict means that we allow the MaxEnt learner to freely adjust the weight of the parameter '*D# $-$ IDENT-IO (voice)'. In such a model, the individual 'factors' *D# and IDENT-IO (voice) would not be parameters of the model. A large positive value for the sole '*D# $-$ IDENT-IO (voice)' parameter indicates that the conflict is going well for *D#, and it is converting a good majority of /grad/ to [grat]. A large negative value for this parameter indicates that the conflict is going well for IDENT-IO (voice), and it is retaining a good majority of /grad/ as [grad]. A value close to zero indicates that it is a rather even fight, and there are roughly equal numbers of [grat] and [grad].

Both levels of modeling have their advantages. The former 'factorized' model is closer to the view of the grammar as a system of conflicting forces, and probably closer to reality. The latter is a closer equivalent of the OT model and therefore less likely to be over-parameterized. Reading the results of the latter model, the analyst is also less likely to miss the point that what truly matters is the weight difference between *D# and IDENT-IO (voice)[29].

6.6.5 Regularization to the rescue?

In the preceding section, we have seen how moving from a world in which strict rankings are used to model the conflicting forces (OT) to a world of weighted constraints (MaxEnt) results a

[29]Although this begs the question on who the latter model is useful for. The analyst who comes up with the latter model has already done some work in order to determine that '*D# $-$ IDENT-IO (voice)' is *the* relevant parameter. In which case, this analyst is unlikely to miss the point that it is the weight difference between *D# and IDENT-IO (voice) that truly matters when they use the former model.

particular flavor of over-parameterization. One symptom of this flavor of over-parameterization is when it is the weight difference between constraints, rather than a constraint's individual weight that matters. This symptom implies non-identifiability: we are unable to recover the true weights of the constraints that produce the data. Instead, we have to settle for an infinite number of weight vectors that all produce equally good fits to the data.

In §6.5, we have seen that the introduction of regularization to the solution curve with the ridge results in a solution curve with a single global maximum. In addition, Wilson (2022) claims that non-identifiable models can be avoided by augmenting the model with regularization. Can regularization help us get an identifiable model? Can it help us recover the true values of the constraints that produce the data? Unfortunately, the answer for the particular case we are looking at is "no".

I follow Wilson (2022)'s definition of identifiability: a model is identifiable "if different parameter values necessarily specify different probability distributions". In other words, "the relation between θ and $p(x;\theta)$ is one-to-one".

Let us consider a specific counter-example. Let us say that after introducing regularization, instead of an infinite set of global maxima occurring where *D# is exactly $ln(\frac{7}{3}) = 0.847$ greater than IDENT-IO (voice), we find that a single peak emerges in the solution curve. The new optimal solution (there is now only one!) is when *D# is 0.606 and IDENT-IO (voice) is 0 (right plot of Figure 6.10). The weight difference between *D# and IDENT-IO (voice) is now 0.606. This corresponds to the following probabilities: 65% [grat], 35% [grad]. However, all constraint weight settings in which *D# is exactly 0.606 heavier than IDENT-IO (voice) will produce the same probability: 65% [grat], 35% [grad]. We still do not know the true value of *D# despite the learner misleadingly returning only one solution now. (*cf.* Previously, the learner could have returned multiple solutions all corresponding to *D# being 0.847 heavier than IDENT-IO (voice)). In other words, we have made the fit to the data worse without solving the identifiability problem.

In fact, the introduction of regularization has done nothing to change the relationship between the parameter values, θ, and the probability distribution, $p(x;\theta)$. The constraint weights '*D#: 0.847, IDENT-IO (voice): 0' still produce the distribution: 70% [grat], 30% [grad]. As do the

different constraint weights '*D#: 1.847, IDENT-IO (voice): 1'. This is the same probability distribution that these constraint weights produced before regularization. Rather than getting closer to finding the one true value of each constraint that would produce the data, all regularization has done is to further obscure these values.

