Title
Phonotactic probability in Amharic : : a psycholinguistic and computational investigation

Author
Colavin, Rebecca Irene Victoria

Publication Date
2013

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Phonotactic probability in Amharic: a psycholinguistic and computational investigation

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

in

Linguistics

by

Rebecca Irene Victoria Colavin

Committee in charge:

Professor Sharon Rose, Co-Chair
Professor Roger Levy, Co-Chair
Professor Eric Bakovic
Professor Gary Cottrell
Professor Robert Malouf

2013

The dissertation of Rebecca Irene Victoria Colavin is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Co-Chair

University of California, San Diego

2013

DEDICATION

To Francis, who always says *"be happy"*

EPIGRAPH

"With the exception of the 'explosive sounds' represented by p̣, q, ṭ, č̣, ṣ and the guttural χ (frequently pronounced k by Abyssinians themselves) the pronunciation of Amharic presents little difficulty to an Englishman".

—Armbruster (1908)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have long looked forward to writing the acknowledgements section of my dissertation. I did not foresee how difficult it would be to actually put into words my appreciation for all the wonderful people who have inspired, encouraged, supported and occasionally (deservedly) scolded me.

First and foremost, I must thank my advisors, Sharon Rose and Roger Levy. Sharon, you are really a wonderful advisor. Your passion for linguistics and your incisive, meticulous analyses are a high standard that I will strive to live up to. If this was not enough, your sensible advice has guided me throughout the gradutate student experience. There is nothing I can do to adequately thank you. Likewise, it has been a privilege to work with Roger Levy, who generously accepted to co-advise a student working on a subject outside his usual area of interest. With almost limitless patience, Roger introduced me to a view of linguistics that, while challenging, changed my view of the field.

For the other members of my committee, Eric Bakovic, also my advisor for a short while, has always found time to listen to my concerns and try to help find solutions and Gary Cottrell has shown amazing patience with a project that in no way evolved the way we thought it would when he signed on.

Rob Malouf was my advisor at SDSU. Rob, this is all your fault. There is a way you have of believing in your students that is quite irresistible. If I had not met you, I would not be where I am today. You made everything seem, if not easy, at least doable and fun. Thank you.

Bruce Hayes and Colin Wilson, whose work is the foundation of my thesis, have been endlessly kind and patient and their advice has been invaluable. Bruce in particular, has never failed to respond to questions with thoughtful and patient answers. I must also thank Adam Albright, not only for his sharing his insights when we have met, but for his wonderful publications. When I am confused about a problem, his publications are my first go-to solution for getting a clear and precise explanation of the issues. Stefan Frisch and Iris Berent have also been prompt and kind respondents to emails about their work.

I am deeply indebted to the kindness and generosity of many members of

the Ethiopian community. Tadesse Sefer translated the first draft of the experiment, Nunu Belette spent many hours trying to make the experiment instructions sound reasonable and reading lists of invented words, eliminating slang. To my experimental subjects, in the US, Canada and Ethiopia, I extend my deepest thanks for doing an experiment that was not only tiresome but also, out of context, must have seemed completely pointless.

There are many other faculty members who have helped me along the way. Andy Kehler was my advisor until I caught the phonology virus from Sharon. One of my greatest regrets is that I could never quite talk him into doing phonology with us. His insights and advice, not just on linguistic but on life in academia have been invaluable.

It is impossible to imagine the importance of the relationship between graduate students without having lived it. I was lucky to have amazing students in my year, my companions in beer and laughter. That first year, we were closer than family. Klinton Bicknell has always been generous with his time and knowledge, Alex del Giudice has been generous with time, knowledge, his wonderful wife Ping and daughter Zoe. These last few weeks, his daughter Zoe has reminded me of the really important things. For Lisa Rosenfelt, there are no words. Dispensing love and warmth to all around her, she has never failed to kick my ass when it needed kicking, to hug me when I was sad and to remind me that we are both awesome. She turned up with home-cooked food and Champagne when there was sickness in the house and metamorphosed into a very, very scary tigress if anyone messed with me. Lisa, I can't believe my luck in having you as my Bestie.

Lucien Carroll, deserves a special mention[1]. Lucien has been a dear friend for more years than either of us care to remember. He has dispensed hugs in times of need and has, over time, learned the importance of caffeine and beer. He has had endless patience in answering my many, many questions about all things linguistic and computational and has never protested being treated like a walking encyclopaedia.

Cindy Kilpatrick was my roomie, my work buddy and my social life until

[1]Yes, Lisa, he gets his own paragraph

she had the bad taste get her degree and wander off with an amazing job, an amazing husband and an amazing child. Cindy, how could you leave all that we had just for that? More seriously, Cindy managed to convince me that models can be useful and inspired me to work on phonotactics.

The members of the San Diego Phonology Interest Group (aka SadPig) are an amazing group. Bozena Pajak, Bethany Keffala, Hope Morgan, Page Piccini and all the others are an example of a group that is both demanding and supportive.

I am also a member of the Roger Levy's Computational Psycho-Linguistics lab and I must acknowledge the generosity of the lab members, Gabe Doyle, Emily Morgan, Victoria Fossum, Mark Myslin and all the others, in helping me with many thorny mathematical questions. Not only are these people endlessly knowledgeable, but they are always ready to discuss serious theoretical issues involving cylons and bus routes. I have absolutely loved working with you, guys.

Many members of the staff at UCSD have shown me immense kindness. Ezra Van Everbroeck, our IT manager,and Dennis Fink, our multi-media specialist have saved me from the consequences of my own dopiness more times that I care to remember. Our Grad. Admin, Alycia Randol, has been a rock of kindness and practical advice.

For a while, Corie Gochicoa worked with Alycia in our department and she was both kind and competent. However, her great gift to me (and the community) was to encourage us to join her on her lunch time beach walks. Nothing puts your life into perspective like a walk to the beach and perhaps the glimpse of a dolphin or two. The walks have continued since Corie abandoned us for sunnier climes and I hope they will continue after I leave, an enduring tribute to the importance of remembering the beauty around us.

There are non-academic friends in particular who have helped me over the last few years. Teri Gray and I have been friends for so long, we don't ever discuss it. She is my touchstone to the real world outside of academia. Sotiria Lampoudi has been a bracing, sometimes very bracing, and supportive friend, often taking time to explain computational, statistical and mathematical issues in really small words. Thank you also to Osvaldo Colavin, who must often have wondered what

the hell I was up to yet managed to always sound kind and supportive.

I don't believe any thesis on the west side of campus could possibly be written without the support of Sam and his awesome staff at the Art of Espresso coffee cart at Mandeville. Most of us have worked there, met with friends and advisors and even held office hours. Working in the sunshine with a cup of the good stuff made life significantly ($p<.001$) less grim.

Finding the words to thank my family for putting up with me all these years, through all the ups and downs is perhaps the most difficult thing of all. Thank you to my sons, Alex and Vincent, for always encouraging me and never, ever, letting me doubt that I could do it. I think we can say that you raised me real good.

To Francis Cheung, my partner and personal hero, I say nothing here because words are inadequate. However, Francis wasn't just there for me, he helped many of my fellow grads with his kindness and generosity. He fed them, lent them cars, listened to their problems, hosted bridal showers, baby showers and parties and generally spoilt them to death. So, I want to take this opportunity of thanking him, not just in my name but for all the behind-the-scenes work he did to make things a little easier for the community.

VITA

1984	Diplome d'Infirmières, Ecole d'Infirmières, Villeneuve St Georges, France
2002	Bachelor of Arts, San Diego State University, CA
2004	San Diego State University, CA, course work for MA.
2007	Master of Arts, University of California, San Diego, CA
2005-2013	Teaching assistant University of California, San Diego
2010	Candidate of Philosphy, University of California, San Diego, CA
2013	Doctor of Philosphy, University of California

PUBLICATIONS

Rebecca Colavin, Sharon Rose and Roger Levy. 2012 Under-representation and word-acceptability in Amharic: Evidence from a judgement task. LSA 86, Portland.

Rebecca Colavin, Roger Levy and Sharon Rose. 2010. Modeling OCP-Place in Amharic with the Maximum Entropy Phonotactic Learner. Proceedings of CLS 46.

Rebecca Colavin, Roger Levy, and Sharon Rose. Feb. 2010. Modeling OCP-Place with the Maximum Entropy Phonotactic Learner. Workshop on Computational Modeling of Sound Pattern Acquisition, Edmonton, AB.

Rebecca Colavin, Roger Levy, and Sharon Rose. Apr. 2010. Modeling OCP-Place with the Maximum Entropy Phonotactic Learner. CLS 46, Chicago.

ABSTRACT OF THE DISSERTATION

Phonotactic probability in Amharic: a psycholinguistic and computational investigation

by

Rebecca Irene Victoria Colavin

Doctor of Philosophy in Linguistics

University of California, San Diego, 2013

Professor Sharon Rose, Co-Chair
Professor Roger Levy, Co-Chair

In this dissertation, the robustness of the relationship between the lexical frequency of phonotactic patterns and word-acceptability is examined for words of Amharic, an Ethiopian Semitic language. The patterns under investigation span the whole verb root and include both under-represented and over-represented consonant distributions in the lexicon. A state-of-the-art probabilistic model, the Maximum Entropy phonotactic learner, is used to acquire a phonotactic grammar from the input (the lexicon) and the predictions of that grammar are compared with the results of two Amharic nonce-word rating tasks designed specifically to investigate a range of consonantal phonotactic patterns. The first task investigates consonant

co-occurrence patterns (homorganic consonants, identical consonants, and fricatives). In the Amharic verb lexicon, identical consonants are under-represented in some locations and over-represented in others whereas homorganic consonants and fricatives (a previously unknown pattern independently acquired by the model) are under-represented. The phonotactic learner successfully learned the under-represented patterns and the comparison between the model predictions and the experimental results show evidence for a relationship between lexical frequency and word acceptability for under-representation. However, speaker judgements show no preference for over-representation. The second task examines the distribution of single consonants within the verb root with respect to under-representation, over-representation and positional restrictions. Evidence for a relationship between lexical frequency and phonotactic probability was observed for both under-represented and over-represented consonants, but tied to a particular location. The correlation between speaker judgments and model predictions is low for this task, due in part to the way the model deals with over-representation. This investigation demonstrates not only that word acceptability is influenced by phonotactic probability for both under-represented and over-represented patterns, but also that probabilistic models can be used to investigate the phonotactics of a language, even in the absence of speaker judgement data. These models can therefore be used to assess the phonotactics of languages where experimental data is difficult to obtain and broaden our knowledge of phonotactic typology.

1 Introduction

Sounds are distributed within words according to particular patterns. The range of possible patterns is extremely diverse; they may be simple, such as the prohibition of the sound [ŋ] at the beginning of English words (although it occurs frequently in word final position) or complex (for example, in Semitic languages, consonants from the same place of articulation tend to not co-occur within verb roots). These *phonotactic* generalizations, or rules that constrain the allowed patterns, are part of the grammar of a language. Furthermore, research shows that these patterns are not just fossilized evidence of diachronic change; speakers are psychologically aware of the sound distributions that pertain to their language. In word rating tasks, when speakers are asked to assign ratings according to how acceptable a word (real or invented) would be in their language, results show a relationship between the frequency of phonotactic patterns in the lexicon and word acceptability such that the acceptability of a word is a function of the lexical frequency of the patterns that it contains (all else being equal) (Albright and Hayes, 2003; Bailey and Hahn, 2001; Coleman and Pierrehumbert, 1997; Frisch et al., 2000; Hammond, 2004; Hay et al., 2003; Ohala and Ohala, 1986; Treiman et al., 2000; Vitevitch et al., 1997).

Regarding experimental phonotactic research, there is ample evidence that speakers can reliably distinguish between nonce words containing only patterns that occur in the lexicon (*attested* patterns) and those that contain patterns that do not occur in the lexicon (*unattested*) patterns)(Scholes, 1966). For example, English speakers are likely to prefer the nonce word ‘blick’ that contains only sequences that occur in the lexicon over ‘bnick’ that contains the sequence [bn] which never occurs at the beginning of English words. Speakers have also been shown to be sensitive to differences in acceptability among words with patterns that do not occur in the lexicon, based on the lexical frequency of their constituent natural class sequences. For example, speakers reliably prefer 'bnick' to 'bzick' even though [bn] and [bz] are unattested as onset sequences and this is predictable because other stop-sonorant onset patterns do occur in the lexicon, but stop-fricative sequences do not (Albright, 2009).

Experimental studies that evaluate gradient speaker judgements for nonce words and control for phonotactic probability have primarily focused on specific sub-parts of English words. For example, Scholes (1966), Coleman (1996), Albright and Hayes (2003), Hay et al. (2004), (Treiman et al., 2000) and Albright (2011a) have collected English speaker judgements for nonce words that differed in the probability of their onsets, rimes or consonant clusters. Judgements tasks for whole words (where the difference in probability between experimental items is not limited to a specific sub-lexical unit) are somewhat rarer (Frisch et al., 2000; Vitevitch et al., 1997). Research focusing on whole words in languages other than English include Frisch and Zawaydeh (2001) for Arabic, Myers and Tsay (2005) for Mandarin, and Kirby and Yu (2007) for Cantonese. The goal of this thesis is therefore to examine the robustness of the relationship between the lexical frequency of phonotactic patterns and word acceptability for whole words of Amharic, a Semitic language that is under-studied compared to Arabic and Hebrew and that has complex and interacting phonotactic patterns. We use a state-of-the-art probabilistic model, the Maximum Entropy phonotactic learner (Hayes and Wilson, 2008) to acquire a phonotactic grammar from the input (the lexicon). The predictions of that grammar are compared to the results of two Amharic nonce-word rating tasks designed specifically to investigate a broad range of phonotactic patterns. The comparison between the model predictions and the experimental results shows 1) evidence for the breadth of the relationship between lexical frequency and word acceptability in terms of both under-representation and over-representation of patterns, 2) the importance of the choice of representational systems used by the model and 3) that when interpreted with prudence, the predictions of models such as the Maxent learner can be used to discover previously unstudied restrictions.

In Chapter 2, we review the assumptions and performance of the most influential computational models of word acceptability (Albright, 2009; Bailey and Hahn, 2001; Coleman and Pierrehumbert, 1997; Hayes and Wilson, 2008; Vitevitch and Luce, 2004). Phonotactic probability models are designed to acquire a phonotactic grammar based on the frequency of sound patterns in the lexicon.[1] We determine that the Maximum Entropy (Maxent) phonotactic learner of Hayes and Wilson (2008) has characteristics that recommend it compared to other models; using a statistically well-founded methodology, the Maxent learner acquires a phonotactic grammar of weighted constraints from the input that assigns penalties to patterns that are under represented in the lexicon. The phonologically motivated representational system is rich and purposely designed to permit the modelling of languages other than English. Finally, the performance of the model in predicting speaker judgements, as measured by the correlation between the model predictions and speaker judgements of unattested English onsets, is higher than for any other model to date.

Chapter 3 describes the pertinent lexical, morphological and phonotactic characteristics of the lexicon of Amharic verb roots drawn from Kane (1990). Amharic recommends itself for a number of reasons. First, it is an under-studied language in many respects, especially for investigations of its phonology and lexicon, and we believe that wherever possible, researchers should attempt to add to the body of linguistic knowledge in the broadest way possible. Secondly, like other Semitic languages (such as Arabic and Hebrew), Amharic has root-and-template morphology. In this paradigm, semantically related words share a common consonantal *root* and different derivations are obtained by modifications of the intervening vowels and consonants according to specific templates. The assumption is that if speakers are presented with verb roots in a given derivational form, the differences in judgements between them will be motivated only by the consonants of the root as all other material is constant. Finally, other well-documented Semitic languages, Arabic and Hebrew, provide a comparative counterpoint to our results.

For our purposes, the most important characteristic of Amharic (like other

[1] Although word types are the most commonly used learning data, word token frequency may also be taken into consideration (Vitevitch and Luce, 2004).

Semitic languages) is that verb roots are subject to two robust and well-studied restrictions. The first is a restriction over homorganic consonants called the Obligatory Contour Principle as applied to place of articulation (OCP-Place). This is a disharmony restriction that prohibits segments from the same place of articulation from co-occurring within the same verb root, whether it be at the left edge (C_1C_2X) of a tri-consonantal root, the right edge (XC_2C_3) or in non-adjacent (C_1XC_3) position.[2] Crucially, OCP-Place is a *gradient* restriction; although there are few cases where homorganic consonants co-occur in the lexicon, they do exist and the frequency of their occurrence is modulated by place of articulation and location within the word. A second important characteristic of OCP-Place is that it is a long distance restriction in two respects: 1) it is active over the consonants of the verb root, unaffected by intervening materials and 2) root consonants in non-adjacent position (C_1XC_3) are also restricted, showing that the restriction is active even across an intervening root consonant.

In Amharic, there is also a restriction over the co-occurrence of *identical* consonants within verb roots. As for other Semitic languages, identical consonants co occur rarely within verb roots at the left edge (C_1C_2X) or in non-adjacent (C_1XC_3) position but very frequently at the right edge (XC_2C_3). This pattern of identical consonants, often analysed as arising through a the restriction on the co-occurrence of identical consonants in underlying representation (McCarthy, 1981, 1986, 1988), poses a challenge to the modelling of OCP-Place. The model must learn that homorganic consonants are generally restricted in left edge and non-adjacent locations, but that only *non-identical* homorganic consonants are restricted in the right edge location, where identical consonants are frequent. How to represent identity becomes an important issue in the assessment of the model's performance.

Chapter 4 presents two series of simulations using the Maxent learner to model speaker judgements for nonce words with OCP-Place violations. We evaluate models both on their statistical fit (log-likelihood of the learning data) and by measuring the correlation between the model predictions and the averaged word

[2]Although the examples given involve tri-consonantal roots, these restrictions also holds for longer roots.

ratings from a pre-existing word rating task (King and Rose, 2003). We compare the results of two models, an automatically learned model and a model with hand-written constraints capturing the restriction on OCP-Place as it occurs in the lexicon of Amharic verb roots and completed with automatically acquired constraints. Unlike the Hayes and Wilson simulations for English, our models are extremely large (each contains 1000 constraints) in an attempt to capture the full range of phonotactic probability. In the first simulation, the consonants of the verb roots in the training data are encoded in their surface-true form and the results show that the automatic model performed significantly less well in predicting speaker judgements than the hand-written model. Our analysis indicates that the automatic model was not deficient in learning OCP-Place, but rather that the frequency of identical consonants at the right edge of words leads the model to over-estimate the importance of an independent harmony restriction on the co occurrence of voiceless stops. A second simulation with a modified encoding of successive occurrences of identical consonants so that they are no longer surface true showed that the automatic model was no longer deficient compared to the hand-written one.

The results of the simulations provide evidence for greater flexibility in the representational systems of computational models and in particular that the second occurrence of a consonant at the right edge of verb roots should not include the same information load, in terms of distinctive features, as the first occurrence.

The results of the simulations also indicate the specific conditions for which the model makes predictions but for which no test data exists for Amharic. Based on the predictions from the automatically learned models, we determined two important areas to investigate through word rating tasks: 1) nonce words with the segments that tend to not co occur and 2) nonce words with segments representing a range of distributional patterns, including segments that are either rare or very frequent and segments that occur primarily in specific locations in the word.

Chapter 5 describes Experiment I, an investigation of speaker judgements for differing patterns of consonantal co-occurrence in verb roots. The conditions include OCP-Place violations, the co-occurrence of identical consonants in all locations, and the co-occurrence of non-homorganic fricatives (a co-occurrence restriction predicted by the automatic model and confirmed by the statistical analysis of the lexicon, but previously unstudied). In the discussion of the methodology, we motivate the selection of experimental items for the task. Unlike other researchers in this area, we distinguish the selection of controls (defined as nonce words that are statistically similar to the members of the set of phonotactically legal real words) from the selection of stimuli (defined as nonce forms that are are statistically representative of the set of nonce words available for a narrowly defined experimental condition). The experimental items thus selected are presented in an online rating task. The speaker ratings show that under-represented co-occurrence restrictions (OCP-Place, non-homorganic fricatives and identical consonants in left edge and non-adjacent locations) are rated as less acceptable than controls, but that the over-represented patterns of identical consonants at the right edge are rated similarly to controls, rather than as more acceptable as the frequency $\sim$ acceptability hypothesis predicts.

Chapter 6 describes experiment II which investigates speaker ratings for differing levels of segmental frequency. The conditions include nonce stimuli with segments that are generally over-represented, under-represented or distributed irregularly over the three possible root locations. To avoid the possibility that the low acceptability ratings for very rare segments could over-shadow the more nuanced rating differences for moderately under- and over-represented segments, the task is divided into two sections. In the first part, speakers are exposed only to nonce forms with over-represented and moderately under-represented segments and the second part includes stimuli for a broader range of segmental frequencies. The analysis of results of Experiment II shows that speakers may, in optimal experimental circumstances, assign higher acceptability ratings to nonce words with with over-represented segments than to controls. As the Maxent learner performs poorly in predicting these results, they provide evidence that computational models should be equipped to make predictions for the full range of phonotactic gradiency, rather than just predicting that under-represented patterns are less acceptable..

The final chapter is a general discussion of the results, both from the simulations and the experimental tasks, and presents avenues for future research.

2 Phonotactic models

2.1 Phonotactics

It has long been observed that for a given language, sounds do not occur freely in all positions of a word or syllable. Consider the case of the segment [h] in English. [h] occurs 1153 times as the onset of a word in the CMU online pronouncing dictionary (http://www.speech.cs.cmu.edu) but it never occurs in word or syllable final position. Restrictions over sound *sequences* may also be defined in a specific word or syllable position. For example, in English, the consonant cluster [rk] is perfectly acceptable in word and syllable final position, but never appears word or syllable initially. Restrictions may operate across intervening material. In Shona, a Bantu language, the vowel [e] may only occur in non-initial position if the preceding vowel in the word is [e] or [o] (Beckman, 1997; Riggle, 1999). Restrictions may be *absolute*, as in the case of [h] in word and syllable final position in English, or non-absolute, in cases where a pattern occurs in the lexicon, but less frequently than would be expected, all else being equal. These last are called *gradient* restrictions. For example, in many Semitic languages there is a general restriction against the co-occurrence of homorganic (same place of articulation) consonants in verb roots but this is not an absolute restriction; there are some verb roots in the lexicon in which labial or coronal consonants do co-occur (Bender and Fulass, 1978; Frisch et al., 2004; Greenberg, 1950; Rose and King, 2007). The restrictions on the distribution of sounds within a language (of which these examples are but a small sample) are called the phonotactics of the language.