In sum, identifiability is important because we can recover the true value of a constraint if we have an identifiable model. Unfortunately, having single global maximum is a different concept than identifiability. While we have seen a case in which regularization can help us achieve the former, it has not been able to help us achieve the latter.

6.7 An interpretation of models with multiple best solutions

Let us take stock of what we have seen so far. In the preceding sections, we have been concerned with the solution curve that has a ridge. We have seen that the type of over-parameterization that produces the ridge comes about because a single constraint ranking in OT is converted to two independent constraints in MaxEnt. In other words, a single Boolean-valued[30] OT parameter's equivalent in MaxEnt is two real-valued[31] parameters. Unfortunately, this flavor of over-parameterization produces a non-identifiable model. In the pseudo-Bulgarian toy example that produces the ridge, we are unable to recover the one true value of *D# and the one true value of IDENT-IO (voice) from the word-final devoicing data set. We have also seen that for this particular flavor of over-parameterization, regularization is unable to help us with non-identifiability (in fact, regularization further obscures the values of *D# and IDENT-IO (voice) that produce the data).

One way to approach the unrecoverability issue is to ask what we are actually doing when we train a single MaxEnt model. I suggest applying my approach of simulating a population of language learners to the identifiability problem. This results in the following interpretation: a single trained MaxEnt model represents a single speaker.

[30] Or ternary-valued if we allow the null ranking, where neither of the constraints of the pair are ranked with respect to the other.

[31] Non-negative real-valued to be precise.

When a single speaker of pseudo-Bulgarian learns their language, they are trying to find a single weight for *D# and and a single weight for IDENT-IO (voice) that reproduces the observed data. Any one weight setting that reproduces the observed data will do. When we train a single MaxEnt model, we are really trying to find a solution that best fits the training data – any one solution will do. We are not concerned with how many other solutions produce equally good fits.

In other words, when we look at the weights that a single trained model has learned, what we are really looking at is a single weight vector that is learned by just one speaker. From an individual speaker's point of view, all that they need to do is to find a single weight for *D# and and a single weight for IDENT-IO (voice) that reproduces the observed data: 70% [grat], 30% [grad]. One speaker might have learned the weights {*D#: 0.847, IDENT-IO (voice): 0}, yet another speaker may have learned the weights {*D#: 1.847, IDENT-IO (voice): 1}. It really is not an individual speaker's job to recover the one true value of *D# and the one true value of IDENT-IO (voice) that produces the observed data.

In fact, from this perspective, 'the one true value of *D# and the one true value of IDENT-IO (voice) that produce the observed data' may not even be meaningful. Let us consider where the observed data comes from. The observed data is produced by a community of pseudo-Bulgarian speakers. It is not the output of a single speaker. The observed data is not a monolithic object that is produced from a single source, from which we can recover the true values of this lone source's parameters. The observed data is the combined output of many individual speakers each running their own unique MaxEnt grammar.

An alternative MaxEnt model is the one that uses only one parameter '*D# $-$ IDENT-IO (voice)'. This alternative model is identifiable because the true value of its single parameter can be recovered from the observed data. The true value of this parameter is 0.847 – this is the single best solution. Nevertheless, while this alternative MaxEnt model is an excellent model of the observed data, it is not a very useful model of human speakers. Human speakers may under- or over-learn certain constraints owing to a constraint's naturalness, as in Hungarian backness harmony (Hayes *et al.* (2009)). Different constraints may have different favorite weights which encode substantive bias (Wilson (2006), White (2017)). Such phenomena are more naturally modeled when each constraint is a separate parameter whose weight can be freely adjusted. It is difficult to see how such

phenomena can be modeled when the parameter that is used is the MaxEnt analog of strict ranking (*i.e.* Constraint A − Constraint B).

But perhaps, there is still a lingering sense of dissatisfaction – why are there so many options available to an individual speaker when there is only one surface pattern? Here, I offer a consolatory viewpoint. Having many options available to an individual speaker can be valuable in phonological modeling. A case in point: my model is able to correctly predict inter-speaker variation in Velar Softening precisely because there are multiple equally good options available to an individual speaker.

With the pseudo-Bulgarian word-final devoicing model, it may feel intuitively "wrong" that so many equally good options are available. Yet, in the case of Velar Softening, the multiple equally good options turn out to be exactly what's needed. Different speakers engage different strategies to learn the same surface pattern. This is not obvious from looking at the surface pattern of existing words alone – in other words, the multiple equally good options look superfluous when the training task is viewed in isolation. Nevertheless, we have empirical data from a generalization task (the wug test) showing that some speakers employ a morphophonological approach to Velar Softening and learn the /k/ → [s] grammar while others just store the unanalyzed URs in their lexicon. With Velar Softening, we are lucky to find a case in which the multiple equally good options turn out to make predictions about inter-speaker variation that have been confirmed with empirical data. Unfortunately, we are less lucky with word-final devoicing: its multiple equally good options will never produce a prediction that can be confirmed or refuted with empirical data.

Having multiple equally good options is useful elsewhere in phonology. Hence, I forward that 'having multiple equally good options' alone is not a good reason to believe that the model of pseudo-Bulgarian with separate *D# and IDENT-IO (voice) parameters is flawed just because it has multiple equally good options to fit the observed data.

6.8 Summary of highlights

In this chapter, we have seen that my over-arching model is able to leverage the superior utility of Rich Base Grammars (over non-Rich Base Grammars), such that the learner always experiences

an incentive to learn the Rich Base Grammar no matter where it is in the solution curve. In other words, the overwhelming preference for Rich Base Grammars is emergent.

We have also looked at two scenarios that commonly arise in phonological modeling. First, we have explored how the conversion of one OT constraint ranking into two MaxEnt constraints is a source of over-parameterization. Second, phonologists often include unobserved SR candidates into their models, whether necessarily (when modeling categorical phenomena) or by choice (by including additional unobserved candidates when modeling data with variation).

I then presented two toy cases to illustrate these scenarios. The first case featured over-parameterization, and produced a solution space with an infinite number of global maxima (*i.e.* the ridge). The second case featured both over-parameterization and unobserved SR candidates, and produced a solution curve without any global maxima (*i.e.* the asymptote).

I refined Goldwater and Johnson's generalization in order to make room for these two solution curve shapes.

I elaborated why both of these solution curve shapes are compatible with an iterative method of optimization.

For these two toy cases, I demonstrated that the introduction of a Gaussian prior as a regularization term caused both of their solution curves to have a single peak.

Converting one OT constraint ranking into two MaxEnt constraints whose weights can be freely adjusted results in a particular flavor of over-parameterization. One issue with over-parameterization of this sort is that a constraint's true weight cannot be recovered from the observed data (*i.e.* such models are non-identifiable). I showed why regularization cannot help the analyst recover the 'one true weight' of each constraint from the observed data, but instead returns a solution that is even further from the 'one true weight'.

I then applied my sampling technique, which simulates a population of language learners, to the identifiability problem. Under this approach, each trained MaxEnt model is interpreted as a single human speaker. Each speaker only needs to learn one setting of weights that works for them (*cf.* Velar Softening), they do not need to recover the 'one true value' of each constraint, so the existence of multiple equally good solutions is not so bad after all.

CHAPTER 7

Conclusion

In this dissertation, I explored the learning of hidden structure.

In Chapter 2, I reported the results of an experiment that confirmed the productivity of CiV Lengthening. I re-analyzed CiV Lengthening as the emergence of the unmarked Stress-to-Weight Principle (rather than a derived environment effect).

In Chapter 3, I introduced my hidden structure model. The over-arching model architecture was composed of two Maximum Entropy sub-models that were chained together via the product rule. This model was generative – being able to predict the frequecies of words. Its ability to generalize probability to unseen words was also demonstrated. In addition, I introduced my sampling methodology, which simulated a population of language learners.