2.2 The generative linguistic tradition

An implicit assumption in generative linguistics is that the systematic gaps in the lexicon of a language are the manifestation of a grammar or set of rules that prohibits specific patterns.[1] The theory also assumes a representation where the phonemes [2] of a language can, at a more abstract level, be represented as a set (or *bundle*) of sub-segmental phonetic characteristics called distinctive features. Distinctive features reference characteristics such as place of articulation and voicing, for consonants, or height and rounding, for vowels. Sets of sounds that share one or more distinctive features are called a *natural class* and have been shown to behave similarly in a given context. For example, in English, /p, t, k/, the members of the natural class of voiceless stops, become aspirated [p^h, t^h, k^h] at the beginning of words and stressed syllables.

Phonotactic rules in linguistics are therefore expressed in terms of distinctive features and natural classes and may refer to syllable positions (such as onset, coda and rhyme) as well as word edges. Any word violating one or more of the rules would be considered illegal. For example, the fact that English syllables cannot begin with sonorant-stop sequences might be expressed as *$[+sonorant][-continuant] (where '$' indicates a syllable boundary).

Figure 2.1 shows a graphical representation of the space of possible words of English which is defined as the set of all possible combinations of the sounds in the native sound inventory[3]. Note that, for consistency, all strings (segment combinations) are called words, regardless of their phonotactic well-formedness. The set of all possible words, Ω, is framed by the rectangle. This is an infinite set because it contains all possible combinations of segments with no upper bound on length. Within that space, the pink area contains words with phonotactic patterns that are attested in the lexicon. This area includes the words of the lexicon itself. The words in the pink area that are not a part of the lexicon are considered to be

[1]There does not appear to be a specific reference for the pervasive assumption that the grammar describes illegal forms only.

[2]Phonemes are the sound units of a language that can convey a contrast in meaning. For example, /p/ and /k/ are different phonemes in English because the minimal pair /pæt/ (*'pat'*) and /kæt/ (*'cat'*) do not have the same meaning.

[3]This representation assumes that syllabification rules are part of the phonotactic grammar.

accidental gaps in the lexicon; they are not in the lexicon but neither are they ruled out by the grammar. The white area is the set of words that contain sequences that are unattested in the lexicon. According to the traditional generative view, the phonotactic grammar is the set of rules or generalizations over unattested patterns that eliminates words in this area as legal words of English. The words of the lexicon (hereafter *real words*) necessarily contain only attested phonotactic patterns. All the other words in the word space are *nonce words*. Because they are not part of the lexicon, nonce words do not have an associated meaning. Nonce words may contain attested and/or unattested patterns.

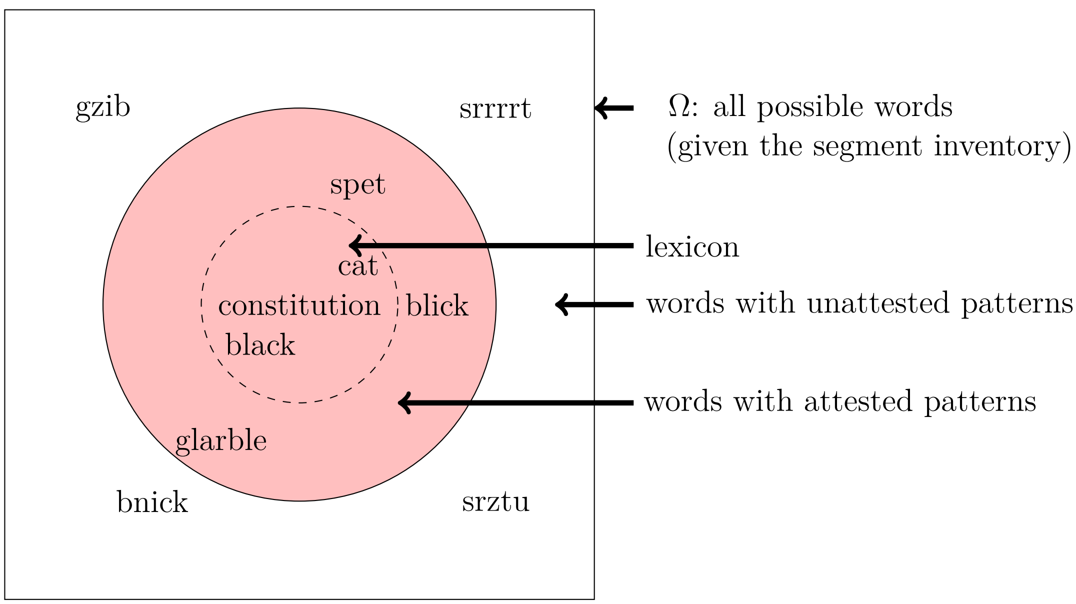


Figure 2.1: Structure of the (English) word space

The generative tradition does not attempt to account directly for the *frequency* of patterns in the lexicon. Words are either legal with regard to the grammar, or illegal; that some patterns are more frequent than others (for example, [k] is a more frequent onset than [g]) is not meaningful. However, the presence or absence of a segmental pattern in the lexicon is not sufficient to determine its status with regard to legality. Consider the case of the English onset [sf]. This pattern could be regarded as illegal because it is rare (‘sphinx’, ‘sphincter’) and because there are no other s-fricative onset combinations in the language. However, the onset [stw] never occurs in words of English but could be analysed as an *accidental gap* in the lexicon because [s] can combine with many other two-segment onset clusters where the first segment is a voiceless stop ([skw], [str], [spr]...). Such analyses are controversial because the predicted boundary between legal and illegal patterns does not correspond to the observable data.

2.3 Experimental evidence

The goal of phonotactic research is to provide evidence that speakers have access to a *phonotactic grammar* containing knowledge about the sound patterns of their language, distinct from the mental lexicon. Over the past 25 years, researchers have have evaluated the extent and structure of speaker’s phonotactic knowledge by investigating the *acceptability* of nonce words, the degree to which a nonce word is an acceptable word of the language according to native speakers.

In such studies, speakers are asked to rate the acceptability of nonce words with specific statistical and phonotactic characteristics. The results show that, given a scalar judgement task, speakers do not assign binary legal/illegal acceptability ratings as a function of whether or not a nonce word violates the phonotactic grammar of the language. In experiments by Bailey and Hahn (2001); Frisch et al. (2000); Hammond (2004); Hay et al. (2004); Ohala and Ohala (1986); Treiman et al. (2000); Vitevitch et al. (1997) speakers were asked to rate words on a scale of acceptability and although each study asked slightly different questions (e.g. “how far from English is this word?” or “how good a word of English would this be?”) and the scales varied (between 7 and 11 points), speakers chose to give a gradient acceptability rating rather than a binary legal/illegal judgement. Note that real words are not usually included in these tasks because, 1) by definition, they are acceptable words of the language and tend to be assigned optimum ratings regardless of their phonotactic composition and 2) their presence tends to disrupt the judgements for nonce forms in the same task (Albright, 2009; Shademan, 2007).

The case might be made that speakers feel obligated to use the gradient scale if one is made available to them and that their judgements are implicitly binary. This appears somewhat unlikely because 1) results show that, within each experiment at least, speakers have similar gradient judgements of nonce words, and 2) as will be discussed, gradient speaker judgements are reproduceable and *predictable* based on the statistics of the lexicon.

In a rating task investigating the acceptability of English onsets, we might be unsurprised that speakers would prefer the nonce word [blɪk] compared to [θwɪk], based on the observation that there are more words in the lexicon that begin with [bl] than with [θw]. We might also expect that speakers would assign identical low scores to nonce words with sound sequences that do not occur in the lexicon, such as [bnɪg] and [bzɪg], and for which they therefore have no frame of reference. In fact, it appears that speakers prefer nonce words with some unattested sequences over others. For example, Berent et al. (2007) showed that English speakers were more likely to reanalyze [lbɪf] as disyllabic than [bdɪf], showing that [lb] is a less acceptable onset for English speakers than [bd] even though neither sound sequence appears in any English syllable. Albright (2011b) obtained English speaker judgements for nonce words containing a range of unattested sequences and showed that speakers reliably preferred the onset [bn] over [bd] and [bz][4].

2.4 Accounting for gradient speaker judgements

Given the experimental evidence, researchers have attempted to create models accounting for gradient speaker judgements. A crucial issue for many researchers is whether gradient speaker judgements of nonce words are motivated by an underlying phonotactic grammar. According to the traditional generative linguistic view, gradient speaker judgements for nonce words with patterns that are under-represented or absent from the lexicon are accounted for by the phono-

[4]Albright attributes this preference to the fact that [-sonorant][+sonorant] natural class sequences occur in onsets, whereas [-sonorant][-sonorant] sequences do not. However, it might be argued (as Berent et al. (2007) do) that speaker knowledge of the sonority scale is innate or that speakers can learn the sonority scale from the input (Daland et al., 2011).

tactic grammar but gradient speaker judgements for nonce forms that contain patterns that occur commonly in the lexicon are accounted for by other mechanisms (Frisch and Zawaydeh, 2001; Hayes and Wilson, 2008). Note that according to this analysis, speaker judgements for nonce words that violate the grammar might be expected to be qualitatively different than for other irregularities (see Becker et al. (2011) and Hayes et al. (2009) for a discussion of natural versus unnatural restrictions). For other researchers, gradient speaker judgements always reflect of the statistics of the lexicon (Albright, 2009; Vitevitch and Luce, 2004).

Although the theoretical debate between these points of view (and the hybrids between them) is not without interest, the goal of this thesis is to investigate the relationship between the input (the lexicon) and speaker judgements. In the rest of this chapter, we present the principal models of gradient speaker judgements and examine their performance in predicting speaker judgement data.

In the literature, models accounting for gradient speaker judgements are usually divided into two groups, analogical models and phonotactic probability models. For analogical models, gradient speaker judgements arise through analogy with existing words and reflect the *wordlikeness* rather than the *acceptability* of a word. A metric commonly used to evaluate wordlikeness is Neighborhood Density (Luce, 1986). In its most basic formulation, the Neighbourhood Density (ND) value of a word (real or nonce) is the number of its nearest neighbours where a neighbour is defined as a real word within a single segment edit distance. For example, the nonce word 'blick' has many neighbors ('flick', 'black', 'blip'...) but 'sfip' has very few ('slip', 'skip') so 'blick' is predicted to be more wordlike that 'sfip'. Ohala and Ohala (1986) show a relationship between speaker ratings for nonce words and ND using this single segment edit distance metric. For phonotactic probability models, (which are also called lexical models by some authors), gradient speaker judgements arise through speaker knowledge of the statistical qualities of the lexicon. For example, the English onset [skl] ('sclera') occurs much less frequently than the onset [st] (1 and 521 occurrences respectively[5] based on the 125,000 plus words of the CMU Pronouncing Dictionary) and this likely accounts

[5]We use the CMU dictionary, www.speech.cs.cmu.edu, for consistency with Hayes and Wilson (2008).

for the intuition that [sklɪp] is a much less likely word of English than [stɪp]. The phonotactic probability approach is supported by experimental evidence showing a relationship between gradient speaker ratings of nonce words and the probability of their constituent sound sequences (Albright, 2009; Hammond, 2004; Vitevitch et al., 1997). For example, in a task where subjects were asked to indicate a preference between two nonce words, Vitevitch et al. (1997) showed that speakers preferred forms containing high-probability sounds or sounds sequences (such as [fʌlˈʧʌn], for example) to those with low-probability sounds or sound sequences (such as [ðaɪbˈʤaɪz]). Phonotactic probability models formalize such observations by assigning a probability distribution (based on some version of the lexicon) over possible words of the language. The hypothesis is that, at least for nonce words, the probability of a form is monotonically related to its acceptability such that higher probability implies higher acceptability.

Historically, phonotactic probability models have really been taken to mean "n-gram models" but if that definition is broadened to mean any model that assigns a probability distribution over possible phoneme strings, then one can construct a phonotactic probability model that subsumes analogical (neighbourhood density) effects. For example, the simple phonotactic probability models such as Vitevitch and Luce (2004) and Albright (2009), described in the next sections, compute the probability of a word based on the probability of the component unigrams or bigrams and do not reflect neighbourhood density in their predictions. However, a more sophisticated phonotactic probability model would assign an acceptability rating to a word based on the probability of *all* the component sequences and therefore subsume the analogical model. For example, for such a model the predicted acceptability for a word such as [blɪk] ('blick') is a function not only of the probability of the component unigrams [b], [l], [ɪ], [k], and bigrams [#b], [bl], [lɪ], [ɪk], [k#], but also of longer (possibly discontinuous) sequences such as [lɪk], [#b_ɪk], [#bl_k#], and [#blɪ] that directly reflect neighbourhood density. The construction of such a model would be complicated by the large number of possible forms to consider, given the learning data (lexicon). For a description of how such a model could be implemented using kernel density estimation see Levy (2012).

In the rest of this chapter, we present the most influential models of gradient speaker judgements and examine their performance in predicting speaker judgement data.

2.4.1 The Generalized Neighbourhood Model

Bailey and Hahn (2001) note that one of the issues with the use of ND as an analogical metric is that many nonce words have few or no neighbours. They propose the Generalized Neighborhood Model, adapted from Nosofsky (1986), which computes wordlikeness using a more sophisticated metric of neighbourhood density based on the *phonological similarity* between words. Phonological similarity is a modified version of the similarity metric proposed by Frisch (1996) and Frisch et al. (2004). Equation 2.1 shows how that the similarity between two words i and j is a function of the natural classes of which they are both members, divided by the sum of the number of natural classes that they both belong to and that they each belong to individually (non-shared natural classes). The GNM similarity score (equation 2.2) is a function of the frequency-weighted phonological similarity between a word i and all of the other words j in the lexicon.

$$\text{Similarity}_{ij} = \frac{\text{Shared natural classes}}{\text{Shared natural classes} + \text{non-shared natural classes}} \tag{2.1}$$

$$\text{Similarity Score}_i = \sum_{j \in lexicon} \text{frequency weighted similarity (i, j)} \tag{2.2}$$

The GNM has two powerful advantages over the usual GN metric:

- GNM is sensitive to sub-segmental information. For ND, the similarity between [tʌb] ‘tub’ and [dʌb] ‘dub’ is the same as between [tʌb] ‘tub’ and [rʌb] ‘rub’, as there is a one segment difference between both pairs. However, the GNM finds greater similarity between [tʌb] and [dʌb] because [t] and [d] are more similar, in terms of natural classes (both are coronal stops that differ only by the feature [voice]), than [t] and [r] (a coronal stop and a coronal sonorant, respectively) which differ by a larger set of features (e.g. [sonorant], [voice], [continuant]).

- Under ND, two words are either neighbours or they are not. For example, [trʌk] ‘truck’ and [trɪk] ‘trick’ are neighbours but [trʌk] ‘truck’ and [trɪm] ‘trim’ are not, because there is more than one segment difference between them. However, under GNM, the similarity metric is much more sensitive, so both [trɪk] and [trɪm] (and many other words) make a contribution to the GNM value of [trʌk].

The strength of the GNM is that it makes phonologically motivated predictions for both nonce words with attested sequences and nonce words with unattested sequences. This contrasts with the two models which are described next, the onset-rime model of Coleman and Pierrehumbert (1997) and the Phonotactic calculator (Vitevitch, 1998; Vitevitch and Luce, 2004) which cannot, given their representational systems, make predictions that differentiate between unattested sequences.

2.4.2 Low-order n-gram models

The Onset-Rime model Coleman and Pierrehumbert (1997) accounts for gradient speaker judgement data with a probabilistic model that uses a sub-lexical representation. These researchers calculate the probability of all the onsets and rhymes in English (modulated by word position and stress) and define the acceptability of a novel word as the the product of the probabilities of the constituent parts determined according to the best parse:

$$\text{Acceptability [blɪk]} \propto \text{P(onset [bl])} * \text{P(rime [ɪk])} \tag{2.3}$$

They find a significant correlation between their model predictions and speaker ratings. The relationship between gradient speaker judgements for English nonce words and the probability of their onset and rime constituents was confirmed by Frisch et al. (2000) and for VC pattern rimes by Treiman et al. (2000) (both of these studies used their own independently created nonce words).

A weakness of the Coleman and Pierrehumbert model is that it cannot account for differences in speaker judgements between nonce words with unattested sequences because all unattested onsets and rimes are assigned the same low probability by default. For example, the model predicts the same acceptability rating for both [bwɪp] and [bzɪp] although according to Albright (2011b), the onset [bw] is reliably rated as more acceptable than [bz].

The phonotactic probability calculator. Segmental n-gram models of speaker judgements express statistical regularities as probability distributions over segments (unigrams) and segment sequences (usually bigrams). An example of a simple model in this category is the phonotactic probability calculator (Vitevitch, 1998; Vitevitch and Luce, 2004). Based on the Kucera and Francis (1967) frequency dictionary,[6] the phonotactic probability calculator computes positional unigram and positional bigram frequencies. Based on the methodological descriptions in Vitevitch and Luce (2004), Equation 2.4 shows the positional unigram frequency for segment i in position a:

$$\mathrm{KF}_{i_a} = \text{set of word types in KF that have segment } i \text{ in position } a$$
$$\mathrm{N}_a = \text{number of word tokens in KF that have a segment in position } a$$

$$F(i_a) = \frac{\sum_{x \in \mathrm{KF}_{i_a}} log_{10}(\text{frequency x })}{log_{10}(\mathrm{N}_a)} \tag{2.4}$$

The log frequencies of all the word types in the dictionary that have segment i in position a are summed and then divided by the frequency of word tokens in the dictionary that have a segment in position a. Similarly for positional bigram frequencies, the log frequencies of all the word types that have i in position a and j in position b are summed and divided by the frequency of word tokens in the dictionary that have segments in a and b:

[6] The Kučera and Francis frequency dictionary is a list of the unique words (types) and their associated frequencies (tokens) in the Brown corpus.

$$\mathrm{KF}_{i_a j_b} = \text{set of word types in KF that have } i \text{ in } a \text{ and } j \text{ in } b$$
$$\mathrm{N}_{ab} = \text{number of word tokens in KF having segments in } a \text{ and } b$$

$$F(i_a j_b) = \frac{\sum_{x \in \mathrm{KF}_{i_a}} log_{10}(\text{frequency x })}{log_{10}(\mathrm{N}_{ab})} \tag{2.5}$$

This transcription of the computations is accurate based on the description in Vitevitch and Luce (2004) but it seems likely that, as the phonotactic probability calculator never returns values above 1, the computation is based on relative frequencies rather than counts.

$$\begin{gathered}\text{acceptability (blɪk)} \approx \mathrm{F}(\mathrm{b}_1) + \mathrm{F}(\mathrm{l}_2) + \mathrm{F}(\text{ɪ}_3) + \mathrm{F}(\mathrm{k}_4) \\ \text{and / or} \\ \text{acceptability (blɪk)} \approx \mathrm{F}(\mathrm{b}_1\mathrm{l}_2) + \mathrm{F}(\mathrm{l}_2\text{ɪ}_3) + \mathrm{F}(\text{ɪ}_3\mathrm{k}_4)\end{gathered} \tag{2.6}$$

Vitevitch and Luce use log-values of relative frequencies rather than just raw relative frequencies to approximate type based estimates because these have been shown to correlate better with performance.[7] The acceptability of a nonce word, according to this model, is a function of both the sum of the constituent positional unigram frequencies and the sum of the constituent bigram log frequencies (equation 2.6). These two numbers are compared to the mean unigram and bigram frequencies for the dataset under consideration. Words for which both measures are above the median are predicted to have high acceptability; those below the median are predicted to have low phonotactic acceptability.

The phonotactic probability calculator is sometimes used as a baseline model against which to test more sophisticated models of speaker judgements (Albright, 2009) or to design stimuli (Berent, 2008) and, in the limited domain

[7]The use of type frequency was also shown to produce models that better predicted speaker judgements than token frequency in both Hayes and Wilson 2008 and Bailey and Hahn 2001.

of nonce words with attested English onsets at least, its performance is comparable to more sophisticated models (Albright, 2009). However, it was not really designed to capture fine differences in acceptability and has limited capacity to distinguish between two words with unattested sequences (where at least some of the constituent segment sequences have a frequency of 0). For example, experimental evidence (Moreton 2002) shows that speakers have a strong perceptual bias against [dl] compared to [bw] (and therefore would likely judge [bw] to be a better onset than [dl]). For the phonotactic probability calculator, the bigram frequency measure is the same for both words (because in both cases, a constituent bigram has a probability of 0) and the log-probability based on the constituent unigrams assigns a higher acceptability score to [dl] than [bw], contrary to the predicted speaker preference.

2.5 Accounting for unattested sequences

The phonotactic probability models presented in the previous section (the onset-rime model and the phonotactic probability calculator) are ill-equipped to distinguish between nonce words with unattested sequences ([bnɪk] and [bzɪk], for example). This is because the elemental unit of representation of these models is the segment (or onsets and rimes for the onset-rime model). If two patterns are unattested, these models are silent with regard to the predicted acceptability difference between them. Note that there might still be differences in predictions between two nonce words having different unattested sequences but that difference would be not be motivated by the unattested sequences themselves. For example, the phonotactic probability calculator makes different predictions for [bnɪk] and [bzɪk] based on the difference in unigram frequencies $F(\mathrm{n}_2)$ and $F(\mathrm{z}_2)$ and bigram frequencies $F(\mathrm{n}_2\text{ɪ}_3)$ and $F(\mathrm{z}_2\text{ɪ}_3)$ but $F(\mathrm{b}_1\mathrm{n}_2)$ and $F(\mathrm{b}_1\mathrm{z}_2)$ bigrams themselves both have a probability of 0.

Recent phonotactic probability models (Hayes and Wilson 2008, Albright 2009) have combined the phonotactic probability approach to predicting speaker judgements with a sub-segmental distinctive feature and natural class representation. The key intuition is that although onsets [bz] and [bn] are both unattested, we do have frequency information about the constituent natural class sequences. For example, we might show that [-sonorant][+sonorant] sequences occur in the lexicon of English onsets but that [-sonorant][+strident] sequences do not and infer that speakers would prefer [bn] over [bz].