In Chapter 4, I modeled inter-speaker variation in English Velar Softening. In this phenomenon, speakers learned the same surface pattern for words that currently exist in English, but behaved differently on a generalization task. Empirically, 80% of speakers were found to generalize while the remaining 20% did not. My simulated population of language learners performed similarly to humans: 78% generalized while the remaining 22% did not. Taking a closer look a the lexicon, URs and grammars of the generalizing models showed that they had acquired an alternation analysis. That is, they built up the UR of the *electricity* from the URs of its component morphemes *electric* and *-ity*. They also learned the requisite grammar with its appropriate morphophonological conditioning: turning underlying /k/ to surface [s] only in the presence of a Velar Softening triggering suffix and before the [ɪ, aɪ] sounds. The lexicon, URs and grammars of the non-generalizing models likewise confirmed that these models had acquired a memorization analysis. Instead of building the UR of the *electricity* from the URs of its component morphemes, the UR of *electricity* was stored in the lexicon /ɪlɛktɹɪsɪti/. No grammar that transformed underlying /k/ to surface [s] was

learned.

In Chapter 5, I modeled the learning of Rich Base grammars. The Rich Base problem is as follows: there is one surface pattern that can be produced by two grammars. Humans have been observed to learn only one of these grammars – the Rich Base Grammar – as evidence by loan word adaptation. The Rich Base Grammar is the one that generalizes the surface pattern to problematic URs while the non-Rich Base Grammar does not have this capacity. In my simulated population of language learners, the overwhelming majority learned the Rich Base Grammar (100% for Experiment 1, 98% for Experiment 2). This result was confirmed by submitting these two collections of trained models to a generalization task with the problematic UR.

In Chapter 6, I showed that the overwhelming preference for the Rich Base Grammar was emergent. It emerged from my over-arching model's ability to leverage the superior utility of the Rich Base Grammar (over the non-Rich Base Grammar), such that the learner always experiences an incentive to learn the Rich Base Grammar no matter where it is in the solution curve. I also looked at two common scenarios in phonological modeling: the inclusion of unobserved SR outcomes and the over-parameterization that results from the conversion of one OT constraint ranking into two MaxEnt constraints whose weights can be independently adjusted. These two scenarios can variously result in solution curves that have an infinite number of global maxima (*i.e.* the ridge) or solution curves that do not have even one global maximum (*i.e.* the asymptote). I explained why both of these solution curve shapes were compatible with an iterative method of optimization, and also demonstrated that the usage of a Gaussian prior as regularization resulted in a single peak. I proposed a refinement of Goldwater and Johnson (2003)'s generalization that leaves room for these two solution curve shapes.

APPENDIX A

CiV Lengthening: Test stems (Experiment 1)

		Backness	
Tenseness	**Stress**	*Front*	*Back*
ə	*None*	ˈkudəb	ˈt͡ʃɑkəl
		ˈkâɪɹək	ˈlʌnəm
		ˈmitəv	zɪməp
		ˈnâɪsəg	ˈmɛzən
		ˈsɛpəv	ˈzæləʒ
		ˈtɛpən	ˈlɪzəm
		ˈkʌtəf	ˈzɑbək
		ˈbɑməd͡ʒ	ˈdɑgən
		ˈzɑt͡ʃəp	ˈwiməl
		ˈtɑləp	zipəd͡ʒ
		ˈlæbəp	ˈbɪnən
		ˈsɑsəv	ˈmêɪsəm
		ˈpɪləb	ˈgufəl
		ˈsó͡ʊsəb	ˈnuʒəp
		ˈlêɪt͡ʃət͡ʃ	ˈhɔ̂ɪnəm
		ˈgɪməv	ˈhɪnəv
Tense	*Primary*	səˈnêɪm	ləˈwó͡ʊk
		zəˈpêɪb	pəˈbó͡ʊk
		səˈzêɪp	ləˈʃó͡ʊn
		səˈzêɪk	nəˈzó͡ʊk
		zəˈpêɪl	dəˈló͡ʊn
		pəˈsêɪp	zəˈtó͡ʊb
		jəˈtêɪk	kəˈbó͡ʊl
		d͡ʒəˈdêɪk	ʃəˈsó͡ʊn
	Secondary	ˈnɛˌtêɪl	ˈsɑˌzó͡ʊm
		ˈmâɪˌdêɪf	ˈbɪˌhó͡ʊn
		ˈtó͡ʊˌsêɪg	ˈsɑˌɹó͡ʊk
		ˈtuˌdêɪt͡ʃ	ˈtæˌsó͡ʊp
		ˈzɛˌsêɪb	ˈvɑˌnó͡ʊl
		ˈzɪˌfêɪn	ˈdæˌló͡ʊb
		ˈʒæˌkêɪm	ˈzæˌsó͡ʊk
		ˈθɑˌlêɪn	ˈbâɪˌdó͡ʊn
Lax	*Primary*	ɹəˈnæn	θəˈt͡ʃɑl
		ləˈkæn	kəˈkɑk
		kəˈzæʃ	təˈɹɑk
		səˈdæl	zəˈzɑb
		ɹəˈlæk	ɹəˈlɑn
		həˈzæf	ləˈkɑl
		kəˈkæl	səˈɹɑn
		ʃəˈkæv	təˈlɑp
	Secondary	ˈnɑˌlæg	ˈtiˌdɑn
		ˈdɛˌzæn	ˈt͡ʃɪˌtɑk
		ˈdɪˌkæk	ˈmêɪˌbɑl
		ˈgɪˌtæk	ˈbɪˌnɑm
		ˈzâɪˌtæd͡ʒ	ˈsêɪˌsɑg
		ˈpɑˌtæf	ˈpɛˌkɑg
		ˈnɪˌmæb	ˈlæˌpɑk
		ˈkɪˌt͡ʃæm	ˈluˌsɑn

Table A.1: *Distribution of test stems across conditions (Expt 1)*

References

Akers C (2012). *Commitment-Based Learning of Hiddden Linguistic Structures*. Ph.D. thesis, Rutgers University.

Alderete J, Brasoveanu A, Merchant N, Prince A, Tesar B (2005). "Contrast analysis aids the learning of phonological underlying forms." In C hye Han, A Kochetov (eds.), "Proceedings of the 24th West Coast Conference on Formal Linguistics," pp. 34–42. Cascadilla Proceedings Project, Somerville, MA.

Baković E (2013). *Blocking and Complementarity in Phonological Theory*. Equinox Publishing, Sheffield & Bristol, Conn.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting linear mixed-effects models using lme4." *J. Stat. Softw.*, **67**(1). ISSN 1548-7660. doi:10.18637/jss.v067.i01. URL `http://dx.doi.org/10.18637/jss.v067.i01`.

Becker M, Levine J (2014). "Experigen – an online experiment platform." URL `http://becker.phonologist.org/experigen`.

Berger AL, Pietra VJD, Pietra SAD (1996). "A maximum entropy approach to natural language processing." *Computational Linguistics*, **26**, 39–71.

Berko J (1958). "The Child's learning of English Morphology." *WORD*, **14**, 150–77.

Boersma P (2001). "Phonology-semantics interaction in OT, and its acquisition." In R Kirchner, W Wikeley, J Pater (eds.), "Papers in Experimental and Theoretical Linguistics," volume 6, pp. 24–35. University of Alberta, Edmonton.

Boersma P, Pater J (2016). "Convergence properties of a Gradual Learning Algorithm for Harmonic Grammar." In JJ McCarthy, J Pater (eds.), "Harmonic Grammar and Harmonic Serialism," Equinox Press, London.

Burzio L (2005). "Sources of paradigm uniformity." In LJ Downing, TA Hall, R Raffelsiefen (eds.), "Paradigms in Phonological Theory," pp. 65–106. Oxford University Press, Oxford.

Chomsky N, Halle M (1968). *The sound pattern of English*. Harper and Row, New York.