The challenge of this approach is that every segment represented as a distinctive feature bundle belongs to a number of natural classes, so for a given sequence, the number of possible natural class representations is $C_1 \times C_2 \times ... \times C_n$ where C_1 is the number of natural classes containing the first segment, C_2 is the number of natural classes containing the second segment, etc. Selecting the most predictive natural class sequence from that set is a non-trivial problem. For example, consider again the onsets [bz] and [bn]: both of these patterns could also be described with the natural class sequence [+voice][+voice] but this choice of generalization is unhelpful because it predicts them to be equally acceptable.

Albright, 2009 For Albright, the goal is to select a natural class sequence that reflects both the frequency of that natural class sequence in the lexicon and the specificity of each natural class to the segment it represents. The probability of a segmental sequence ab where $a \in \text{Class}A$ and $b \in \text{Class}B$ is shown in equation 2.7:

$$\text{Probability of segmental sequence } [ab] \text{ as } [\text{class } A][\text{Class } B] =$$

$$\frac{\text{occurrences of } [\text{class } A][\text{class } B] \text{ in the corpus}}{\text{total number of biphones}} \times P(a|A) \times P(b|B) \tag{2.7}$$

To take advantage of the success shown by segmental models (such as the phonotactic probability calculator) for attested sequences, Albright includes the fully specified distinctive feature bundle as a possible natural class for each segment. For example, [t] is a member of a number of natural classes including hidden states (natural classes)

outputs (segments)

$$\begin{tikzcd} X_2 \arrow[r, "t"] \arrow[d, "e"'] & X_1 \arrow[d, "e"'] \\ y_2 & y_1 \end{tikzcd}$$

Figure 2.2: Graphical representation of a bigram HMM

the fully specified bundle [-continuant][+coronal][-voice] of which it is the unique member.

Equation 2.8 shows the probability of the very general representation [+consonantal] [+consonantal] for the sequence [tl]. Note that in Albright's example for English data, there are 20 segments with the feature [+consonantal] so the probability of the representation is multiplied by 1/20 for [t] and 1/20 for [l].

$$\text{Probability of sequence [tl] as [+consonantal][+consonantal]} =$$

$$\frac{\text{occurrences of [+consonantal][+consonantal] in the corpus}}{\text{total number of biphones}} \times \frac{1}{20} \times \frac{1}{20} \tag{2.8}$$

Compare this to equation 2.9 showing the probability of the more specific [-voice, -continuant][l] representation where there are only three members in the class [-voice, -continuant] and [l] is the only member of its unique class.

$$\text{Probability of sequence [tl] as [-voice,-continuant][l]} =$$

$$\frac{\text{occurrences of [-voice,-continuant][l] in the corpus}}{\text{total number of biphones}} \times \frac{1}{3} \times \frac{1}{1} \tag{2.9}$$

More formally, Albright's model appears to be an implementation of a bigram Hidden Markov Model (HMM, figure 2.2) where the segments are outputs generated by the (hidden) natural class states. $P(X_1 = A, y_1 = a, X_2 = B, y_2 = b)$, the joint probability of a segment sequence output and a "hidden" natural class sequence is the product of the transitional probabilities from each state (natural class) to the next and the emission probabilities from states (natural classes) to outputs (segments) as shown in equation 2.10. Note that this is equivalent to equation 2.7.

$$
\begin{aligned}
&P(X_1 = A, y_1 = a, X_2 = B, y_2 = b) \\
&= P(A) \times P(B|A) \times P(a|A) \times P(b|B) \\
&= P(AB) \times P(a|A) \times P(b|B)
\end{aligned}
\tag{2.10}
$$

However, Albright's model cannot be described fully as an HMM because of the way in which bigrams are combined to create longer sequences. In a bigram HMM, the probability of a natural class sequence ABC instantiated as the segment sequence [abc], is $P(A) \times P(B|A) \times P(B|C) \times P(a|A) \times P(b|B) \times P(c|C)$, but in Albright's model, the segment [b] may be assigned to one natural class in for the [ab] bigram and to a different one in the [bc] bigram. For example, in the sequence [stl], the most probable representation for the [st] bigram might be [+sibilant, -voice][-voice, -continuant] and the most probable representation for the [tl] representation might be [-sonorant][+sonorant]. In that case, combining the two requires two different states for [t] and this is inconsistent with an HMM formulation.

Albright compared the performance of several models on the same dataset and showed that although his own model performs best overall in predicting speaker judgements involving both attested and unattested sequences, it performs less well in predicting judgements for nonce words with attested sequences than a simple bigram probability segmental model (such as the phonotactic probability calculator), and less well in predicting judgements for nonce words with unattested sequences than the Maximum Entropy (Maxent) phonotactic learner (Hayes and Wilson, 2008) that slightly predates it.

2.5.1 The Hayes and Wilson Maxent Phonotactic learner

The Maxent learner is considerably more sophisticated in its concept and design than any other model. Conceptually, the Maxent learner is a generative model in the traditional sense because it is designed to learn, from positive evidence only, a Maximum Entropy weighted grammar that describes the systematic gaps in the lexicon. Although the rime-onset model, Phonotactic Probability Calculator and Albright model can be conceptualized as grammars where the rules are probability weighted sequences (in either onset-rime, segmental or natural class representation), the Maxent learner is explicitly designed to account only for gradient speaker judgements of phonotactic patterns that are unattested (or very under-represented) in the lexicon.

Table 2.1: Sample of Maxent learner grammar for English onsets

constraint	penalty weight
*[+sonorant,+dorsal]	5.64
*[+continuant,+voice,−anterior]	3.28

Description: A trained model consists of a Maximum Entropy weighted set of constraints that can be used to assign speaker rating predictions to real words or nonce forms. As the model is designed to penalize under-represented patterns, it assigns only positive (penalty) weights.[8]. A part of the grammar for English onsets, a set of constraints and their associated weights,[9] is shown in table 2.1. The *score* of a word x is the sum of the penalty weights that apply to it (equation 2.11). In this example, the model has learned that the dorsal sonorant [ŋ]

[8]This is not a general characteristic of Maxent models which can be designed to assign both positive (penalty) and negative (goodness) weights.

[9]From Hayes and Wilson 2008.

and the coronal fricative [ʒ] do not occur in English onsets and assigns them both a penalty weight. The model predicts that, all else being equal, speakers would prefer words beginning with any segment other than [ŋ] or [ʒ], and that a word beginning with [ʒ] (score 3.28) would be preferred to a word beginning with [ŋ] (score 5.64). As these constraints are cumulative, both of these would be preferred to words with the onset [ʒŋ] (score 5.62 + 3.28) or [ŋŋ] (score 5.62 + 5.62).

$$\begin{aligned} &\text{Score }(x) = \sum_{i=1}^{n} \mathrm{w}_i \mathrm{C}_i(x) \\ &\mathrm{w}_i : \text{ weight of the } i\text{th constraint} \\ &\mathrm{C}_i(x) : \text{ number of times that the phonological form } x \\ &\ \text{violates a constraint in the grammar} \end{aligned} \tag{2.11}$$

The model learns from two sources, (a) an inventory of the segments that occur in the language expressed as distinctive feature bundles and (b) dictionary data (for example, the list of onsets that appear in a dictionary of English and their type frequencies). Before further describing the model, it's important to clarify vocabulary distinguishing the three contexts in the model where phonotactic patterns are described in terms of natural classes and natural class sequences.

- *restrictions* are statistical sound patterns present in the lexicon (learning data).
- *phonological forms* are members of the set of natural class sequences that could potentially be picked to incorporate into the grammar as constraints.
- *constraints* are the restrictive phonological rules in the grammar —the set of phonological forms which are penalized.

In a preliminary phase, the model generates the complete list of natural classes for the given segment inventory necessary to represent the learning data (lexicon) and creates all possible sequences of those natural classes. Note that the word boundary condition [#] is treated as a natural class. The maximal length of these phonological forms is a parameter of the model. For example, if there were just 2 natural classes [a][b] and the word boundary condition [#], and the maximal sequence length was set to 3[10] then the set of phonological forms would be :

- All sequences of length 1: [a], [b], [#]
- All sequences of length 2: [a][a], [a][b], [a][#], [b][b], [b][a],[b][#], [#][#], [#][a], [#][b]
- All sequences of length 3: [a][a][a], [b][b][b], [#][#][#], [a][a][b], [a][a][#], [a][b][a], [a][#][a], [a][b][b], [a][#][#]... [#][#][#]

These phonological forms correspond to possible constraints[11]. For example, the sequence [+voice][+voice] is one of the phonological forms generated by the learner for English onsets. If the learning data shows evidence that adjacent voiced segments are less frequent than would be expected, all else being equal, then *[+voice][+voice] is a potential constraint and may be incorporated into the grammar. The size of this set of phonological forms is $C^n + C^{n-1}... + C$ where C is the number of natural classes and n is the maximal length of the forms.

The learner selects a model, a grammar that describes the training data, by iterating over the following two phases:

- Assigning weights: the constraints in the grammar are assigned penalty weights. When the grammar is complete, it can be used to predict speaker judgements of new words.
- Constraint selection: the set of phonological forms is evaluated to determine which constraint should next be added into the grammar

[10]The length of a constraint is defined here as the number of sounds that it spans; for a constraint over a single segment, the length is one, for a sequence over two segments, the length is 2 etc.

[11]Presumably, redundant or useless phonological forms such as [#][#][#] are either not generated or eliminated from the set. However, this is not discussed in Hayes and Wilson's description of their model.

Assigning weights The goal of the model is to learn a grammar that maximizes $P_\theta(D)$, the probability of the data given the constraint weights θ.[12] $P_\theta(D)$ is the product of the probability of every word in the data, given the weights (equation 2.12).

$$P_\theta(D) = \prod_{x \in D} P_\theta(x) \tag{2.12}$$

Each time that a constraint is added to the grammar, the entire grammar is retrained and a penalty weight is assigned to each constraint θ such that $P_\theta(D)$ is maximized.[13] As there is no single-step method for determining the best weight for each constraint, each weight must be incrementally modified until the probability of the data can no longer be improved.

The weight adjustment process can be visualized as climbing a hill where the constraint weights define a multi-dimensional surface. At each incremental iteration, the constraint weights are adjusted until they reach the highest point on the surface.

Maximizing $P_\theta(D)$ requires determining, at each iteration, how each constraint weight should be adjusted to improve it. The solution to this is to consider, for each constraint C_i in the grammar, the difference between the number of Observed violations of the constraint and the number of Expected violations given the current grammar, $O[C_i] - E_\theta[C_i]$. This number corresponds to a (positive) slope that indicates the peak of the multi-dimensional surface (Della Pietra et al., 1997). The weight for a given constraint is then adjusted according to the slope at that point.

[12]This is a simplification of the weight assignment. To avoid over-fitting in Maxent models, it is usual to penalize large weights (Duda et al., 2001). The goal is therefore to maximize the penalized log-likelihood of the data rather than the raw log-likelihood. Hayes and Wilson use a gaussian prior on the weights with μ set to 0 and σ set to 1 based on the discussion in Goldwater and Johnson (2003). As the gaussian is an exponential of a sum of squares, the equation that is being optimized has two summation terms, one of which is the likelihood and the other the sum of the squares of the weights.

[13]For reasons related to the difficulties of computing very small numbers, and as the weights that maximize $P_\theta(D)$ are also those that maximize log $P_\theta(D)$, log $P_\theta(D)$ is used rather than $P_\theta(D)$ itself.

$$
\begin{aligned}
&E_\theta[C_i] = \sum_{x\in\Omega} P_\theta(x) C_i(x),\\
&\textit{where}\\
&P_\theta(x) \text{is the probability of the representation } x\\
&C_i(x) \text{ is the number of times } x \text{ violates } C_i \text{ and}\\
&\sum_{x\in\Omega} \text{is the summation over all } x \text{ in } \Omega
\end{aligned}
\tag{2.13}
$$

$O[C_i]$ is simply the count of Observed violations of $[C_i]$ in the training data. $E_\theta[C_i]$ (equation 2.13), however, poses computational issues because it is the product, for every word x in Ω (the set of all segment combinations shown in figure 2.1), of $\mathrm{P}_\theta(x)$, the probability of the word x, and the number of times that word violates constraint i.

$$
\begin{aligned}
P_\theta(x) &= \frac{\text{Maxent value}(x)}{Z}\\
Z &= \text{a normalizing factor}
\end{aligned}
\tag{2.14}
$$

$\mathrm{P}_\theta(x)$ (equation 2.14) is the Maxent value x shown in equation 2.15[14] divided by a normalization constant Z to ensure that the distribution is proper.

$$
\begin{aligned}
&\text{Maxent Value } (x) = \exp\ (-(\sum_{i=1}^{n} \mathrm{w}_i \mathrm{C}_i(x))\\
&\mathrm{w}_i : \text{ weight of the } i\text{th constraint}\\
&\mathrm{C}_i(x) : \text{ number of times that the phonological form } x\\
&\ \text{violates the } i\text{th constraint in the grammar}
\end{aligned}
\tag{2.15}
$$

Even the normalization constant Z is computationally non-trivial because it is the sum of the Maxent values of all the words y in Ω (equation 2.16).

[14]Note that the Maxent value is the negative score exponentiated (equation 2.11).

$$Z = \sum_{y \in \Omega} \text{Maxent value}(y) \tag{2.16}$$

As Ω is an infinite set, it is common practice to consider only strings no longer than the longest word in the training data. Even so, the number of strings to be considered is exponentially large. For English onsets, which are never longer than 3 segments, and a segment inventory of 24 consonants, the size of the set of possible segmental sequences is $24^3 + 24^2 + 24$.

Hayes and Wilson solve this problem by using a finite state machine (FSM) to find the Expected values of constraints. Each constraint is represented as a weighted finite-state acceptor and these are then combined into an FSM where each through-path corresponds to a string and the list of constraints (with their weights) that it violates. The $E_\theta[C_i]$ values are obtained by summing through all the paths of the FSM (Eisner 2001, 2002).

All constraints are initialized with a weight of 1. At each iteration and for each constraint, the algorithm computes $O[C_i] - E_\theta[C_i]$ and the weight of each constraint is then adjusted as a function of the slope at that point. The process reiterates until the slope for each constraint is close to 0. Once the grammar is complete, according to some criteria[15], it can be used to evaluate both the words in the lexicon (training data) and novel words. As noted previously, the predicted acceptability of a word x is its *score*, the weighted sum of its constraint violations (equation 2.11). As the weights penalize under-represented sequences, higher Maxent values predict lower speaker acceptability.

Constraint selection The goal of constraint selection is to add the next most useful constraint to the grammar. The model searches the set of phonological forms and determines which one describes the most significantly under-represented

[15]The criteria for completion of the the model is determined by the researcher. Possible options are to stop the learning when the grammar reaches a specified size or when there are no more constraints with O/E values below a certain threshold. Although it is theoretically possible to let the model acquire all the relevant constraints with O/E values < 1, the current implementation does not allow this).

sequence, given the current state of the grammar. The selected form is then added to the grammar as a constraint.

Searching the set of phonological forms is not trivial. The model must compute, for each phonological form, $O[C_i]/E_\theta[C_i]$, the ratio between the Observed number of violations of the constraint and the Expected number of violations, given the current state of the grammar. Consider the phonological form [+word boundary][+consonantal, +continuant, +coronal, +anterior, -strident] which describes a generalization over [#][ð] and [#][θ]. In terms of type (rather than token) frequency, these are rare onsets in English so the phonological form [+word boundary] [+consonantal, +continuant, +coronal, +anterior, -strident] may be incorporated into the grammar as the constraint *[+word boundary][+consonantal, +continuant, +coronal, +anterior, -strident], and all words that contain that sequence are assigned a penalty weight. Now, the number of times that the onset [#ðr] and [#θr] would be expected to occur, given that they are subject to a constraint in the current grammar, is reduced. As the Expected value is smaller, $O[C_i]/E_\theta[C_i]$ is higher so the phonological form [+word boundary] [+consonantal, +continuant, +coronal, +anterior, -strident][+retroflex] is less likely to be picked as a constraint.

To summarize the difficulty, to select from the set of phonological forms the one that has the lowest $O[C_i]/E_\theta[C_i]$ value given the current grammar, the model must compute $E_\theta[C_i]$ for every word of Ω and this, *for every member (potential constraint) of the set of phonological forms*. Computationally, this is orders of magnitude more intensive than computing $E_\theta[C_i]$ values in the grammar weighting phase because (in the Hayes and Wilson simulations at least) even a completed grammar contains less that 100 constraints while there are up to 100 million natural class sequences in the set of phonological forms.

Hayes and Wilson determined that building a finite state machine to evaluate the Expected value (equation 2.13) of the phonological forms, the method used for constraint weighting, would be impractically slow for constraint selection (given the size of the set of phonological forms). Instead, they elect to use Monte Carlo sampling to estimate the Expected value for each constraint as the average number of times that a constraint is violated in the sample is proportional to the number of times it is violated in the whole set.

Constraint selection proceeds over successive $O[C_i]/E_\theta[C_i]$ intervals starting with the interval 0 to .01, then 0 to .1, 0 to .2 etc. Within a given $O[C_i]/E_\theta[C_i]$ interval, the model is equipped with a heuristic that selects the most general constraint; short constraints are preferred over long ones and constraints with few natural features are preferred over those with many.

Hayes and Wilson 2008: results The Maxent learner acquires a constraint grammar learned from the input, iterating over constraint selection and constraint weighting until the researcher-determined criterion for cessation of learning is reached. In the case of the English onset simulations of Hayes and Wilson, the grammar was deemed to be complete when there were no more phonological forms available with $O[C_i]/E_\theta[C_i]$ values below the threshold of .3.

The acquired grammar is then used to assign penalty weights both to the lexicon (training data) and to novel words. In the case of English onsets, Hayes and Wilson compared the weights assigned by the grammar to the experimental results of Scholes (1966). Scholes presented 33 7th grade English-speaking schoolchildren with 66 nonce forms with onsets of varying legality and asked them to judge whether they were acceptable words of English or not. These binary judgements were averaged for each word to produce a gradient measure of acceptability[16].

The predictions of the Maxent learner were shown to be very strongly correlated to the gradient measures obtained by Scholes (Pearson's correlation: r = .946). Hayes and Wilson also compared the predictions of their model with implementations of Coleman and Pierrehumbert's onset-rime model (1997), a segmental n-gram model, the Bailey and Hahn GNM model (2001) and a hand-written grammar (Clements and Keyser, 1983). The correlation between the Scholes experimental results and the Hayes and Wilson predictions was greater than for any other model.

The predictions of the Maxent learner for unattested sequences are also generally in line with experimental results other than Scholes that rank the ac-

[16]Frisch et al. (2000) show that there is a high correlation between pooled binary judgements and judgements that directly assign a gradient score such as King and Rose (2003).

ceptability of English onsets (Albright, 2009; Berent et al., 2007), However, there are indications that the predictions may not always be accurate. The Maxent learner assigns weights of 7.925 and 4.396 to [bw] and [dl] respectively, predicting that [dl] is preferred over [bw]. However, Moreton (2002) showed in a perception experiment that speakers have a stronger perceptual bias against [dl] than [bw] indicating that in a word judgement task, speakers would likely prefer [bw] over [dl].

One somewhat controversial aspect of the Maxent learner's performance on English onsets is that very few weights are assigned to attested onsets. This behaviour reflects the traditional generative assumption that gradient speaker judgements for nonce words with attested sequences do not need to be accounted for by the phonotactic grammar (which only describes systematic gaps). The model is restricted in its predictions for nonce words with attested sequences by the interaction of two mechanisms:

- The model assigns only positive (penalty) weights. This is because assigning negative (goodness) weights would make the prediction that the repetition of a favoured pattern would increase acceptability. For example, if the pattern [ta] is assigned a goodness weight, then [tatata] should be even better.[17]
- The $O[C_i]/E_\theta[C_i]$ threshold for stopping learning is relatively low (.3 in the simulations for English), corresponding to strongly under-represented patterns.

While it is true that the fit between the Maxent learner predictions and the Scholes experimental data, which contains some nonce words with attested onsets, is high, this is a questionable result, both with regards to the model predictions and the experimental data. The fact that nonce words with attested onsets were almost always assigned a perfect rating in the Scholes judgement task is most likely an artefact of the experimental design. Scholes used both attested and unattested onset sequence stimuli and speakers were asked to assign binary

[17] With thanks to Bruce Hayes (pc) for this example.

acceptable-unacceptable judgements. Albright (2011b) points out that in such conditions, asking speakers whether or not a nonce form is a possible word of English may have diminished speaker sensitivity to the fine-grained difference between attested sequences because the difference in acceptability between attested sequences is much smaller than the difference between attested and unattested sequences.

However, a grammar that assigns positive weights only is not incompatible with assigning differential weights to words with attested sequences. The key concept is that if a pattern is over-represented, some other pattern or set of patterns must be under-represented.

Example 2.17 shows a very simple demonstration of how the model could predict a difference in acceptability between nonce forms with attested onsets [k] and [g] (2764 and 537 occurrences respectively, in the CMU online pronouncing dictionary). Given that [k] and [g] are both over-represented and [k] is more over-represented than [g], it must be the case that the complement set of all the segments that are not k ([ˆk][18]) is more under-represented than the complement set of g ([ˆg]). If a grammar is acquired such that, given the weights, there are no under-represented sequences, [ˆg], which includes [k], receives a smaller penalty weight than [ˆk], which includes [g], and [k] is therefore predicted to be preferred by speakers over [g].