Chong A (2019). "Exceptionality and derived environment effects: A comparison of Korean and Turkish." *Phonology*, **36**, 543–72.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society*, **Series B (Methodological), 39**(1), 1–38.

Eisenstat S (2009). *Learning underlying forms with MaxEnt*. Master's thesis, Brown University.

Eisner J (2000). "Review of Kager: "Optimality Theory"." *Computational Linguistics*, **26**, 286–90.

Farnetani E, Kori S (1986). "Effects of syllable and word structure on segmental durations in spoken Italian." *Speech Communication*, **5**, 17–34.

Goldwater S, Johnson M (2003). "Learning OT constraint rankings using a maximum entropy model." In J Spenader, A Eriksson, O Dahl (eds.), "Proceedings of the Stockholm Workshop on Variation within Optimality Theory," pp. 111–20.

Hayes B, Zuraw K, Siptár P, Londe Z (2009). "Natural and unnatural constraints in Hungarian vowel harmony." *Language*, **85**, 822–63.

Jarosz G (2006). *Rich Lexicons and Restrictive Grammars – Maximum Likelihood Learning in Optimality Theory*. Ph.D. thesis, John Hopkins University.

Jarosz G (2015). "Expectation Driven Learning of Phonology, Ms." University of Massachusetts, Amherst.

Jarosz G (2017). "Learning with Hidden Structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing." *Language*, **93**, 1–36.

Johnson M (2002). "Optimality-theoretic Lexical Functional Grammar." In P Merlo, S Stevenson (eds.), "The Lexical basis of sentence processing: Formal, computational and experimental issues," pp. 59–72. John Benjamins, Amsterdam, The Netherlands.

Kager R (1999). *Optimality Theory*. Cambridge University Press, Cambridge, UK.

Kang Y (2011). "Loanword phonology." In M van Oostendorp, C Ewen, E Hume, K Rice (eds.), "Companion to Phonology," pp. 2258–82. Wiley Blackwell, Malden, MA.

Kato H, Tsuzaki M, Sagisaka Y (2003). "Functional differences between vowel onsets and offsets in temporal perception of speech: Local-change detection and speaking-rate discrimination." *The Journal of the Acoustical Society of America*, **113**, 3379–89.

Kiparsky P (2000). "Opacity and cyclicity." *The Linguistic Review*, **17**, 351–67.

Legendre G, Miyata Y, Smolensky P (1990). "Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An application." *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 884–91.

Legendre G, Sorace A, Smolensky P (2006). "The Optimality Theory–Harmonic Grammar connection." In P Smolensky, G Legendre (eds.), "The Harmonic Mind," pp. 339–402. MIT Press, Cambridge, MA.

Lubowicz A (2002). "Derived environment effects in Optimality Theory." *Lingua*, **112**, 243–80.

Mayer C (2021). *Issues in Uyghur backness harmony: Corpus, experimental, and computational studies*. Ph.D. thesis, UCLA.

McCarthy J (2000). "Harmonic serialism and parallelism." In "NELS 30," pp. 501–24.

McCarthy J (2003). "Comparative markedness (long version)." In "Papers in Optimality Theory II [University of Massachusetts Occasional Papers in Linguistics 26]," .

McCarthy J (2007). "Restraint of analysis." In "Freedom of analysis?", pp. 203–31. Mouton de Gruyter, Berlin & New York.

McCrary K (2004). *Reassessing the Role of the Syllable in Italian Phonology*. Ph.D. thesis, UCLA.

Merchant N (2008). *Discovering underlying forms: Contrast pairs and ranking*. Ph.D. thesis, Rutgers University.

Myers S (2003). "Vowel shortening in English." *Natural Language and Linguistic Theory*, **5**, 485–518.

Nazarov A, Jarosz G (2017). "Learning parametric stress without domain-specific mechanisms." In "Proceedings of the 2016 Annual Meeting on Phonology," .

Nazarov A, Pater J (2017). "Learning opacity in stratal maximum entropy grammar." *Phonology*, **34**, 299–324.