[18]The diacritic ˆ designates the complement class, the set of all segments not defined by the matrix. For example [ˆ-voice,+labial] is the set of all segments other than [p] and [f].

$$
\begin{aligned}
&\textit{given}:\\
&1 < \mathrm{O/E}\ [\mathrm{g}] < \mathrm{O/E}[\mathrm{k}]\\
&\textit{therefore}\\
&\mathrm{O/E}[\hat{\ }\mathrm{k}] < \mathrm{O/E}[\hat{\ }\mathrm{g}] < 1\\
\\
&\textit{a grammar is acquired such that}\\
&\textit{there are no under-represented patterns}\\
\\
&\textit{weights assigned to}\ [\hat{\ }\mathrm{k}] > [\hat{\ }\mathrm{g}]\\
&\textit{as}\ \ [\mathrm{k}] \in [\hat{\ }\mathrm{g}]\ \textit{and}\ [\mathrm{g}] \in [\hat{\ }\mathrm{k}]\\
&[\mathrm{k}]\ \ \textit{is predicted to be preferred to}\ [\mathrm{g}]
\end{aligned}
\tag{2.17}
$$

Hayes and Wilson also report simulations for other languages but they only compare their model's predictions with speaker judgements for English onsets which is a fairly simple dataset to characterize. For example, for simple onsets, all consonants except [ŋ] are allowed. For two segment sequences, a subset of obstruent-sonorant sequences are allowed and three consonant onsets can only start with [s] followed by a subset of the permitted two consonant onsets. It is not clear how well the model would fare on languages where restrictions are more complex or encompass a larger number of segments, such as vowel harmony across entire words[19].

2.5.2 Conclusion

This review of phonotactic models of speaker judgements shows the evolution of the field towards models linking the probability of a sub-segmental sequence to its acceptability for speakers. With regard to English onsets, at least, bigram

[19] Although Hayes and Wilson did train the model on data from Shona, a language with vowel harmony, they did not compare the predictions of their model to speaker judgements.

models perform well in predicting nonce words with attested sequences, the Maxent learner reliably predicts speaker judgements for nonce English words with unattested sequences, and the Albright model predicts judgements for both attested and unattested sequences, though with less success in each specific domain than the other two. However, the Maxent learner stands out for its statistical wellfoundedness, its phonological representational system and its capacity to be extended in the direction of predicting attested sequences.

Most of the models presented here were compared on their performance in predicting English speaker judgements of nonce words with differing attested onsets. This is the benchmark task in the discipline because it is an easy and well-understood data set with available test data. However, there is still much work to be done for longer sequences (particularly whole words) and languages other than English. In the upcoming chapters, we will address these limitations by evaluating the performance of the Maxent learner in predicting whole word judgements for existing speaker judgement data from Amharic, a Semitic language that has phonotactic characteristics that are different from English. Based on the results of those simulations, we collect new Amharic speaker judgement data designed to clarify both the model predictions and our understanding of the frequency∼acceptability cline.

3 Overview of Amharic morphology and phonotactics

Amharic, an Ethiopian Semitic language, has characteristics that recommend it as a test case for phonotactic modeling:

- A complex and well-studied phonotactic restriction on the co-occurrence of consonants is active within Semitic verb roots, including Amharic.
- The morphology of Semitic languages makes it possible to model whole words (verbs) with a relatively small consonant inventory[1] and to keep the vowels constant across all words so that they do not impact consonant phonotactics.
- The Amharic lexicon of verb roots is not homogeneous; there are many roots that do not conform to the canonical triliteral length. This allows us to evaluate whether the generalizations that are true for canonical roots also hold for other root shapes.
- There is appropriate experimental speaker judgement data against which to test the model predictions (King and Rose, 2003).

The goal of this chapter is to provide an overview of the morphology and phonotactics of Amharic, demonstrating that a Semitic language, and particularly Amharic, makes an appropriate and interesting test case for the Maxent learner. The first section describes the root-and-template morphology that is common to Semitic languages and the shape of the lexicon of verb roots, in terms of the variety of

[1] The Maxent learner is impractically slow when the number of phonological forms, a function of the segment inventory, is in excess of 100 million.

root shapes and their frequencies. We review the frequency of verbs containing identical consonants and the theoretical analysis for their distribution. The following section discusses the Obligatory Contour Principle as applied to place of articulation, a restriction over homorganic consonants within Semitic verb roots and provides a dictionary analysis of the co-occurrence of homorganic consonants in the lexicon of Amharic verb roots. The chapter concludes with a description of the experimental word rating data used to test the model.

The dictionary analysis of Amharic presented in this chapter is based on an electronic database of 4243 verb roots extracted from the Amharic-English dictionary (Kane, 1990) using only verb roots that could be conjugated (not the fixed forms with auxiliaries such as 'alə').

3.1 Morphology

Amharic, like other Semitic languages, has root-and-template morphology (Bender and Fulass, 1978; Leslau, 1995). In the root-and-template system, morphemes are not simply added to a stem in a process of linear affixation but rather morphological derivations are accomplished through templates combining a set of root consonants with vowels and other consonants according to particular shapes. In this way, a typical 3-consonant, or *triliteral*, root (represented hereafter as $C_1C_2C_3$) can, depending on the template selected, form a series of semantically related nouns, verbs or adjectives[2]. Table 3.1 shows a group of semantically related words for the Amharic root /lwt'/, which has the lexical semantics '*change*' and the root /dfn/ which has the lexical semantics '*fill in*'. The key observation here is that there is a relationship between the consonants of the verb root that is independent of intervening material. A difference in judgement between two verbs conjugated in the same paradigmatic form should therefore be dependent only on their respective roots. The advantage is that we can model judgements

[2]Not all researchers acknowledge the root as a morphemic entity (Bat-El, 1994; Bat-El, 2003; Benmamoun, 1999; Ratcliffe, 1997; Ussishkin, 2000). However, this research is neutral with regard to the actual status of the root.

Table 3.1: Two example of verb roots and words derived from them

Root /lwt'/	gloss *'change'* (Type B)
ləwwət-t'ə	'he changed' (tr.)
jɨ-ləwwɨt'-all	'he is changing' (tr.)
ləwwɨt'	'change!' (tr.)
ləwawwət'-ə	'he changed completely'
ləwt'	'change, alteration'
lɨwwɨt'	'changed'
mə-lləwawət'	'interchange'
lɨwwatʃ'	'thing given in return'

Root /dfn/	gloss *'fill in'* (Type A)
dəffənə	'he filled in (hole), blocked up, buried'
jɨ-dəfn-all	'he is filling in'
dɨfən	'fill in!'
dəfaffənə	'he buried by digging holes here and there'
dɨfan	'stopped up, plugged, buried'
dəfaɲ	'one who plugs up or fills in a hole'
dəfəna	'plugging up, act of covering over'

for whole verbs by keeping the vowels identical (for example, all [ə]), and altering the consonants of the root.

We are aware that this is a simplifying assumption as there may be frequency effects related to the number of derived words available for a given verb root. The phonotactic characteristics of the non-verb root lexicon may also have an influence on the judgements of verbs[3].

In Amharic, triliteral verb roots (which are the most common at 44% of the lexicon of verb roots) are divided into three lexical classes characterized primarily

[3]We are undertaking a statistical investigation of non-verb roots but this requires a hand-annotated transcription of the lexicon that is not yet available.

Table 3.2: Examples of gemination in different lexical classes

Type	Type A	Type B	Type C	4-consonantal
gloss	*'break'*	*'want'*	*'take prisoner'*	*'mix, confuse'*
perfective[4]	səbbərə	fəlləgə	marrəkə	dəballək'-ə
imperfective	jɨ-səbɨr	jɨ-fəllɨg	jɨ-marrɨk	jɨ-dəballɨk'-ə
jussive	jɨ-sbər	jɨ-fəllɨg	jɨ-mark	jɨ-dəbalk'

by the occurrence of gemination (or lengthening) of the middle consonant (C_2) (see table 3.2).[5] In Type A, gemination occurs only in the perfective form, in Type B, gemination occurs throughout the paradigm and in Type C, gemination occurs in the perfective and imperfective forms. Types B and C are further characterized by segmental regularities: for Type B, palatal or labio-velar consonants are common in C_1 position and for Type C, the first vowel (between C_1 and C_2) is [a] for the standard aspectual forms (perfective, imperfective and jussive, but other moods and aspects are also possible). This is a lexical distinction, so some homophonous verb roots may occur in more than one type without necessarily being semantically related. For example, the Type A root /blg/ ('bəlləgə'), means '*rain in the season of small rains; send out shoots*' while the Type C root /blg/ ('balləgə'), means '*misbehave, be naughty, be rude*'. The frequency of the three classes is shown in table 3.3.

Table 3.3: Frequency of verb types

Type	Type A	Type B	Type C
count	770	627	210

Amharic also has a large number of roots with 4 ($C_1C_2C_3C_4$) or even 5 consonants. With respect to gemination, they are similar to Type C roots, as the

[5]Triliteral verb roots where the last two consonants are identical may belong to any of the three lexical classes and are included in the counts of table 3.2. Verb roots with identical consonants will be discussed in detail in section 3.2.

penultimate consonant (C_3) is geminate in the perfective and imperfective forms. For example, /bdrg/ surfaces as `bədərrəgə', (‘*to rise quickly*’) and /dnbʃ/ surfaces as `dənəbbəʃə' (‘*go bad, be spoiled*’).

Note that lexical class of Type A and Type B triliteral roots cannot be determined based on the citation form (3ms perfective) because both classes have gemination in the perfective aspect.[6] This is a useful characteristic in designing word judgement experiments as nonce verbs presented in citation form could be either Type A or B which are the most frequent overall (see table 3.3). The presence of a specific consonant might further bias speakers towards one class of those classes over the other. For example, a nonce root with an initial palatal is likely to be classified as Type B. However, nonce consonantal roots are unlikely to be interpreted as type C, which has a distinctive vowel.

Table 3.4: Examples of weak verbs

missing consonant	C_1	C_3	C_2
gloss	*‘pass’*	*‘measure’*	*‘kiss’*
perfective (class)	alləf-ə (A)	ləkka (B)	sam-ə (hollow)
gloss	*‘think’*	*‘hobble’*	*‘send’*
perfective (class)	assəb-ə (B)	gadda (C)	lak-ə (hollow)

Like other Semitic languages, Amharic has some bi-consonantal roots, labeled ‘weak roots’. However, synchronic and diachronic evidence indicates that a third consonant was once present (or depending on the analysis, the third consonant is still present in the underlying representation but absent from the surface realization (SR)). Table 3.4 shows examples of how a missing root consonant is synchonically marked through the location of the geminate consonant. When the C_1 consonant is missing, the geminate appears as the left-most root consonant. Likewise, when the C_3 consonant is missing, the geminate appears as the right-most root consonant. Finally, when the middle (and therefore geminate) consonant

[6]Type C has the characteristic [a] vowel, so while its perfective form resembles the others in terms of gemination, the [a] vowel creates a distinction.

is missing, there is no geminate consonant in the root. In each case, a missing guttural consonant is replaced by [a] (as shown in the examples). Although a consonant may be missing from any position, roots with missing C_2 are called *hollow* and gemination does not occur. Note that because of the absence of gemination, the lexical class of hollow roots cannot be determined 1) because the gemination cue is absent and 2) the presence of [a], a cue for Type C, could also mark a vestigial guttural. There are also triliteral weak forms similarly related to historically quadriliteral roots. Weak roots are not a rare occurrence in the lexicon of Amharic, 17% of Amharic verb roots show evidence of one or more missing consonants.

Table 3.5: Examples of w-medial roots

gloss	*'run'*	*'think'*	*'be restless'*
perfective (class)	rot'-ə (A)	ləwwət'-ə (B)	nawwəz-ə (C)

The glides (/j/ or /w/) may also be missing from some roots. For type A roots, a round vowel appears instead of medial [w] (table 3.5). For [j] in medial position of type A roots, a modified vowel appears (mid in perfective and imperfective, high in jussive). However, if the first consonant is a coronal, there is no vowel change and the coronal consonant becomes palatalized as shown in table 3.6 (Hudson, 1979). Type C roots are not included in the table because there is only one type C root with [j] in medial position, where it appears as a geminate.

Table 3.6: Examples of j-medial roots

gloss	*'go'*	*'sell'*	*'ask, visit'*
perfective (class)	hed-ə (A)	ʃət'-ə (A)	t'əjjək'-ə (B)

3.2 The Obligatory Contour Principle

There is a strong cross-linguistic trend to avoid the co-occurrence of identical material within a word. This trend, formalized as the Obligatory Contour Principle (OCP) was first used to account for the behaviour of tone (Leben, 1973) and later expanded (McCarthy, 1979, 1981) to state that adjacent identical material cannot co-occur in Underlying Representation (UR) or through derivation. Further research (McCarthy, 1986; Myers, 1997), argues that the OCP functions as a constraint operative throughout the derivation.

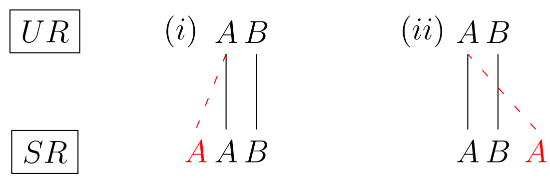


Figure 3.1: Prohibited derivations

OCP accounts for the distribution of identical consonants in Semitic verb roots. The theory assumes that:

i. material in UR can be repeated in SR through a process of spreading or copy

ii. consonants map to template positions one-to-one from left to right; a remaining template slot is filled by spreading rightwards.

iii. association lines between phonemes in UR and segments in SR cannot cross.

Under the theory, verb roots with identical left edge and non-adjacent consonants in AAB and ABA patterns are underlying:

- AAB cannot be derived from AB because this would require right to left association or edge-in-association (Buckley, 1990; Yip, 1988) followed by leftward spreading (or copy) of the A as shown in figure 3.1(i).

- ABA cannot be derived from AB because spreading would require crossing the association line connecting the UR and SR of B as shown in figure 3.1(ii) and copy is licit only from the rightmost consonant of the stem and to the adjacent position.

As both AAB and ABA contain identical material, they represent violations of OCP and this accounts for their rarity (note that to account for ABA forms, the analysis requires a formalization of OCP that prohibits identical material in any position rather than only in adjacent position)[7].

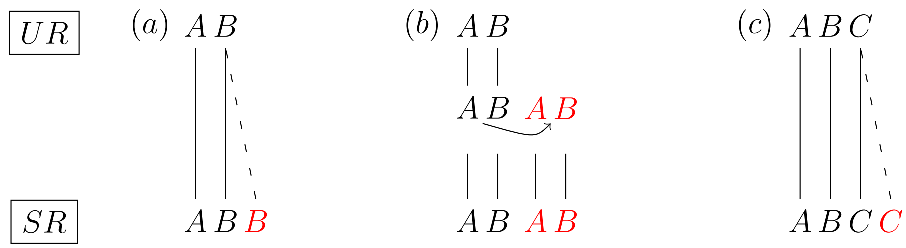


Figure 3.2: Licensed derivations

Roots with patterns of identical consonants such as ABB, ABAB, ABCC are more frequent:

- ABB roots, called *doubled verbs*, are common in Semitic languages. It is assumed that the underlying root is AB with a spreading or copying of the B to fill the tri-consonantal template (see figure 3.2 a).

- ABCC and ABAB type patterns are assumed to arise from 3-consonantal and 2-consonantal roots in UR (ABC and AB, respectively). ABAB forms are assumed to arise not by spreading (given the prohibition on crossing association lines) but through a process of root copy (figure 3.2 b) whereas ABCC forms can be derived similarly to ABB forms (figure 3.2 c).

For terminological convenience, I will refer to these roots with identical consonants that can be analysed as rightward copy or spreading as *reduplicative*, and those that have identical consonants in other positions as non-reduplicative, even though we do not have a particular theoretical stance on the issue of whether the repetition is due to a formal process of reduplication (copy) or spreading.

Table 3.7 summarizes reduplicated verb root patterns in Amharic. These patterns are also attested in Tigre (Raz, 1983; Rose, 2003a,b), Tigrinya (Buckley,

[7]See Gafos (1998) for alternate views on the derivation of these forms

Table 3.7: Semantics of reduplicative verb roots

Type	Status	Semantic notion of repetition or intensity
ABB	phonological	No
ABAB	lexical	Yes
ABCBC ABCDCD	lexical or morphological	Yes
ABCC	lexical	Yes
ABBC ABCCD	morphological	Yes

1990; Leslau, 1941), Harari (Leslau, 1958), Chaha (Prunet and Petros, 1996; Rose, 2000, 2003b), Inor (Prunet and Chamora, 2001) and Muher (Rose, 2003b). A number of them (such as ABB, ABAB, ABCBC) are also attested in other Semitic languages. Note that with the exception of doubled verbs (ABB), there may be a relationship between reduplication and a semantic notion of repetition.

ABAB and ABCC roots are not usually derived from some unreduplicated form that exists independently as a surface form (ABB[8] and ABC, respectively) and are therefore considered to be cases of lexical reduplication. Both of these forms may have a semantic connotation of repetition, intensity or local movement and ABCC roots may also encode a semantic notion of physical defect. ABCBC patterns are assumed to be morphologically derived from ABC forms although many do not have a related ABC form. In verbal derivations, these forms usually have a tə- or a- prefix and may have a semantic connotation of repeated or intense

[8] Underlyingly AB

Table 3.8: Examples of reduplicative verb roots encoding a semantic notion of repetition or intensity

pattern	examples	gloss
ABAB	**təbəttəbə**	‘twist, tie, wind around’
	fədəffədə	‘be in excess, exceed, run over’
ABCBC	**tə-lmət’əmmət’ə**	‘be chewed continually at length’
	tə-mləgəlləgə	‘become slippery, slimy, viscous’
ABCC	**ʃəbəllələ**	‘wrap up, roll up’
	fərəttətə	‘swell up’

action. Some examples of reduplicative verb roots that have a related semantic connotation are shown in table 3.8.

Note that although ABB and ABCC forms share a characteristic repetition of the final consonant, ABB forms do not share the semantic connotation of ABCC forms. ABCBC and ABAB forms, which both have copies of two consonants, are called ‘bi-consonantal reduplication’ by Unseth (2002) who documents their usage in Amharic and other languages.

Table 3.9: Examples of verbs in frequentative aspect

1ms perfective	gloss	frequentative	gloss
dəggəmə	‘repeat’	**dəgaggəmə**	‘review, repeat again and again’
t’ərrəgə	‘sweep, clear, wipe clean’	**t’ərarrəgə**	‘sweep up everything, clean completely’
gərəbbədə	‘to open a door wide’	**gərəbabbədə**	‘to open several doors wide’

There are also cases of verb roots with identical consonants encoding intensification or repetition of a related base form arising through productive and regular conjugation[9]. For Amharic, in the frequentative aspect (illustrated in table 3.9), the penultimate root consonant is copied, producing ABBC and ABCCD forms and the vowel 'a' appears between the consonants (Leslau 1939, Rose 2000, 2003, Schluter 2008). In a related Semitic language, Hebrew, the (rare) pəʕalaʕal binyan, triconsonantal roots are copied into a 'ABCBC' template that confers some notion of semantic intensification (root: /sħr/, first binyan form: [saːħar] 'to go about', pəʕalaʕal binyan form: [səħarħar] 'to palpitate') (McCarthy, 1981).

These cases are different from those mentioned above because the identical consonants involve productive infixation or productive templatic selection derived from an existing base form.

It is important to understand the distribution of identical consonants in verb roots because the frequency of identical consonants (which are homorganic by definition) in some locations of verb roots interacts with OCP-Place, a general restriction over homorganic consonants described in the next section.

3.3 The Obligatory Contour Principle as applied to place of articulation

The Obligatory Contour Principle as applied to place of articulation (OCP-Place) is a special case of OCP. First identified as a restriction in Semitic languages (Cantineau, 1946; Greenberg, 1950), OCP-Place is based on the observation that in some languages, consonants from the same place of articulation[10] tend not to co-occur in verb roots. For example, triliteral roots such as /kgf/ and /brm/, which have two consonants from the same place of articulation (dorsal and labial, respectively) are rare.

The formulation of OCP-Place is dependent on the specific definition of place of articulation. In his analysis of Semitic languages, with special attention to 3775 Arabic roots, Greenberg (1950) defines four broad places of articulation; back consonants (dorsals, pharyngeals, laryngeals), coronal sonorants, front con-

[9]Conjugation here is different from the affixation process of Romance languages, but refers instead to the consonantal and vocalic modifications around the stem segments.

[10]In OCP terms, these roots contain *identical* material with regards to place of articulation.

sonants (coronal stops and sibilants) and labials. Greenberg provides statistical evidence that segments co-occur freely across, but not within groups, with two exceptions: velars co-occur freely with pharyngeals and laryngeals, and dental stops and fricatives tend to co-occur more frequently with each other than they do within each group (ie, dental stops co-occur rarely with other dental stops and coronal sibilants rarely with coronal sibilants).

Based on an analysis of 3330 verb roots from a dictionary of Modern Standard Arabic, McCarthy (1988) repeats Greenberg's findings and directly defines five groups of non-co-occurring consonants, noting, as did Greenberg, that coronal obstruents are restricted according to manner:

- labials: m, f, b
- coronal sonorants: l, r, n
- coronal obstruents: θ, ð, t, d, s, z, ʃ, sˤ, dˤ, tˤ, zˤ
- dorsal obstruents: k, g, q
- pharyngeals and laryngeals: ʁ, χ, h, ħ, ʕ, ʔ

For both Greenberg and McCarthy, the consonants /w/ and /y/ are excluded from the general analysis. Greenberg notes that these glides do not pattern with any group and this this is likely because /w/ and /y/ replaced /u/ and /i/ in older root forms. For McCarthy, the absence of co-occurrence of these glides is motivated not by OCP-Place but by "conspicuous phonological irregularities" elsewhere in the language.

Although first identified in Semitic languages, OCP-Place has also been studied in languages as diverse as English (Berkley, 1994), Russian (Padgett, 1995), Javanese (Mester, 1986) and Muna (Coetzee and Pater, 2006). For Semitic languages, OCP-Place is attested in Arabic (Elmedlaoui, 1995; Frisch et al., 2004; Frisch and Zawaydeh, 2001; Greenberg, 1950; McCarthy, 1979, 1981, 1988, 1994; Pierrehumbert, 1992; Yip, 1988), Hebrew (Bachra, 2001; Koskined, 1964; Kurylowicz, 1972), Tigrinya (Buckley, 1997), Amharic (Bender and Fulass, 1978; Rose and King, 2007) and Chaha (Rose and King, 2007).

	very rare	$\rightarrow$		*less rare*
POA	Dorsals	$\rightarrow$	Labials $\rightarrow$	Coronals
distance	adjacent	$\rightarrow$		non-adjacent

Figure 3.3: OCP gradiency

OCP-Place in Semitic languages is a considered to be a *gradient* restriction because it may be violated. This means that while homorganic segments do co-occur in the consonantal verb root template, they do so less often than would be expected, all else being equal.