Nelson M (2019). "Segmentation and UR acquisition with UR constraints." *Proceedings of the Society for Computation in Linguistics*, **2**, 60–68. doi:10.7275/zc9d-pn56.

O'Hara C (2017). "How abstract is more abstract? Learning abstract underlying representations." *Phonology*, **34**, 325–345.

Pater J (2009). "Weighted constraints in generative linguistics." *Cognitive Science*, **33**, 999–1035.

Pater J (2016). "Universal grammar with weighted constraints." In J McCarthy, J Pater (eds.), "Harmonic Grammar and Harmonic Serialism," pp. 1–46. Equinox Press, Sheffield.

Pater J, Staubs R, Jesney K, Smith B (2012). "Learning probabilites over underlying representations." *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pp. 62–71.

Pierrehumbert J (2006). "The statistical basis of an unnatural alternation." In L Goldstein, DH Whalen, CT Best (eds.), "Laboratory Phonology VIII, Varieties of Phonological Competence," pp. 81–107. Mouton de Gruyter, Berlin.

Potts C, Pater J, Jesney K, Bhatt R, Becker M (2010). "Harmonic Grammar with linear programming: From linear systems to linguistic typology." *Phonology*, **27**, 77–117.

Press W, Teukolsky S, Vetterling W, Flannery B (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, England.

Prince A, Smolensky P (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Oxford.

Prince A, Smolensky P (2004). *Optimality Theory: Constraint interaction in generative grammar*. Blackwell Publishing, Malden, MA.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Riad T (1992). *Structures in Germanic prosody: A diachronic study with special reference to Nordic languages*. Ph.D. thesis, Stockholm University.

Ryan KM (2000). "Onsets contribute to syllable weight: Statistical evidence from stress and meter." *Language*, **90**, 309–41.

Smolensky P (1986). "Information processing in dynamical systems: Foundations of harmony theory." In D Rumelhart, J McClelland, the PDP Research Group (eds.), "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations," pp. 194–281. Bradford Books/MIT Press, Cambridge, MA.

Staubs R, Pater J (2016). "Learning serial constraint-based grammars." In J McCarthy, J Pater (eds.), "Harmonic Grammar and Harmonic Serialism," Equinox Press.

Steriade D (2019). "CiV Lengthening and the weight of CV." In TM M Bowler P Duncan, H Torrence (eds.), "Schuhschrift, Papers in Honor of Russell Schuh," University of California.

Tesar B (2004). "Using inconsistency detection to overcome structural ambiguity." *Linguistic Inquiry*, **35**, 219–53.

Tesar B, Alderete J, Horwood G, Merchant N, Nishitani K, Prince A (2003). "Surgery in language learning." In G Garding, M Tsujimura (eds.), "WCCFL 22 Proceedings," pp. 477–90. Cascadilla Press, Somerville, MA.

Tesar B, Prince A (2007). "Using phonotactics to learn phonological alternations." In "Proceedings of the Thirth-ninth conference of the Chicago Linguistics Society," pp. 209–37. Chicago Linguistics Society, Chicago.

Tesar B, Smolensky P (2000). *Learnability in Optimality Theory*. MIT Press, Cambridge, MA.

White J (2017). "Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias." *Language*, **93**, 1–36.

Wilson C (2006). "Learning phonology with substantive bias: An experimental and computational study of velar palatalization." *Cognitive Science*, **30**, 945–82.

Wilson C (2022). "Indentifiability, log-linear models, and Observed/Expected (response to Stanton & Stanton, 2022)."

Wu CFJ (1983). "On the Convergence Properties of the EM Algorithm." *Annals of Statistics*, **11**, 95–103.

Zhu C, Byrd RH, Lu P, Nocedal J (1997). "L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization." *ACM Transactions on Mathematical Software*, **23**, 350–60.

Zuraw K (2000). *Patterned exceptions in phonology*. Ph.D. thesis, UCLA.