The strength of the restriction (and therefore the possibility that verb roots with homorganic consonants are attested in the lexicon) is modulated by two factors:

- distance: homorganic consonants within the root are adjacent[11] or non-adjacent (Greenberg, 1950)
- place of articulation (POA) (Frisch et al., 2004)

The consequence of this granularity is that for triliteral roots of the form $C_1C_2C_3$, the restriction is stronger for homorganic consonants in adjacent position C_1C_2X or XC_2C_3 than in non-adjacent position C_1XC_3. With regard to POA, the restriction is stronger for dorsals or gutturals (note that gutturals do not occur in all Semitic languages) than for coronals (according to manner of articulation), with labials in an intermediate position. Figure 3.3 shows the strength of the restriction as a function of POA and violation location.

3.4 Dictionary study

As discussed in the previous section, OCP-Place is a well-studied restriction in Semitic languages generally. However, the dictionary of Amharic verb roots is

[11] Where adjacency is defined with respect to the root, not intervening vowels.

the data that is used to train our model. The statistical analysis of that data provides important information at two levels:

- precise knowledge of the shape of the lexicon is necessary to analysing the performance of the Maxent learner as a function of the representational system. For example, the number of forms with identical (and therefore homorganic) consonants affects the generalizations that can be used to describe OCP-Place.

- OCP-Place is a gradient (rather than categorical) restriction, so the evaluation of the relationship between pattern frequency and speaker ratings requires a careful evaluation of co-occurrence patterns as a function of place of articulation and location of violation.

Table 3.10: Composition of the lexicon of verb roots

root type	count	percentage
roots without identical consonants	2688	63%
roots with identical consonants	1555	37%
total	4243	

Table 3.10 shows that for Amharic, 37% of Amharic verb roots contain identical consonants. and table 3.11 shows that patterns that are compatible with a spreading (ie, ABB, ABCC) or copy (ie, ABAB, ABCBC) analysis account for a total of 1467 of the 1555 verb roots with identical consonants roots (94%). Patterns such as AAB, ABA are much rarer. Note that the lexical ABBC pattern (for which there is no related ABC form), may be historically related to the productive frequentative form, but is rare.

This summary of the shape of the lexicon of Amharic verb roots shows that a large proportion of them do not conform to the canonical triliteral pattern most

Table 3.11: Distribution of Amharic verb roots with identical consonants

pattern	count	%
ABB, ABCC, ABCDD	665	15.7%
ABAB, ABCBC, ABCDCD	802	18.9%
AAB, ABA, ABAC...	111	2.6%
ABBC	10	.23%

commonly associated with Semitic verb roots. Accordingly, we made the decision to base our analysis on the database of verb roots including weak roots, roots with four or more consonants and roots with identical consonants regardless of their assumed derivation. The reduplicative verbs pose a problem in how a model might assess them – based on their assumed underlying representation without repetition, or based on their surface realization. This is particularly problematic for lexical reduplication, since a form without reduplication does not surface.

Our decision to assess OCP-Place over the entire lexicon of verbs (including those with identical consonants) contrasts with (Frisch et al., 2004) and similar studies because we assume that all verbs in the lexicon influence phonotactics and speaker's awareness of phonotactics, and that it is not just confined to a subset of the data[12]. Note that as roots also allow for a large number of other derived forms such as nouns and verbs (see table 3.1), the lexicon of verbs actually covers a large proportion of the overall Amharic lexicon.[13]

3.4.1 Consonant inventory

According to Leslau (1995), Amharic has the inventory of 43 consonants[14] shown in table 3.12. Leslau notes that many consonants may occur with labial-

[12]Although the subset of verb roots included in the analysis is not always explicitly described in other studies, based on the database used for Frisch et al. (2004) available to us, the analysis considers only tri-consonantal verbs in UR (these forms lack reduplicative consonants and if weak roots are included, the missing consonant is present in the representation).

[13]The specific percentage is currently undetermined.

[14]As we are modelling words derived from the same basic template, vowels are not included in the analysis.

Table 3.12: Segment inventory

Place	non-labialized	labialized
labial	p’, b, m, w, f	b^w, m^w, f^w
dorsal	g, k ,k’	g^w, k^w, $k’^w$
glottal	h	h^w
coronal-stop	d, t, t’	d^w, t^w, $t’^w$
coronal-fricative	s, s’, z, ʃ, ʒ	s^w, z^w, $ʃ^w$
coronal-affricate	ʧ, ʧ’, ʤ	$ʧ^w$, $ʧ’^w$, $ʤ^w$
coronal-sonorant	n, l, r, ɲ, j	n^w, l^w

ization but that it is likely contrastive only in the case of dorsal stops and [h]. However, the number and phonemic status of labialized consonants varies widely for different authors. Armbruster (1908) records no labialized consonants at all and in the phoneme inventories of Cohen (1970) and Maddieson (1984)[15], the only labialized consonants are dorsal stops [k^w, $k’^w$ and g^w] and, for Maddieson, [h^w]. In later work, Hayward and Hayward (1992) find evidence for [$p’^w$ b^w, m^w, f^w, k^w, $k’^w$, g^w, $t’^w$, h^w] but do not indicate if all of these are contrastive. Unseth (2002) provides the same inventory as Leslau (1995) and similarly notes that labialization appears to be contrastive only for dorsal stops.

Leslau also notes that [p’] occurs only in older Greek loanwords and that [p] and [v] occur only in modern load words such as ‘vino’ (‘*wine*’), ‘viza’ (‘*visa*’), ‘polis’ (‘*police*’) and ‘posta’ (‘*mail*). Note that unlike Arabic, there are no glottal or pharyngeal consonants other than [h][16]. For this dissertation, we will assume the inventory of Leslau (1995).

[15]Maddieson’s inventory is based on Leslau (1968), Klingenheben (1966) and Sumner and François (1957).

[16]According to Leslau (1995), there is a marginal glottal stop that occurs between vowels that is also the reduction of [k’] for some speakers.

3.4.2 OCP-Place

The metric generally used in phonology to evaluate the co-occurrence of consonants is the Observed/Expected (hereafter O/E) ratio[17], a measure of how often two consonants appear together in words given how often they appear separately[18]. Equation 3.1 shows the O/E ratio for $A_x B_y$, the co-occurrence of segment A in position x and B in position emphy (and N is the number of words in the lexicon). O/E was first used to evaluate OCP-Place in Arabic verb roots by Pierrehumbert (1992). O/E is equal to 1 for consonants that co-occur as many times as would be expected (O = E), between 0 and 1 (O < E) for consonants that occur less often than would be expected all else being equal, and greater than 1 (O > E) for consonants that co-occur more often than would be expected.

$$\mathrm{O/E}(A_x B_y) = \frac{\frac{\text{occurrences}(A_x B_y)}{N}}{\frac{\text{occurrences}(A_x)}{N} * \frac{\text{occurrences}(B_y)}{N}} \tag{3.1}$$

Based on the discussion in King and Rose (2003), the coronal obstruents are analysed as two groups, fricatives and stops, contrary to the McCarthy (1988) analysis of Arabic. Affricates are analysed as stops based 1) on the "affricates as stops" approach to affricates described in Lin (2011), 2) the necessity, for purposes that will be described in detail in Chapter 4, of restricting the number of natural classes necessary to describe the segment inventory and 3) the fact that the affricates in Amharic can be shown to be historically derived from stops via palatalization (Leslau, 1957; Lowenstamm, 1986). [h], the sole (and rare) glottal, is not included in the analysis.

As our test data (discussed below) contains triliteral forms only, we are concerned with how the model will reflect the co-occurrence of homorganic consonants in left edge adjacent (C_1C_2X), right edge adjacent (XC_2C_3), and non-adjacent (C_1XC_3) locations. However, as we assume that all verb roots, including weak roots and roots with more than three consonants, influence speaker judgements

[17]Not to be confused with $O[C_i]/E_\theta[C_i]$ used in the Expectation Maximization algorithm of the Maxent Learner which is computed quite differently.

[18]A case might be made for the use of mutual information as a baseline metric, but O/E values allow us to directly compare our results with those of other researchers.

for triliteral forms, we must also determine how the presence of non-canonical roots in the learning data will influence statistical learning of triliteral patterns.

To include the statistical contribution of weak roots, we do not include the missing consonant in the representation, but, as speakers have cues (through the presence of vowels and location of gemination) to the location of the missing consonant, we use a place marker X in the place of the missing consonant to ensure that the edges of the verb root are correctly aligned. For example, the root /dm/ of 'addəmə', (*plot*) has the same consonantal representation as /dm/ of 'dəmma' (*bleed*). By replacing the missing consonants with a place marker X, we obtain the forms /Xdm/ and /dmX/ and the /dm/ sequence in 'addəmə' contributes to the statistics for the right edge, and the /dm/ in 'dəmma' contributes to the statistics for the left edge.

Table 3.13: O/E values of homorganic non-identical consonants in triliteral roots

POA	adjacent left edge	adjacent right edge	non-adjacent
Labial	0.15	0.18	0.5
Dorsal	0.14	0.0	0.0
Coronal-stops	.2	0.5	0.56
Coronal-fricatives	0.0	.04	0.18
Coronal-sonorant	0.8	0.2	0.6

Table 3.13 shows the O/E values for non-identical homorganic sequences in triliteral roots (the contribution of identical consonants is not included in this computation). O/E values are low (<1) in all cases with lowest values for dorsals and segments in adjacent positions, and higher values for some coronals and segments in non-adjacent positions. These results are similar to those previously reported in King and Rose (2003) and Rose and King (2007).

Table 3.14 shows the average O/E values for the co-occurrence of identical consonants in triliteral roots. The O/E value is computed for the co-occurrence of each segment and these are then are averaged across segments of like place of articulation. As O/E values for rare and distributionally irregular segments

Table 3.14: Average O/E values of identical consonants in triliteral roots

POA	adjacent left edge	adjacent right edge	non-adjacent
Labial	0.08	2.92	0
Dorsal	0.92	3.38	0.12
Coronal-stops	0.87	3.47	0.38
Coronal-fricatives	0.66	4.76	0.06
Coronal-sonorant	0.06	1.66	0

are strongly skewed (for example, the O/E value for p'p' at the left is 2294[19]), only the 14 most frequent and evenly distributed segments[20] are included in the computation. The O/E value for non-adjacent and left edge identical consonants is low (< 0), indicating that these forms are under-represented. However, the O/E value for right edge identical forms is high (> 0), reflecting the frequency of right edge reduplicative forms (ABB) in the lexicon.

Table 3.15: O/E values of all triliteral homorganic consonants (including identical consonants)

POA	adjacent left edge	adjacent right edge	non-adjacent
Labial	0.18	0.7	0.4
Dorsal	0.3	0.9	0.01
Coronal-stops	0.4	1.07	0.57
Coronal-fricatives	0.2	1.6	0.19
Coronal-sonorant	0.9	0.78	0.5

Table 3.15 shows O/E values for the co-occurrence of homorganic consonants in triliteral roots, including identical consonants. The O/E values for the right edge position are higher that for right edge homorganic non-identical consonants in table 3.13, showing the influence of identical consonants in that position.

Table 3.16 shows the O/E values for non-identical homorganic sequences

[19]p' occurs only 6 times in the lexicon but there is 1 root with p'p' at the left edge. The observed occurrence of p'p' is therefore much higher than the expected number of occurrences,

Table 3.16: O/E values of homorganic non-identical consonants in quadriliteral roots

POA	adjacent			non-adjacent		
	C_1C_2	C_3C_4	C_2C_3	C_1C_3	C_2C_4	C_1C_4
Labial	0.09	0.04	0.0	0.35	0.0	0.8
Dorsal	0.0	0.0	0.0	0.0	0.0	0.0
Coronal-stops	0.0	0.0	0.24	0.16	0.6	0.4
Coronal-fricatives	0.0	0.0	0.0	0.1	0.0	0.16
Coronal-sonorant	0.24	0.03	.27	2.3	0.98	0.13

in quadriliteral roots. The pattern is similar to the triliterals with low O/E values for all adjacent patterns and higher values for some coronals in non-adjacent position. In particular, the O/E value for coronal sonorants in $C_1C_3 = 2.3$, an indication that the pattern is over-represented. These results shows that OCP-Place as a restriction over non-identical homorganic consonants is active in verb roots of different lengths but that gradiency is an essential component of the description.

Table 3.17: Average O/E values for identical consonants in quadri-literal roots

POA	adjacent			non-adjacent		
	C_1C_2	C_3C_4	C_2C_3	C_1C_3	C_2C_4	C_1C_4
Labial	0.0	4.38	0.05	7.28	8.51	0.18
Dorsal	0.0	4.38	0.0	5.09	20.93	0.11
Coronal-stops	0.0	5.81	0.16	8.29	13.54	0.0
Coronal-fricatives	0.0	10.2	0.0	11.46	23.35	0.0
Coronal-sonorant	0.0	6.71	0.1	14.51	1.35	0.0

The O/E values for identical consonants in quadriliteral roots (table 3.17) reflect the presence of numerous ABCC and ABAB forms. O/E values are low at the left edge and in C_2C_3 (adjacent, non-edge) but O/E values for right edge (influenced by the frequency of ABCC roots) and non-adjacent patterns C_1C_3 and C_2C_4 (influenced by ABAB roots) other than C_1C_4 (where identical consonants

producing an exceptionally high O/E.

[20]These are b, f, m, n, l, r, t, t', d, s, z, k, k', g.

occur rarely) are consistently over-represented (O/E > 1).

Table 3.18: O/E values for all quadri-literal roots (including identical consonants)

POA	adjacent			non-adjacent		
	C_1C_2	C_3C_4	C_2C_3	C_1C_3	C_2C_4	C_1C_4
Labial	0.04	0.9	0.02	1.6	1.7	0.45
Dorsal	0.0	1.0	0.0	1.13	4.4	0.01
Coronal-stops	0.04	1.02	0.24	1.41	2.9	0.26
Coronal-fricatives	0.0	1.7	0.08	2.61	4.4	0.06
Coronal-sonorant	0.17	1.29	0.23	4.0	1.17	0.35

The O/E values for the co-occurrence of homorganic consonants (including identical consonants) in quadriliteral roots (table 3.18) directly reflects the frequency of patterns of identical consonants. O/E values are low in C_1C_2, C_2C_3 and C_1C_4 but high in all other positions.

With regard to longer roots, there are only 16 quinquiliterals without repetition. The 300 or so quinquiliterals with repeated consonants are predominantly ABCBC patterns (so doubtless contributing to higher O/E values in C_3C_5). Patterns such as ABABA and ABABC are rare.

To summarize, OCP-Place is active over homorganic consonants in verb roots, regardless of length, but is strongly modulated by three factors:

1 Place of articulation: dorsal consonants are most restricted and coronals the least

2 Distance: adjacent consonants are more restricted than non-adjacent

3 Location of violation: OCP-Place is a restriction over homorganic *non-identical* consonants in locations where identical consonants are frequent.

3.5 Experimental evidence

The general methodology for evaluating the psychological reality of a phonotactic restriction is to evaluate native speaker ratings for nonce words with specific phonotactic characteristics. Such experiments have shown that OCP-Place is an active constraint over homorganic non-identical consonants in Hebrew (Berent and Shimron, 1997), Arabic (Frisch and Zawaydeh, 2001) and Amharic (King and Rose, 2003). These results are compatible with models of phonotactics that predict low acceptability for words with patterns that violate a phontactic restriction (all else being equal).

For Hebrew, Berent and Shimron (1997) investigate the acceptability of nonce words with identical consonants based on the observation that of the 1449 productive triliteral roots in the Even-Shoshan (1993) dictionary, only 12 have initial identical consonants in AAB form. They create 24 nonce word triplets in each of the three word classes (206 items total). Each triplet is composed of one form with left edge identical consonants (AAB), one with right edge identical consonants (ABB) and one with no identical or homorganic consonants (ABC). The first experiment, a relative rating task where speakers are asked to assign a grade of 1 (possible), 2 (less possible) or 3 (least possible) to each member of a triplet, shows a dispreference for identical consonants in any position (AAB, ABB) compared to ABC forms. The authors interpret these results as an indication that both AAB and ABB forms violate OCP, according to an analysis under which OCP constrains not only lexical representations (UR) but also derivations (McCarthy, 1986). For the second experiment, speakers are asked to rate the nonce words freely on a 1-5 scale with 1 = "impossible" and 5 = "excellent" and the results show that only AAB forms are dispreferred compared to ABC forms, with no significant difference between ABB and ABC forms. A possible interpretation of these results is that in the first experiment, the difference in acceptability between the first and second points of the relative scale is not significant.

$$\text{Similarity} = \frac{\text{Shared natural classes}}{\text{Shared natural classes} + \text{non-shared natural classes}} \tag{3.2}$$

In an analysis proposed by Pierrehumbert (1992), the strength of the OCP-Place restriction is a function of segment *similarity*. Based on the metric developed by Pierrehumbert (1992), identical consonants are more similar than homorganic but non-identical segments and are therefore predicted to be more strongly constrained. Note that this analysis assumes that ABB patterns of identical consonants are derived and so not subject to OCP-Place. Frisch and Zawaydeh (2001) investigate this claim for Arabic by comparing speaker ratings for nonce word pairs with differing levels of similarity, ranging from no similarity (non-homorganic) to complete identity.

Using the set of native and assimilated triliteral roots from the Wehr and Cowan (1971) dictionary of Standard Arabic as a statistical foundation, Frisch and Zawaydeh (2001) compared nonce words with OCP-Place violations to violation-free controls, nonce words with accidental gaps (an accidental gap is a set of non-co-occurring segmental pairs that does not correspond to a coherent natural class) and evaluated the strength of the OCP-Place restrictions as a function of the similarity between segments. Speakers were asked to rate the acceptability of 254 nonce words balanced for expected probability (high-low) and neighbourhood density (dense-sparse) on a 1-7 scale with 1= "Impossible. This word sounds terrible" and 7 = "Definitely. This word sounds very much like a verb of Arabic".

The results showed that nonce words with OCP-Place violations were dispreferred compared to controls and that nonce words with OCP-Place violations were dispreferred compared to nonce words with accidental gaps. For the similarity comparisons, the small number of stimuli in each group makes the analysis somewhat tentative. Although the similarity metric holds as a predictor of relative badness within the group of non-identical homorganic consonants, identical consonants in AAB and ABA patterns were rated better than non-identical OCP-Place violations, contrary to the prediction. Finally, there was a trend such that ABB patterns were rated on a par or better than controls.

The King and Rose (2003) word acceptability data for Amharic is the test data that we use to evaluate the predictions of the model and is therefore described here in detail.

The King and Rose experiment was designed to investigate the psychological reality of two phonotactic restrictions in Amharic: OCP-Place and Laryngeal

Agreement (LA). LA is a restriction such that oral stops must agree in laryngeal qualities (voicing[21] and glottalization). Note that unlike OCP-Place, LA is a harmony restriction. LA is attested in verb roots in both Amharic (Bender and Fulass, 1978; Wedekind, 1990) and Chaha, an Ethio-Semitic Gurage language related to Amharic (Banksira, 2000; Rose and Walker, 2004). In Chaha, the restriction is broad, such that oral coronal and dorsal stops in any position (adjacent or non-adjacent) are either both voiceless (for example, /tkm/), both voiced (/fgd/) or both ejectives (/k'rt'/). King and Rose evaluate the presence of LA in the Amharic lexicon of verb roots and find that O/E values are low only for adjacent voiceless stops that differ in glottalization (ie, /kt'm/ or /mtk'/) in left edge (≈ 0.33) and right edge position (≈ 0.24). We will call this restriction *narrow* LA because it is a subset of the general LA restriction.

The task contained 270 items:

- 135 controls
- 45 stimuli containing OCP-Place violations (balanced for place of articulation and location of violation)
- 45 stimuli containing LA violations (balanced for location of violation).[22]
  - stops differing in voice (/nkd/)
  - stops differing in glottalization (/tnk'/)
  - stops differing in both voice and glottalization (/t'gl/)
- 45 stimuli containing both an OCP-Place violation and a broad LA violation (balanced for place of articulation and location of violation)

The nonce words were presented to 19 native Amharic speakers in Addis Ababa (the capital of Ethiopia) in a paper-and-pencil rating task. The speakers,

[21]See Gallagher (2010) and Mackenzie (2011) for a discussion of whether the restriction involves [voice].

[22]The stimuli included both broad and narrow LA violations even though the dictionary analysis suggested that only narrow LA violations would be active in Amharic. This was to facilitate comparison with similar test items in Chaha.

recruited through word of mouth, were asked to rate each word on an 6 point scale with 1 = “very Amharic-like” and 6 = “not like Amharic at all”.

We (re)analysed the ratings of the judgement task with a maximal linear mixed effect model (LMEM) with by-subject random slope for the fixed effect, and by-item random intercepts for each nonce word using the ‘lmer’ function of the statistical software R (package lme4). To estimate the p-value of a fixed effect, we compute the likelihood ratio between two models identical with regard to random effects but only one of which contains the fixed effect of interest. Where the p-value $< .05$, factors with $|t|$-value > 2 are assumed to be significant contributors to the effect[23].

The results show that dual violations (combining both OCP-place and LA violations), are significantly different from controls ($p < .01$) in all locations (t-values: right edge = 3.02; left edge = 4.5; non-adjacent = 3.2). OCP-Place violations are significantly different to controls ($p < .05$) in both left edge and right edge locations (t-values: left edge = 3.47; right edge= 2.68). However, the difference between controls and non-adjacent OCP-Place violations is not significant (t-value = 1.29).

There is no significant difference between LA violations generally and controls ($p = .3$), or between narrow LA violations and controls ($p = 0.2$). Nevertheless, location is a significant effect for LA violations ($p < .05$) though the t-values show that only right edge violations are significantly different to controls (t-values: left edge = -.077; right edge = 2.7; non-adjacent = 0.26). Location is not a significant factor for narrow LA violations overall ($p = .16$). However, the t-values show that narrow LA violations at the right edge are significantly different to controls (t-values: left edge = .16; right edge = 2.28; non-adjacent = 0.4).

To summarize the experimental results, the ratings show that speakers are very sensitive to dual violations in all locations and OCP-Place violations in adjacent locations. There is no significant difference between controls and ratings for non-adjacent OCP-Place violations. With regards to LA and narrow LA (which operates only over voiceless stops difference in [constricted glottis]), there is a

[23]We give a detailed example of this in Chapter 5.

significant dispreference for violations in right edge adjacent locations.

The results for OCP-Place are not unexpected. With regards to location, the restriction is statistically weakest in C_1XC_3 position and this is where we fail to find a significant dispreference in the speaker ratings. For LA, the results are more surprising. According to the dictionary study of King and Rose, only narrow adjacent LA violations are significantly under-represented in the verb roots. However, O/E values for right adjacent violations are consistently slightly lower than for the left edge adjacent violations and this holds for all forms of layrngeal agreement (different voice-different glottis, different voice-same glottis and same voice-same glottis). It may be the case that the speaker judgements are influenced by the frequent ABB forms (which were not included in the King and Rose dictionary study) because these never violate LA at the right edge of the word and necessarily decrease the O/E for the occurrence of stops differing in laryngeal features at the right edge.

3.6 Conclusion

The goal of this chapter was to show that Amharic verb roots are an interesting test case for the Maxent learner. Morphology makes it possible to model whole words with consonants alone and the phonotactic restrictions over identical and homorganic consonants are complex in terms of gradiency and length (because they are even active over non-adjacent consonants). In particular, we foresee that this data will pose at least four specific challenges to the Maxent learner:

1. The presence of numerous roots in reduplicative patterns makes it difficult to state the OCP-Place restriction in terms of homorganicity alone. To capture OCP-Place effects, the learner must be able to state the restriction over homorganic but non-identical segments.

2. The gradient nature of OCP-Place means that although a statement of prohibition against non-identical consonants can express the generalization, the particular ways in which this is stated must be modulated according to POA and location within the root.

3. The relationship between the frequency of LA violations in the lexicon of verb roots and the speaker ratings is unclear.

4. The variety of non-canonical root shapes (weak, reduplicative, quadri- and quiquiliteral) is likely to make the acquisition of a grammar that encodes broad generalizations more difficult.

4 Simulations

In previous chapters, we explained both the principles of the Maxent learner and the specific challenges of modelling Amharic phonotactics. In this chapter, we evaluate the performance of the Maxent learner in predicting (simulating), speaker judgements of nonce words from the King and Rose (2003) word rating task described in Chapter 3. We use this data rather than directly developing our own because it provides us with a baseline for formulating speaker rating tasks that specifically target our research needs.

Our main research goal is to investigate the relationship between the lexical frequency of patterns and gradient speaker judgements so the correlation between the model predictions and the experimental data of King and Rose (2003) is of crucial interest to us. However, the performance of the model in predicting speaker judgements may be affected by a number of factors such as the statistical quality of the model, the representational system and the constraint selection algorithm and each of these must also be evaluated to determine their contribution to model performance.

Our general procedure is to train the model on our database of verb roots, evaluating it incrementally as automatically selected constraints are added to the grammar. In this way, we have a clear picture of how the model performs as a function of grammar size. We also create a second model trained on the same data but equipped with a 'starter set' of hand-written constraints describing OCP-Place as it is expressed in the training data. The comparison of the results of the two models, one automatic and one initialized with hand-written constraints, allows us to evaluate the quality of the automatic constraint selection algorithm. Both models are trained on the complete lexicon of verb roots, including roots with identical consonants, regardless of location. Vowels are not included in the representation because, as discussed in Chapter 3, they are constant across a given verbal derivation[1] and, presumably, do not participate in the restrictions over the consonants of the root.

In the first simulation, both models are trained on the full set of Amharic verb roots expressed in surface-true form (with some modifications for weak roots that we will discuss). The results show that the automatic model performs less well in predicting speaker judgements than the model equipped with hand-written constraints and our analysis suggests that the representation of identical consonants as surface true artificially reinforces the importance of LA (Laryngeal Agreement), a weak harmony restriction.

Building on the results of the first simulation, a second pair of models (again, one with fully automatic constraint selection, the second initialized with a hand-written grammar) is trained on a modified representation of the verb root lexicon where repeated occurrences of a consonant within a verb root are replaced by a featureless place marker. The results show that 1) the hand-written model is no longer significantly better than the automatic one, providing evidence that identical consonants do not contribute to the perception of the harmony restriction, 2) there is evidence for a previously unstudied restriction over non-homorganic fricative consonants in Amharic verb roots and 3) our test data is too narrow in scope to fully explore the performance of the Maxent learner and that more experimental data is needed.

In the next section, we describe our first simulation with particular attention to explaining the rationale for our design choices and the details of the evaluation methodology.

4.1 Simulation I: baseline model

Our goal in the first simulation is to make as few assumptions as possible in the choice and representation of the training data, given the limitations of the

[1]This holds for verb Types A and B, but not Type C which has a distinctive vowel.

software. The results for this simulation are then used to motivate modifications in the second simulation.

4.1.1 Training data

The model has access to the training data through two files. In the first, the consonants of the segment inventory are defined in terms of distinctive feature bundles. The second is simply a list of the words of the lexicon.

Table 4.1: Segment inventory

Place	non-labialized	labialized
labial	p', b, m, w, f	bʷ, mʷ, fʷ
dorsal	g, k ,k'	gʷ, kʷ, k'ʷ
glottal	h	hʷ
coronal-stop	d, t, t'	dʷ, tʷ, t'ʷ
coronal-affricate	s, s', z, ʃ, ʒ	sʷ, zʷ, ʃʷ
coronal-fricative	ʧ, ʧ', ʤ	ʧʷ, ʧ'ʷ, ʤʷ
coronal-sonorant	n, l, r, ɲ, j	nʷ, lʷ

Segment inventory The items in our training data are the consonontal verb roots extracted from Kane (1990). Although vowels are not included in the representation, the consonant inventory is still quite large and creates issues with training the model.[2] For this reason, we collapsed labialized consonants with their non-labialized counterparts, thereby focusing on the major place of articulation. The resulting set (table 4.1 repeated from Chapter 3) contains 25 segments defined using 10 distinctive features. This design decision does not affect the calculation of OCP-Place, which operates over primary place of articulation.[3]

[2]Attempts to use the Maxent learner with the full set of consonants did not produce usuable results. Training was extremely slow and the software encountered internal errors.

[3]To validate this assumption, we computed O/E values for the co-occurrence of labial consonants and labialized dorsal consonants. The average O/E value is 1.58.

Our preference would be to avoid the use of under-specification in the distinctive feature bundle representation of segments because researchers do not agree on how and what to underspecify (Steriade, 1995). However, when every segment is specified for every natural class, the number of possible natural classes is large and makes the model extremely slow to train. We decided to use under-specification sparingly as shown in table 4.2. *Constricted glottis* is defined only for voiceless obstruents because only voiceless consonants can be ejective. The feature *anterior* is defined only for coronals as it is the feature that distinguishes between alveolar coronals and post-alveolar coronals (Clements, 1985). For example, [anterior] is the feature that contrasts the two voiceless coronal strident consonants [s] and [ʃ], as [s] is [+anterior] but [ʃ] is [-anterior]. Features for place of articulation are unary[4] while all the others are binary. This use of underspecification is for implementational reasons only and does not imply a particular theoretical stance.

Table 4.2: Use of underspecification

distinctive features	**specification scheme**
consonantal sonorant continuant nasal strident	*specified for all segments*
anterior	*specified only for coronals*
constricted glottis	*specified only for voiceless obstruents*
dorsal labial coronal	*unary features*

[4]Unary features are specified only where positive. For example, only dorsal segments are specified for the feature *dorsal*.

Lexicon As we have a strong bias towards avoiding theoretical assumptions where possible, our goal was to encode the training data in a form as close to the surface realization as possible, given the constraints of the model. Our representations contrast with the general trend in Semitic phonology where analysis is performed over the (assumed) underlying form of regular triliteral verb roots only (Frisch et al., 2004; Pierrehumbert, 1992). In such a representation, surface consonants assumed to be derived through copy or spreading would be absent, but the consonants missing from weak roots (which are assumed to be present in underlying representation) would be present.

Our model is trained on the complete inventory of 4243 consonantal verb roots drawn from Kane (1990). As our representation is designed to be neutral with regard to the conjugation of the verb, it does not include information about consonant gemination.[5] The training data includes not only the standard Semitic triliteral roots, but also weak roots (those appearing to lack a surface root consonant), roots with 4, 5 or 6 consonants and roots with identical consonants. Roots with 3 or more consonants and no identical consonants were encoded directly as they appear in the dictionary, but the encoding of weak roots and roots with identical consonants required some analytical decisions.

- **Weak roots** As described in Chapter 3, weak roots lack a consonant. However, the placement of the missing consonant can be induced through the position of medial gemination, the presence of an extra vowel, or the palatalization of the preceeding consonant. As some of these cues are not available in the consonantal templatic representation (notably because there is no indication of gemination), our objective is to supply the model with equivalent information. We chose to replace the missing consonant with the place holder 'X' in that position. For example, the weak root '**səbba**', which has final [a] instead of a consonant, is encoded as [sbX], and the weak root '**hedə**', which is presumed to have a medial glide /j/ due to the front vowel, as [hXd]. 'X' is defined as having a single distinctive feature, 'x', and is unspecified for all

[5] The representation could be any conjugation of the verb or derived word, as long as the non-root material is constant across forms.

other features; all other segments are unspecified for 'x'. This representation is designed to be neutral with regard to the identity of the underlying consonant but to encode the information about its position such that word edge information is preserved.

- **Identical consonants** All roots in our database with identical consonants are encoded as surface true. For example, roots such as 'wəttətə' '*wander*' which follows the ABB pattern, is encoded with identical consonants: [wtt].

Table 4.3 summarizes the encoding choices in simulation I by root type.

Table 4.3: Verb root encoding by type

root type	*example*	*gloss*	*root*	*encoding*
surface true	bəggənə	'get furious'	bgn	bgn
weak	awwədə	'perfume'	wd	Xwd
	bədəbbədə	'beat'	bd	bdbd
identical	bəbbətə	'put inside'	bbt	bbt
	wəjjəwə	'lament'	wjw	wjw

Statistical analysis of OCP-Place To summarize the discussion in Chapter 3, *non-identical* homorganic consonants are generally under-represented in all locations for roots of all lengths (O/E values < 1) but less so in non-adjacent locations. If the occurrence of identical (and therefore homorganic) consonants is included in the computation of the O/E values, the description of OCP-Place is modified and homorganic consonants are reliably under-represented only in the left edge adjacent location where identical consonants rarely occur.

4.1.2 Test data

We evaluate model predictions against the King and Rose (2003) speaker judgement data described in Chapter 3. To briefly summarize, a set of nonce verbs, which used only the 14 most frequent and evenly distributed consonants, contained 90 forms with OCP-Place violations representing a range of predicted acceptability according to location of violation (left edge, right edge, non-adjacent) and place of articulation (dorsal, coronal[6], labial). 19 native speakers of Amharic were asked to rate the nonce forms (all conjugated identically as CəC:əCə) on a 1-6 scale with 1 = very Amharic-like and 6 = not like Amharic at all. Speakers significantly dispreferred nonce forms with OCP violations over controls ($p < .05$).

4.1.3 Procedures

The model was programmed to acquire two distinct grammars:

i. an automatically learned model with 1000 constraints

ii. a Maxent weighted hand-written grammar.

We know from the statistical analysis that for our choice of lexicon, the grammar must encode OCP-Place as a restriction over homorganic but non-identical consonants. To evaluate the automatic constraint selection algorithm, we compare a model with automatically selected to constraints to one that is initialized with hand-written constraints. The hand-written grammar describes the restriction using constraints that are available to the Maxent learner and those constraints are then weighted according to the principle of Maximum Entropy. Table 4.4 illustrates the hand-written grammar for the segment [b]. Note that as the restriction must be described specifically for each individual segment, using only those natural classes predefined by the model, the final hand-written grammar for the baseline simulation contains 384 constraints. Once the constraints of the hand-grammar are weighted, automatically selected constraints are added to the hand-written model until grammar size reaches 1000 constraints.

[6] Coronals are subdivided according to manner: stops, fricatives and sonorants.

Table 4.4: Example of hand-written constraints

condition	*expression*
left edge (C_1C_2)	*[word boundary][b][Labial not b]
	*[word boundary][Labial not b][b]
right edge (XC_2C_3)	*[b][Labial not b][word boundary]
	*[Labial not b][b][word boundary]
non-adjacent (C_1XC_3)	*[Labial not b][not Labial][b]
	*[b][not Labial][Labial not b]

4.1.4 Model evaluation

Our main goal is to evaluate the relationship between the frequency of patterns in the lexicon and the speaker ratings. However, the predictive quality of the model is not only a function of that relationship but also of the statistical quality of the model, the constraint selection algorithm, and the available representational system. The statistical quality of the model and the relationship between probabilistic constraints and speakers judgements are evaluated via statistical metrics, and the contribution of the constraint selection algorithm and representational system are evaluated by inspection of the learned constraints and the comparison of the performance of the automatic and hand-written models.

Statistical quality of the model Prior to evaluating the predictions of the model, it is crucial to ascertain that the statistical model itself is functioning correctly. The underlying principle of statistical models is to encode an accurate statistical reflection of the learning data (according to some principle). Log-linear models such as the Maxent learner learn by maximizing $P_\Theta(D)$ as described in Chapter 2. This is usually referred to as maximizing the *likelihood* of the data, given the model. To avoid the problems of computing very small numbers, it is usual to use $logL_\theta(D)$, which varies monotonically with $L_\theta(D)$.

Evaluating $logL_\theta(D)$ allows us both to incrementally evaluate the progression of learning and to compare the statistical quality of one model relative to another. However, as the Maxent learner uses a hill-climbing algorithm to find the constraint weights that maximize $logL_\theta(D)$ without ever actually computing

$logL_\theta(D)$ itself, we need to manually compute $logL_\theta(D)$ for each incremental grammar.

$$P_\theta(D) = \sum_{x \in D} \frac{P_\theta(x)}{Z}$$
$$Z = \text{ a normalizing factor} \tag{4.1}$$

$logL_D(\theta)$, which is equivalent to $P_\theta(D)$ described in Chapter 2 (equation 4.1) equals, for a given grammar, the sum of the Maxent values assigned to the words of the training data divided by the partition function, the sum of the Maxent values assigned to all words in Ω, the set of all possible words y of the language (as this is an infinite set, only words no longer than the longest word in the learning data are considered).

$$Z = \sum_{y \in \Omega} \text{Maxent value}(y)$$
$$\text{Maxent value(y)} = \exp(-\sum \text{weights assigned to y}) \tag{4.2}$$

Monitoring the progression of statistical learning during training also allows us to evaluate over-training, a common issue with statistical models. In over-training, the model is such an accurate reflection of the training data that it lacks generalization power and does not perform well on novel items. We track over-training by computing the 5-fold cross-validated $logL_\theta(D)$. In this procedure, the training data is partitioned into 5 equal sections. The model is trained on 4/5 of the training data and the log-likelihood is evaluated on the final 1/5, which is called the held-out (or *unseen*) data. This is repeated for each of the 4 other partitions and the 5 measures of log-likelihood are then averaged. As learning progresses, the cross-validated log-likelihood should rise; a fall would indicate that the model is becoming too specific to the training data and losing generalization power.

Predictiveness The stated goal of the Maxent learner is to model speaker judgements. The predictive quality of the model is evaluated by measuring the correlation between the weights assigned to novel words by the learned grammar and the ratings assigned by actual speakers.

Linking function The linking function between the log-likelihood of the data and the predictive capacity of the model is a foundational assumption of the Maxent learner. According to this linking function, the grammars with the highest log-likelihood should also be those most predictive of speaker judgements for novel items.

Representational system The well-foundedness of Maxent models for modelling linguistic data is well-established (Berger et al., 1996; Della Pietra et al., 1997; Eisner, 2001; Manning and Klein, 2003; Rosenfeld, 1996). However, the representational system and process of constraint selection used by the Maxent learner are novel. To evaluate the performance of these, we compare the performance of automatically learned grammars with a grammar of Maxent weighted hand-written constraints and we examine the specific constraints that correspond to a rise or fall in log-likelihood or predictiveness.

It is to be noted that these metrics are related in complex ways. For example, a disconnect between the progression of the log-likelihood of the data and model predictiveness would not necessarily mean that the linking function should be called into question. It might simply be the case (among many other things) that the representational system does not allow the model access to information that is available to speakers.

Evaluation schedule The automatic and hand-written models are both evaluated incrementally as new constraints are added to the grammar. The correlation between the automatic model and the speaker judgements is evaluated after the acquisition of each new constraint until the grammar size reaches 100, and every 20 constraints thereafter. The correlation between the hand-written model and the speaker judgements is first evaluated after the 384 hand-written constraints are incorporated into the model, and then after every 20 constraints in the same manner as for the automatic model. The evaluation of the log-likelihood of the training data is considerably more time-intensive as, for each data point, it requires the evaluation of the partition function Z (equation 4.2) for each of the five folds of the training data. For these evaluations we compute Z exactly by exhaustive enumeration of the unnormalized probability for all possible wordforms. For the automatic model, the log-likelihood is evaluated in increments of 40 constraints and then, for those sections that show a sharp change in log-likelihood, for each individual grammar. For the hand-written model, the log-likelihood is first evaluated at 380 constraints (so just before all the 384 hand-written constraints are integrated into the model), in increments of 20 constraints over areas of the curve that show sharp change, and in increments of 40 constraints elsewhere.

4.1.5 Results and discussion

Cross-validated log-likelihood of data For the automatic grammar, figure 4.1 shows that the model is performing as we would expect: as constraints are first added to the grammar, the log-likelihood rises sharply and then tapers off. The constraints learned during the initial period describe strongly under-represented segments and patterns. For example, one of the earliest selected constraints assigns a weight to [p’], which is an extremely rare segment. The log-likelihood peaks at 360 constraints and then starts to fall slowly, an indication that the model is becoming over-trained.

The 384 hand-written constraints of the hand-written grammar do not raise the overall log-likelihood of the model as much as the constraints that are added to it by the automatic selection algorithm. The log-likelihood of the hand-written model peaks at 900 constraints and continues to rise moderately, showing no sign of over-training. The peak of the hand grammar is slightly higher than that of the automatic model (-9888 and -9893, respectively).

Correlation between model predictions and speaker judgements Figure 4.2 shows the correlation between speaker judgements and model predic-

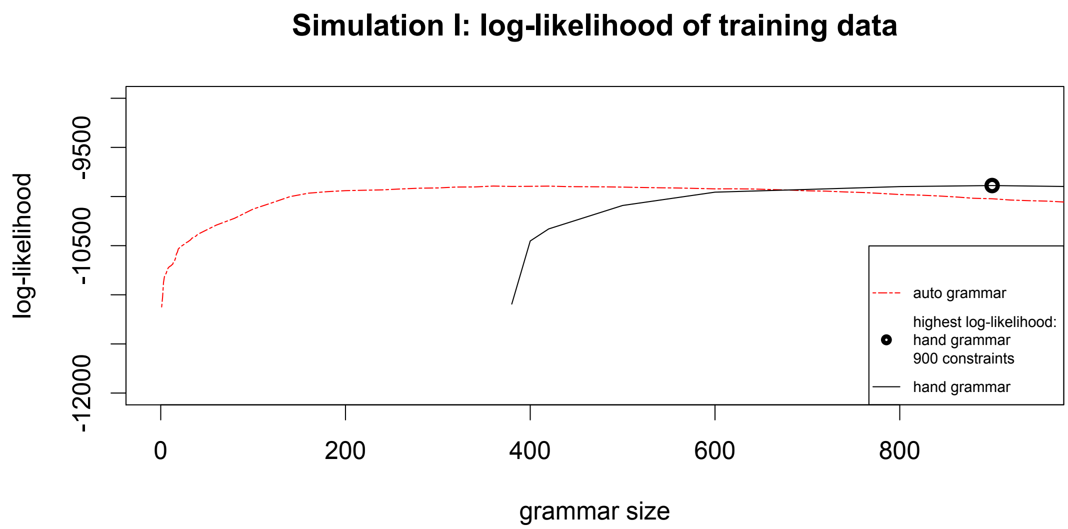


Figure 4.1: Log-likelihood of training data

tions for both models. The peak in the hand-written model at 384 constraints (r = .45) is higher than the peak in correlation for the automatic model (r = .34 at 80 constraints). This difference is statistically significant ($p < .01$, with non-parametric bootstrapping).[7] This implies that automatic constraint selection is less optimal than a phonologically motivated grammar that explicitly encodes OCP-Place.

One unexpected result is that for the hand-written model, the correlation with speaker judgements falls off as automatically selected constraints are added to the grammar. This fall in correlation between the hand-written grammar and the speaker judgements between 384 and 1000 constraints is significant ($p < .01$, with non-parametric bootstrapping).

[7] Bootstrap method used for comparing two models, A and B: For each of N replications, sample the items in the training data with replacement. Calculate the correlation coefficient between the scores of each model with the speaker judgements on that sample. The proportion of samples for which A has a higher correlation coefficient than B can be taken as the p-value of a one-tailed hypothesis test that A is better than B.

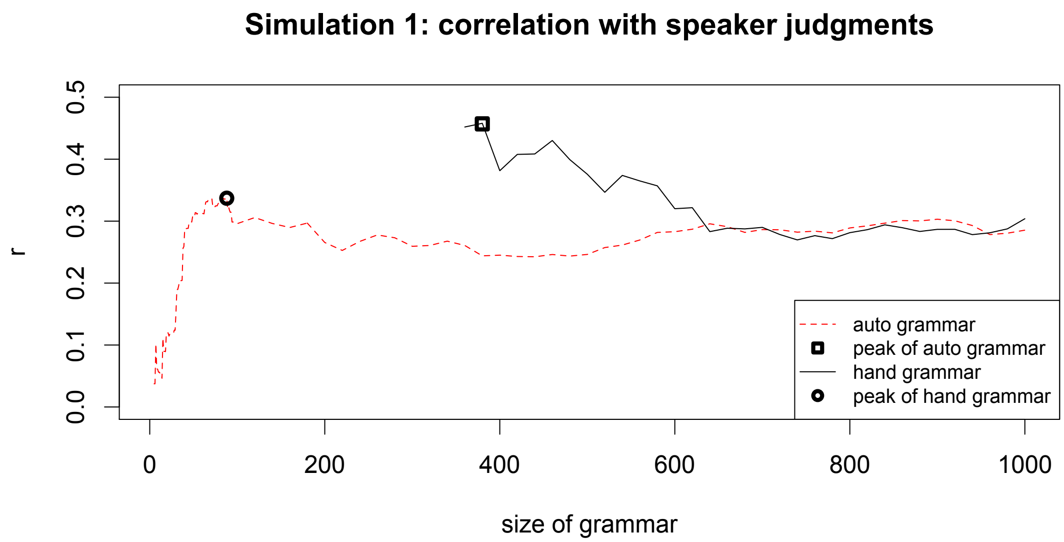


Figure 4.2: Correlation between model predictions and speaker judgements

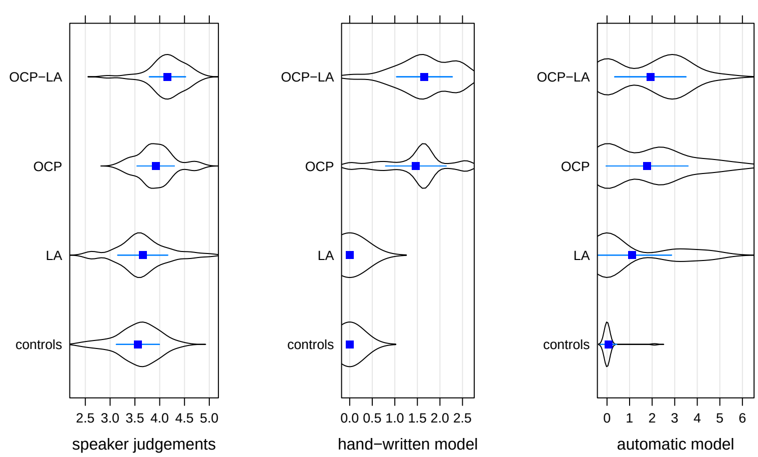


Figure 4.3: Speaker predictions and model performance for OCP-Place

Performance on OCP-Place Figure 4.3 shows violin density plots of speaker ratings grouped together for the main experimental conditions of King and Rose (2003) (OCP violations, LA violations (stops that differ in voice and/or constricted glottis), OCP-LA violations (nonce forms that violate both restrictions) and controls) side-by-side with the predictions of the best (most predictive) hand-written and automatic models. A violin density plot is similar to a boxplot except the width of the box is proportional to the frequency of observations in that area. The blue line designates the second and third quantiles and the square blue box is the median. Note that the speaker judgements for forms with OCP-Place and OCP-LA violations are higher (indicating lower acceptability) than for controls, and that both models make similar predictions.

Analysis of automatically acquired constraints

Table 4.5: First six constraints acquired by the automatic grammar

Constraint number	*Constraint*
1	*[ˆ+dorsal,+constricted-glottis][+consonantal] [+labial,+constricted-glottis] (*[not k’][not j or w][p’])
2	*[+labial,+constricted-glottis] (*[p’])
3	* [+nasal,-anterior] (*[ɲ])
4	*[-sonorant,+coronal][-sonorant,-anterior] (*[ʒ, d, ʤ, s, s’, t, t’, ʧ, ʧ’, z][ɲ, ʒ, ʤ, ʧ, ʧ’])
5	*[-voice,+strident][+voice,+strident] (*[s, s’, ʧ, ʧ’][ʒ, ʤ, z])
6	*[-continuant,-anterior][-sonorant,-continuant,+anterior] (*[ɲ, ʤ, ʧ, ʧ’][d, t, t’])

Automatic constraints corresponding to a rise in log-likelihood The constraints that correspond to a sharp rise in log-likelihood are similar for both models. Table 4.5 shows the six first automatically selected constraints of the automatic grammar. The first three involve rare segments ([p'], [ɲ]). The fourth, fifth and sixth constraints encode a restriction over coronal obstruents in adjacent position without reference to word edge. As discussed in Chapter 3, OCP-Place has been found to be active over coronals generally, although it is stronger within the stop and fricative groups than between (Greenberg, 1950; McCarthy, 1986) so this constraint appears to be a reasonable generalization for a super-set of OCP-Place for coronal fricatives and stops in adjacent position (bearing in mind that, all else being equal, the constraint selection heuristic favours short constraints over long ones and constraints involving many segments over those that involve few) .

Table 4.6: First six constraints acquired automatically by the hand grammar

Constraint number	*Constraint*
385	*[-word boundary][ˆ+sonorant,+continuant,+anterior] [+labial,+constricted-glottis] (*[any segment][not r, l, n][p'])
386	*[+nasal,-anterior][-word boundary][-word boundary] (*[ɲ][any segment][any segment])
387	*[ˆ+voice,-anterior][-voice][+voice,-anterior] (*[not ʤ, ɲ or ʒ][p', f, ʃ , s', s, ʧ, ʧ', t', t, k', k][ɲ, ɲ, ʤ, ʒ])
388	*[+labial,+constricted-glottis] (*[p'])
389	*[-sonorant,+anterior][-sonorant,-anterior] (*[d, s, s', t ,t', z][ʃ, ʒ, ʤ, ʧ, ʧ'])
340	*[-sonorant,-anterior][+continuant,+strident] (*[ʃ, ʒ, ʤ, ʧ, ʧ'][ʃ, ʒ, s, s', z])

For the hand-written model, the correlation peaks at 384 constraints but the log-likelihood rises sharply thereafter. Table 4.6 shows that the first six *automatically* learned constraints are very similar to those of the fully automatic model; the first four are restrictions over rare ([p'], [ɲ]) and distributionally irregular ([ʤ]) segments followed by two constraints generalizing over OCP-Place for coronal obstruents similar to those of the automatic grammar (although they do not improve the correlation with speaker judgments).

Automatically learned constraints corresponding to a fall in predictiveness For both the automatic and hand-written grammars, the acquisition of some constraints corresponds to a fall in predictiveness. In particular, in the hand grammar, the correlation between the grammar and the speaker judgements falls from .45 to .38 with the acquisition of a single constraint over the co-occurrence of ejective voiceless stops and affricates and non-ejective voiceless obstruents:

*[-continuant,-constricted glottis][+constricted glottis]

For the automatic grammar, the acquisition of constraints involving segments with differing values of constricted glottis also corresponds to a fall in predictiveness (the correlation falls by .25 for two combined).

*[-constricted glottis,-strident][-sonorant,+continuant][+constricted glottis]

*[+voice,+coronal][+constricted glottis][+labial]

Our analysis is that although the model is effectively acquiring the restriction over non-identical homorganic consonants as it is encoded in the experimental items of the speaker judgement task, the representation of lexically reduplicated consonants in a manner analogous to ordinary consonants is problematic. In particular, the encoding of identical consonants alters the nature of LA in the lexicon. As discussed previously, in Amharic, LA (both broad and narrow) is a weak restriction both statistically (King and Rose, 2003) and in speaker judgements. The

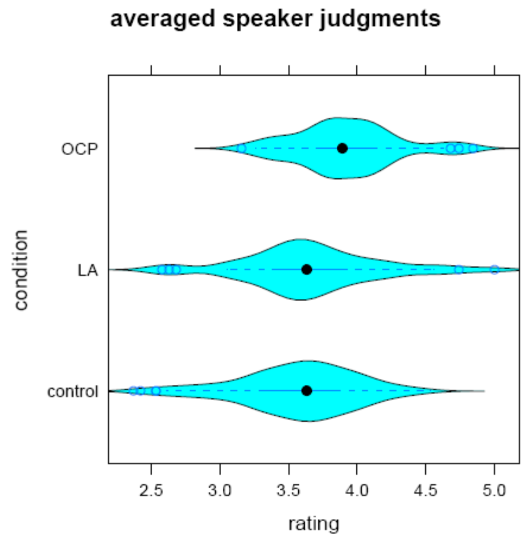


Figure 4.4: Speakers judgements for LA violations

King and Rose (2003) judgement task data investigated Amharic speaker sensitivity to both OCP-Place violations and LA violations and the results show that there is a moderate dispreference for LA violations in the right edge location only.

The results of the judgement task are shown in the violin density plot in figure 4.4 that compares ratings for OCP and OCP-LA violations (combined) to LA violations and controls. Note the similarity between the ratings for controls and for LA violations. This result contrasts with figure 4.5 which shows predicted acceptability ratings for the best automatic model. The predictions for LA violations are stretched more to the unacceptable range than is the case for the speaker judgements.

This over-estimation of the unacceptability of nonce forms with LA violations is directly related to the presence of identical consonants in our training data. Consider the case of the roots [mlt] and [fnk'] and their possible counterparts with identical consonants [mltt] and [fnk'k']. The repeated stops agree in *voice* and *constricted glottis*, strengthening the statistical adherence to LA.

Our general analysis for the baseline simulation is that the model makes predictions for many restrictions over rare and positionally restricted consonants that are not included in the speaker judgement data and that cannot therefore be

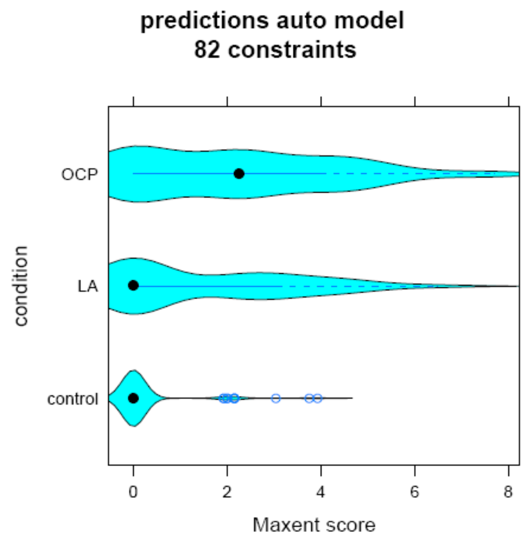


Figure 4.5: Automatic model predictions for LA violations

evaluated. Moreover, the model's prediction that LA violations would be dispreferred by speakers appears to be an artefact of the encoding of identical consonants. It may be the case that identical consonants are produced through a process that is distinct from single feature harmony (such as LA) and that the two processes do not interact (Gallagher, 2008) or that identical consonants are effectively absent from UR where harmony is computed. Our analysis does not allow us to determine which of these hypotheses is correct but it does motivate a second simulation that modifies the representation of identical consonants such that they do not reinforce harmony constraints.

4.2 Simulation II: modified model

In simulation I, every occurrence of identical consonants in the training data is an occurrence of identical distinctive features. For laryngeal features, this poses a problem because it strengthens the presence of LA, an otherwise weak harmony restriction, and the model over-estimates speaker dis-preference for nonce words that violate the restriction. The goal of simulation II is therefore to modify the representation such that an occurrence of identical consonants does not entail the occurrence of identical distinctive features.

In recent work, Berent et al. (2012) propose a representational system that addresses the issue of interaction between reduplication and single-feature harmony and the need to be able to generalize over patterns of identical consonants. These researchers show that Hebrew speakers are capable of generalizing their knowledge of identical consonant patterns to consonants that are not native to their language. For example, speakers dis-prefer identical consonants at the left edge of the root even if the consonants do not occur in Hebrew. Their solution is to enrich the set of phonological forms (potential constraints) with 1) a feature matrix [segment] and 2) a representation of identical consonants where the second consonant is void of features and indexed to the first. For example, in our version of the Maxent learner, the phonological form *#[+labial,+nasal][+labial,+nasal] represents two [m] segments at the left edge of a word, but the Berent et al. version of the model also has access to *#[+labial,+nasal]$_{i\_i}$ and *#[segment]$_{i\_i}$ where the second segment is indexed to the first but has no distinctive features of its own. The *#[segment]$_{\_i}$ representation solves Berent et al.'s problem of generalization to non-native segments and presumably, the *#[labial,+nasal]$_{i\_i}$ representation would allow the model to acquire knowledge of identical consonant patterns without reinforcing single-feature harmony.

Unfortunately, the model used by Berent et al. (2012) was not available at the time we ran these simulations. Our solution was to replace identical consonants after the first occurrence in a word with the featureless placeholder X representation used for weak roots . For example, roots such as 'bələssəsə' *'to bare the teeth'*, which has the root consonants 'blss' (The second occurrence of 's' is a geminate), and follows the ABCC pattern, are encoded as [blsX]. The placeholder method allows reference to the templatic pattern and word edge information of the root, in the case of 'bələssəsə', a quadriliteral pattern, and keeps it distinct from an unrelated triliteral root, [bls] for the verb 'bəlləsə' *'to be very tired'*. In deciding which of the identical consonants in a root should be coded as 'X', we decided that the first occurrence of a consonant (counting from left to right) should be coded as surface true, and that successive occurrences would be coded as 'X'. The purpose of this decision was to preserve information about the left word-edge because, as our learning data contains roots of different lengths, the left edge is the single location shared by all roots where identical consonants are rare and is therefore a strong cue to finding the OCP-Place restriction.

This representation is not ideal. It encodes neither the relationship between the first and successive occurrences of a consonant (so [blX] could encode either 'blb' or 'bll') nor does it distinguish between copies of different consonants (so [blXX] could be 'blll, 'bllb, 'blbl' etc). A better method would be to use a different placeholder representation for each consonant; the placeholder for [r] would be different from that for [t]. However, the limitations of the model in terms of the size of the segment inventory would have made this solution difficult to implement.

Table 4.7: Verb root encoding by type

root type	*example*	*root*	*sim I*	*sim II*
surface true	bəggənə'get furious'	bgn	bgn	bgn
weak	awwədə 'perfume'	wd	Xwd	Xwd
	bədəbbədə	'beat'	bdbd	bdXX
identical	bəbbətə	'put inside'	bbt	bXt
	wəjjəwə	'lament'	wjw	wjX

Table 4.7 contrasts the encoding in the baseline and modified models. The shaded cells show patterns with identical consonants where any occurrence of a consonant after the first is encoded as the placeholder 'X'. Note that this decision deprives the model of any information about the relative frequency of identical consonants, including those in non-reduplicative patterns.

The absence of identical consonants from the representation modifies the description of OCP-Place for the hand-written grammar. For the baseline model, the hand-written grammar encoded a restriction over homorganic but non-identical

Table 4.8: Comparison of hand-written grammars

	condition	*expression*
Baseline model	left edge (C_1C_2X)	*[word boundary][b][Labial not b]
		*[word boundary][Labial not b][b]
	right edge (XC_2C_3)	*[b][Labial not b][word boundary]
		*[Labial not b][b][word boundary]
	non-adjacent (C_1XC_3)	*[Labial not b][not Labial][b]
		*[b][not Labial][Labial not b]
Modified model	left edge (C_1C_2X)	*[word boundary][Labial][Labial]
	right edge (XC_2C_3)	*[Labial][Labial][word boundary]
	non-adjacent (C_1XC_3)	*[Labial][word boundary][Labial]

consonants. However, absent the identical consonants, the hand-written grammar can simply be stated as a restriction over homorganic consonants according to place of articulation and word location. Table 4.8 contrasts the encoding of OCP-Place in the hand-written grammar of the baseline model and the modified model. Given the important role of LA in the first simulation, we included constraints describing the restriction over stops with different features of glottis into the modified hand-grammar. The complete hand-written grammar of the modified model (including constraints for LA) contains only 27 constraints compared to 384 in the baseline.

In all aspects other than the encoding of identical consonants and the expression of the hand-written grammar, the modified simulation is identical to the first: two competing models, a fully automatic model and a model initialized with a hand-written set of constraints, are evaluated and compared as for the baseline.

4.2.1 Results and discussion

Figure 4.6 shows $logL_\theta(D)$ for simulation II. Note that for both models, the curve appears to fall after 360 or so constraints. This is likely a sign of over-training. Similarly to simulation I, the log-likelihood of the hand-written model does not rise until after the incorporation of the (18) hand-written constraints.

After incorporation of the hand-written constraints, the curves for both

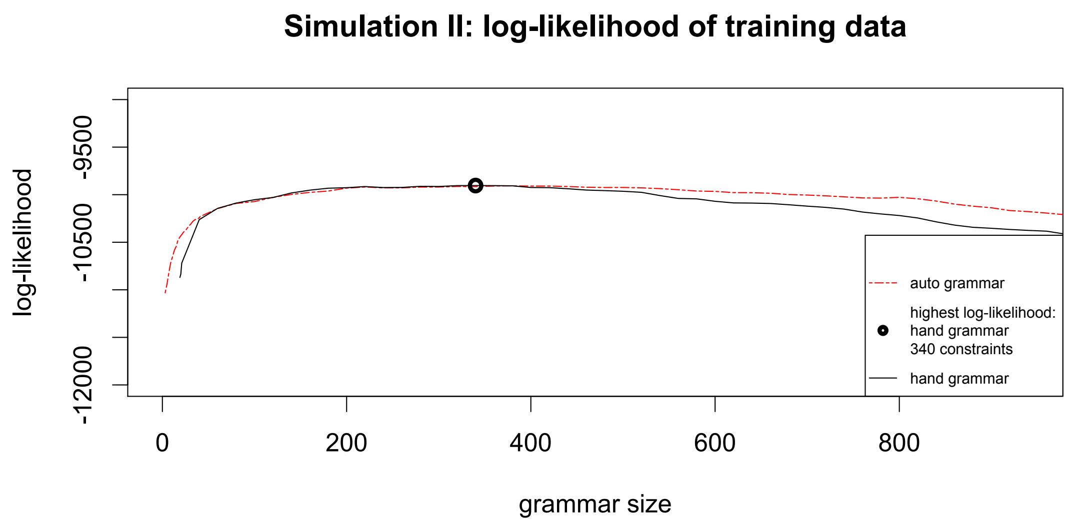


Figure 4.6: Log-likelihood of data: simulation II

models are extremely similar. For both models, and as for simulation I, the sharp rises in log-likelihood corresponds in many cases to the automatic acquisition of constraints over generally rare segments and positionally restricted segments. There is essentially no difference between the two models in terms of log-likelihood as the automatic model peaks at -9907 for 380 constraints and the hand written model peaks at -9904 with 340 constraints.

Figure 4.7 shows the correlation between the models and speaker judgements. The hand-written model peaks at 40 constraints ($r = .45$) but the automatic grammar peaks much later at 380 constraints ($r = .39$). However, the difference between the peak of the automatic model and the peak of the hand-written model is no longer statistically significant $p > .05$).

Model performance on OCP-Place Figure 4.8 shows that, as for simulation I, both the hand-written and automatic models perform well in predicting that nonce forms with OCP and OCP-LA violations are worse than controls. The

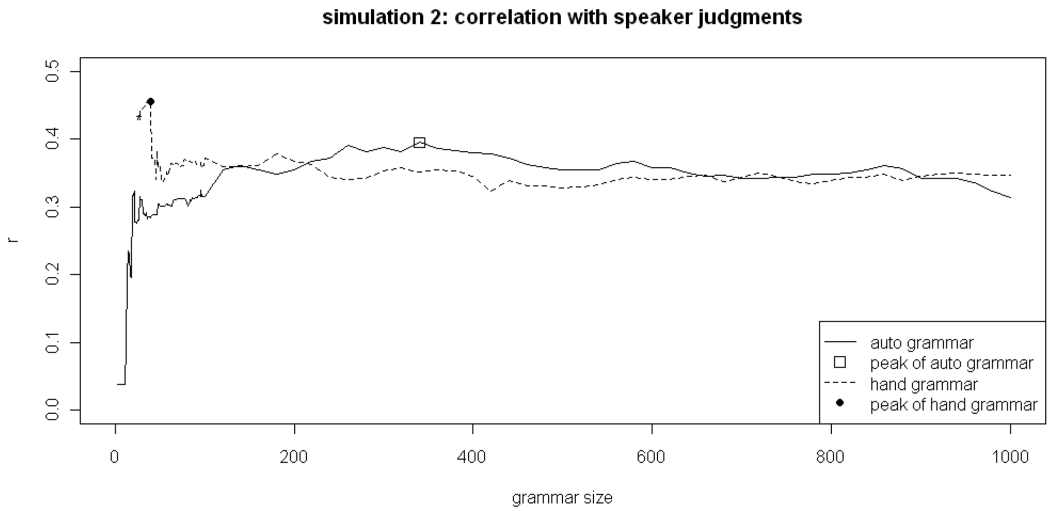


Figure 4.7: Speaker judgment correlation: simulation II

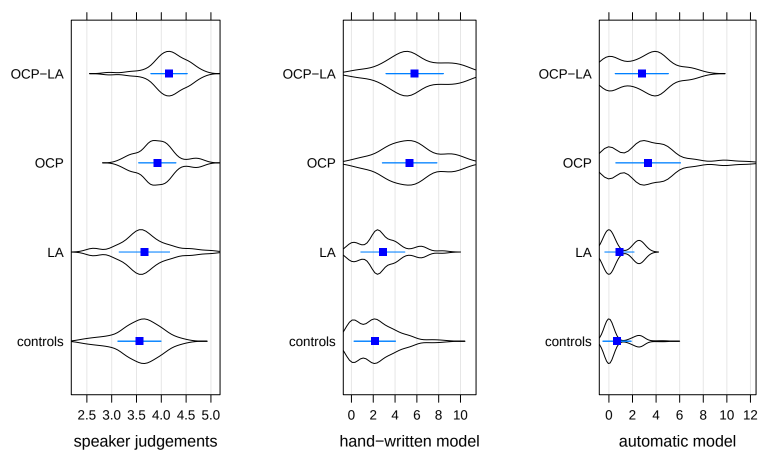


Figure 4.8: Speaker predictions and model performance for OCP-Place

improvement in the general correlation between the best automatic model and the speaker judgements may be a consequence of the modification in the encoding of identical consonants.

Figure 4.9 shows that the automatic model predictions for LA violations are very similar to controls as is the case for the speaker judgements (figure 4.4). This contrasts with the predictions of the automatic model in simulation I where the

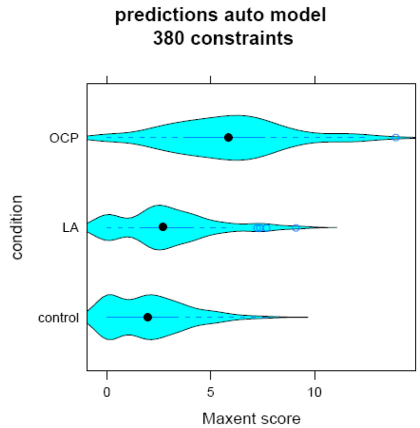


Figure 4.9: Automatic model predictions for LA violations: simulation II

predictions for LA violations were somewhat stretched towards higher (more unfavorable) ratings (figure 4.5). Furthermore, although the hand-written constraints for LA in the most predictive hand-written grammar were assigned weights by the model, their inclusion did not improve model predictiveness compared to simulation I.

Note that we are not claiming that the improvement in the predictions of the modified model is due entirely to eliminating the over-estimation of LA, but that including identical consonants in frequent patterns in the training data (given the available representation) may have consequences for model quality.

We also note that both models quickly acquired the following constraint: *[-sonorant,+continuant][+continuant,+voice] (21st and 42nd constraint for the automatic and hand-written models, respectively). This constraint encodes a general restriction over the co-occurrence of fricatives where the second fricative is voiced, rather than the narrower restriction over homorganic coronal fricatives (s, s', z, ʧ...) that is expected as part of OCP-Place.

The statistical analysis of our verb root database (table 4.9) shows that non-homorganic fricative sequences are effectively under-represented (O/E values < 1) in adjacent position but only when the first fricative is voiceless[8]. Although

[8]This analysis is over triliteral verb roots with the 14 most frequent and evently distributed segments

no such restriction is documented for Amharic, a general restriction against the co-occurrence of fricatives is attested in Chaha, a neighboring language (Banksira, 2000).

Table 4.9: O/E values for coronal/labial fricative sequences

location	C_1C_2X	XC_2C_3	C_1XC_3
fz	0	0.46	0.76
fs	0.78	0.45	1.2
zf	1.64	1.5	1.51
sf	0.53	0.34	1.46

4.3 Discussion

Our simulations suggest that although the OCP-Place restriction is acquired by the automatic models in both simulations, the encoding of identical consonants as they appear in surface realization may be problematic (for example, the root [rtt] is represented with all three consonants although the second occurrence of [t] could be analysed as absent from the surface representation). In some patterns these identical consonants are very frequent and their presence may be reinforcing the otherwise weak LA harmony restriction. It may well be the case that speaker judgements operate over some representation from which identical right edge consonants are absent, or that the process of reduplication does not interact with single feature harmony but, in its current incarnation, the Maxent learner does not provide an appropriate representation.

Our stated goal in this section was to investigate the performance of the Maxent learner on a co-occurrence restriction across whole words, in contrast with the Hayes and Wilson simulations for English onsets. OCP-Place, a statistically robust restriction over verb roots, seemed to be a promising test case for the modelling of a complex restriction. However, although OCP-Place in Semitic languages is well-documented in theoretical and statistical studies as well as experimental work, the learned grammars include heavily weighted constraints for many restrictions other than OCP-Place, such as restrictions over rare and irregularly distributed consonants. There is also evidence for a previously unstudied general restriction over the co-occurrence of fricatives (rather than the usual OCP-Place restriction over coronal fricatives).

A further aspect of the phonotactic grammar of Amharic that we did not evaluate is the acceptability of nonce words with identical consonants in different locations. Based on our hypothesis that the acceptability of a nonce word is a function of the lexical frequency of the component phonotactic patterns, we would expect speakers to have different judgements for nonce words with identical consonants in left edge and non-adjacent patterns (ABA and AAB, for example) which are rare, compared to more frequent patterns such as ABB and ABCC.

The rather inevitable conclusion is that a full evaluation of the performance of the Maxent learner requires experimental data that explores the full range of restrictions that characterize the language rather than our current data which is narrowly focussed on OCP-Place and LA violations.

5 Experiment I: Co-occurrence patterns

5.1 Overview

In the previous chapter, we saw that it is not possible to make a full evaluation of the performance of the Maxent Learner (Hayes and Wilson, 2008) in predicting speaker judgements because our test data (King and Rose, 2003) investigates only OCP-Place and Laryngeal Agreement violations. The goal of the experiments is therefore to obtain judgement data for a broad range of phonotactic restriction types:

- OCP-Place violations
- Non-homorganic fricative sequences.
- Verb roots with identical consonants.
- Distributionally irregular segments

We concluded that a single experiment exploring speaker judgements for those four broad conditions would be uncomfortably long for participants so we decided to break the task into two separate experiments. The first experiment is focused on consonant co-occurrence restrictions, and regroups OCP-Place, co-occurrence of identical consonants and co-occurrence of non-homorganic fricative sequences. The rationale is that these three are co-occurrence restrictions, so they form a natural group on that dimension. It also allows us to compare judgements for nonce words with identical consonants, OCP-Place violations and controls in the same experiment. The fourth condition (distributionally irregular consonants) is reserved for the second experiment.

The choice of conditions for experiment I is motivated by the dictionary analysis of Chapter 3 and the simulations of Chapter 4. To summarize those findings:

- verb roots with OCP-Place violations are under-represented in the lexicon
- verb roots with identical consonants in left edge (AAB) and non-adjacent (ABA, ABAC) patterns are under-represented in the lexicon
- verb roots with adjacent non-homorganic fricatives are under-represented in the lexicon when the first segment is voiceless
- verb roots with identical consonants in right edge patterns compatible with copy or reduplication are over-represented in the lexicon

Collecting speaker judgement data for Amharic poses some difficulties. It was not possible for us to directly collect data in Ethiopia (as King and Rose did) and we considered that the lack of technological infrastructure would make it difficult to obtain data through a web-based experiment targeting speakers living in Ethiopia. For that reason, we collect data through a web-based self-paced acceptability task targeting the diaspora community in the North America. Although this situation is not ideal because 1) our participants do not have the chance to ask questions if they do not understand the nature of the task and 2) they are likely bilingual in English and Amharic, the initial pilot study indicated strong statistical trends in the predicted direction, justifying the use of the methodology.

5.2 Methodology

There are significant differences between our methodology and that used for the King and Rose task (summarized in table 5.1). Our task is presented in an internet based on-line format to speakers who are at least bilingual in English and

Amharic and we assume that speakers are not communicating during the task. King and Rose use a paper and pencil format directly presented to speakers in Ethiopia. As the subjects were educated, they had some knowledge of English but the task instructions, both written and oral, were in Amharic. Several speakers did the task at the same time and were given the opportunity to ask questions about the task with the experimenter before beginning, but they did not communicate between themselves.

Table 5.1: Comparison of methodology

	King and Rose	*Colavin, Levy and Rose*
presentation	paper-and-pencil	online
rating scale	1-6	1-9
randomization	limited	pseudo-randomization
nonce word selection	*see below*	

Our format also made it possible to use pseudo-randomization of experimental items, which was not possible in the King and Rose study. It is because of these methodological differences and to ensure that our test data is consistent across conditions, that we decided to obtain new judgements for OCP-Place violations, as well as the new phonotactic restrictions.

The judgement task of King and Rose contained 270 items and took around 45 - 60mn to complete. As the participants found the task taxing, we decided to limit the number of items in our experiment to 200. The detail of the items is shown in table 5.2.

5.2.1 Novel verb construction

As rare and distributionally irregular consonants are the object of a separate experiment, only the 14 most frequent and evenly distributed segments [m, l, f, t, t', d, s, z, n, l, r, k, k', g] are used. Root length is also likely to affect speaker

Table 5.2: Overview of experimental items

OCP-Place	5 places of articulation (Lab. Dor, Cor-stop,Cor-fric,Cor-son) 3 locations (left edge, right edge, non-adjacent) 3 repetitions	45 items
identical consonants	14 segments 3 locations (left edge,right edge,non-adjacent) - 6 exceptional cases for which are appropriate nonce verbs	36 items
fricative sequences	4 patterns (fs-sf-fz-zf) 3 locations (left edge,right edge,non-adjacent) 3 repetitions	36 items
controls		83 items
Total		200 items

judgements (and should at some point be investigated), so our nonce verbs are all triliteral following the classic Semitic pattern that is also the single most common form in Amharic.

An important goal of our experimental methodology is to ensure that the experimental items encode the target condition with minimal interference from other statistical or phonotactic characteristics. The general procedure in a similar study (Frisch and Zawaydeh, 2001) is as follows:

1) Dictionary analysis. The purpose of the dictionary analysis is to provide a statistical foundation for the creation of experimental items. For Semitic-language word judgement tasks, the dictionary is the lexicon of verb roots[1] and the analysis computes, for each word of the lexicon, those statistical and

[1]For Frisch and Zawaydeh (2001), at least, the analysis is based on triliteral roots only.

phonotactic characteristics that may affect speaker judgements. The list may vary, but would include at least the following:

- positional unigram probability (PP). For a word ‘abc’, PP is defined as
  - P(a in position 1) * P(b in position 2) * P(c in position 3)
- transitional probability (TP). For a word ‘abc’, TP is defined as
  - P(a|#) * P(b|a)* P(c|b) * P(#|C)
- Neighbourhood density (ND) is the number of neighbours for a given word. A neighbour is defined as the number of words that are within a single segment edit distance, through a single substitution, deletion or insertion. For a word ‘abc’, the neighbours are:
  - xbc, axc, abx, xabc, axbc, abxc, abcx, ab, bc, ac.

  This definition of neighbourhood density is broader than that used in Frisch and Zawaydeh (2001) which counts as neighbours only words that share two of the three consonants with the target. However, our lexicon contains such a large proportion of non-triliteral roots that the broader metric seems reasonable.

2) Create the set of all possible non-words with the target phonotactic characteristics.

3) For each nonce word of that set, compute the PP, TP and ND.

4) Select nonce words such that they represent a uniform a distribution over PP, TP and ND.

Although versions of this method have obtained good results in previous studies, we prefer a slightly modified version with a dictionary analysis adapted to the verb root lexicon of Amharic and a method for nonce word selection that recognizes the fundamental difference between controls and stimuli and narrowly defines the necessary characteristics of each.

Dictionary analysis Although our controls and stimuli are all triliteral forms, the lexical analysis is based on the entire lexicon of 4243 verb roots expressed in surface realization. Surface realization differs from the underlying representation in two important aspects.

- Identical consonants in patterns that are assumed to arise through a process of spreading or copying (Gafos, 1999; McCarthy, 1979) are absent from the underlying representation.
- In weak roots, a (specific) consonant absent from the surface realization is assumed to be present in the underlying representation. For the former guttural consonants (but not the glides), it is assumed to be replaced by [a].

Not only do we have a stated preference for avoiding theoretical assumptions where possible, but the use of surface realization rather than underlying representation makes particular sense in this context for two reasons:

- As discussed in chapter 3, the lexicon of Amharic verb roots is very diverse, with a range of root lengths (there are many weak and quadriliteral roots) and roots with identical consonants. Given our assumption that speaker judgements for verb roots are influenced primarily by the lexicon of verb roots, ignoring such a large number of them means that our statistical analysis would be limited to a small subset of the lexicon.
- Our experimental stimuli includes judgements for nonce verbs with identical consonants, so it makes sense to base the statistical analysis on the surface representation where such patterns are present.

In the case of weak roots, as speakers have access to non-root phonological cues to the diachronic presence of a consonant, the consonant missing from SR is replaced by a place marker ”X” to preserve word edge information.

Based on this representation, we compute PP, TP [2]and ND:

- Positional unigram probability (PP) is defined as above.

[2]The probabilities for PP and TP are estimated with relative frequency estimation.

- Transitional probability (TP) For this study, for a root /abc/, TP is defined as:

  - $P(a|\#)*P(b|a)*P(c|b)$

  Note that in computing TP, it is usual to include $P(\#|c)$, the probability of the word boundary given the final segment. We have not done so because there are some differences in right word-edge distributions between triliteral roots and roots with more than three consonants where identical consonants are common.

- Neighborhood density (ND) is defined as above.

Nonce word selection In contrast to similar studies, we make a clear distinction between the selection of controls and stimuli. We define the purpose of controls as a baseline against which the experimental condition stimuli can be compared. The ideal controls are therefore statistically similar to the most statistically representative or *bland* words in the lexicon, words that have none of the phonotactic characteristics that might be assumed to affect speaker judgements. Stimuli, however, are designed to encode phonotactic patterns that are either rare in the lexicon (under-represented) or overly frequent (over-represented) so stimuli should be representative of the experimental condition that they target, rather than the distribution of real words. The full list of experimental items for experiment I is listed in Appendix A.

Controls To select the set of nonce forms that fulfil our criterion of being statistically similar to real words, we proceed as follows:

1. Select all triliteral (because all our nonce forms are triliteral) <u>real</u> words that have none of the phonotactic characteristics that may be assumed to affect speaker judgements in Amharic such as OCP violations, identical consonants, narrow LA violations (voiceless stops differ for the feature CG, constricted

glottis) and fricative sequences.[3] Nasal sequences were also eliminated because nasal assimilation is typologically common between adjacent consonants and in some derivations, root consonants may be in adjacent position.

2. For that set, compute the median and quantiles[4] for ND, TP and PP. We assume that the inter-quartile range (in the 2nd and 3rd quartiles) is most representative of the statistical characteristics of real triliteral verb roots.

3. Create the set of all the possible nonce triliteral roots with no phonotactic violations.

4. Select those nonces for which the ND, PP and TP values fall within the corresponding inter-quartile ND, PP and TP ranges for the real triliteral verb root dataset.

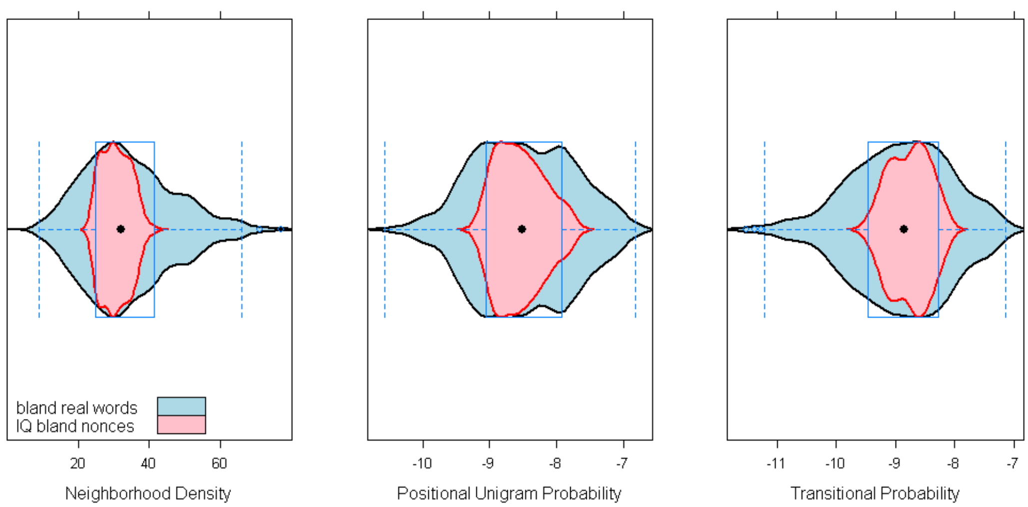


Figure 5.1: Distribution of controls compared to real words

[3]Broad LA violations (stops differing in voice, stops differing in both voice and constricted glottis) violations were not eliminated because 1) to do so would have considerably narrowed the space of possible nonce forms, and 2) the original analysis of the King and Rose task indicated that such violations do not significantly affect speaker judgements.

[4]The inter-quartile range was used rather than the standard deviation because the distribution of the data is often skewed.

In figure 5.1, the set of triliteral real words with no known phonotactic violations is represented in blue. The 96 nonces which fulfil the condition of falling within the interquartile ranges of the real words for ND, TP and PP are in pink. After verification by a native speaker (it is possible that some of the nonces are real words, such as recent slang, that are not included in the dictionary), we select the final 70 controls.

Stimuli For stimuli selection, our goal is to select nonce forms that are statistically representative of the target condition. Ideally, given the entire set of words with OCP-Place violations (both real words and nonces), the selected nonce stimuli would represent equal numbers of high, low and median values of ND, PP and TP for that set. Our preliminary method of stimuli selection was as follows:

1. Find the set of all possible nonce and real words for the condition
2. Compute ND, PP and TP for each word of that set.
3. Define the median values for ND, TP, PP
4. Select nonces such that they are matched with the set of potential experimental items for ND, TP and PP.

However, we found that it is not possible to select a balanced stimuli set with this method. There are several issues but the one that best illustrates the general problem is the distribution of TP. There is experimental work to show that speakers are sensitive to biphones probabilities (Michael S. Vitevitch and Auer, 1999; Vitevitch, 1998; Vitevitch et al., 1997) so it is an important metric. However, TP poses two problems, 1) it is not informative about non-adjacent conditions (such as OCP-Place violations in C_1XC_3) and 2) it varies very widely between conditions. For example, TP values for real words (for triliteral, average TP = -9.13) tend to be higher than those for nonce words (-18.85), left edge OCP-Place violations values tend to be low (-33.65), while values for right edge identical consonants are close to those of real words (-9.83).[5]

[5]Trigram TP would have been a possible solution to this problem, but adding a metric considerably increases the difficulty of selecting a set of nonce forms balanced for all of the metrics under consideration.

A reasonable solution to this problem is to select stimuli by creating all the possible nonce words for each narrowly defined condition, and finding the set of nonces that have TP, ND and PP values that fall within the inter-quartile range *for that set*. Stimuli are then selected from the inter-quartile set. For example, for the condition 'ocp-non-adjacent-labial' (figure 5.2) we create all the possible nonce forms for that condition and compute the ND, TP and PP for each item. We then select stimuli from the subset of nonce forms that have ND, TP and PP values that fall within the interquartile range for that condition. Note that although similar to the procedure for the selection of controls, there is a crucial difference; controls are selected based on the interquartile range for NP, TP and PP of real words whereas stimuli are selected on the interquartile range for ND, TP and PP of the set of available nonces.

This method also solves a similar problem with neighbourhood density. As discussed in Chapter 2, real words (and presumably nonce words with no phonotactic violations) tend to inhabit denser neighbourhoods than nonce words with under-represented sequences. This is certainly true for our dataset. For example, real words have an average of 28 neighbours whereas nonce words with left edge identical consonants usually have 0 neighbours.

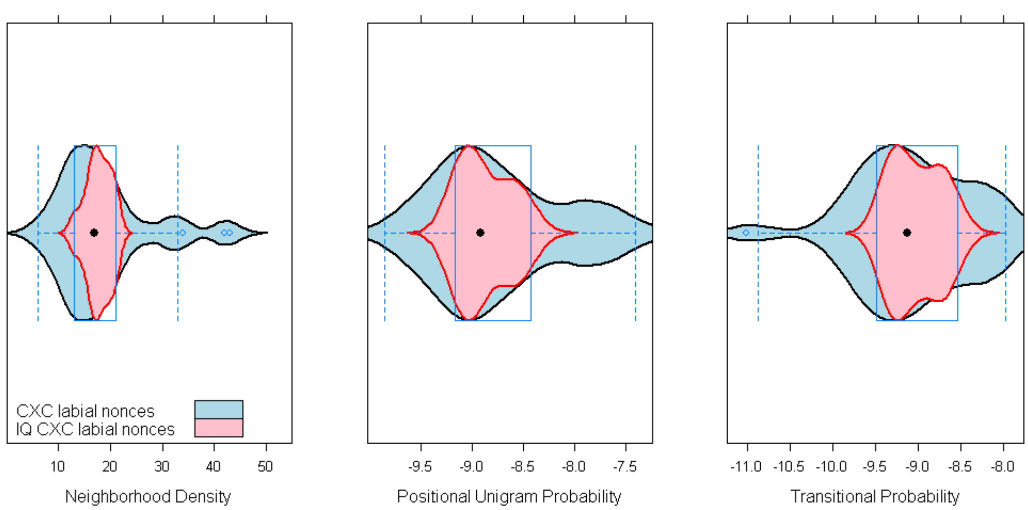


Figure 5.2: Selection of nonce forms for the ocp-labial-left edge condition

To summarize, our search for a motivated and consistent method of stimuli selection has led us to select stimuli that are statistically representative (fall within the interquartile range for PP, TP and ND) of the nonces available for that condition.

The case might be made that our method is deficient on the grounds that our stimuli encode the average statistical values for each condition and that they represent only a narrow sample of the possible statistical variation for a given condition. A possible counter-argument is that our target conditions are more narrowly defined than in similar studies (for example, the stimuli for OCP-Place are broken down by location of violation and by POA, rather than just by location of violation) in a way that effectively represents the gradiency of the restriction.

There were some cases where even this refined method of stimuli selection was not satisfactory. For a few conditions over adjacent consonants, the values of TP were bi-modal and heavily skewed such that choosing all the stimuli from the IQ range was not possible. In those cases, we hand selected the stimuli according to our best judgement to represent the condition, given the available set of nonces.

OCP-Place For each of the 15 specific conditions (3 locations * 5 POA), we followed the general procedure described above. However, in several conditions, there were insufficient nonces fulfilling the requirement of being in the inter-quartile range for ND, PP and TP. This problem arose for patterns that are extremely rare in the lexicon, such as dorsals in the left edge location where the values for TP are extremely skewed. In those cases, we selected nonces that were in the inter-quartile range for ND and PP only.

Identical consonants The general procedure was used to select left edge and non-adjacent reduplication stimuli for each of the 14 segments, as these are strongly under-represented. However, right edge identical consonants are so frequent in the lexicon that there are very few nonce forms available, particularly as we are using only the most frequent segments. Stimuli for the right edge condition were therefore directly hand selected from the list of all nonce forms with right edge reduplication but no other distinctive characteristics. In the case of [b, l, m, n, r], there are no nonce forms available that do not also have an some other phonotactic particularity (for example, [nrr] and [lrr] violate OCP-Place). For this reason, the final stimuli set for the right edge reduplication condition contains only 9 stimuli, instead of 14.

Fricative sequences Fricative sequences were selected with the general procedure: for each of the 12 specific conditions (4 patterns: f-s, s-f, f-z, z-f * 3 locations: left edge, right edge, non-adjacent), we first created nonce forms that fulfill the condition (for example, left edge-s-f) and computed the ND, TP and PP interquartile ranges. We then determined the subset of nonces for which the ND, TP and PP values fall in the inter-quartile range. Each condition is represented by three items, so three stimuli were then selected from that inter-quartile set, such that the consonants not involved in the fricative sequence present some variation over manner and POA.

5.2.2 Presentation

As the participants for the study are geographically dispersed, we developed an internet accessible version using the Ibex software for running self-paced acceptability judgement tasks online through a web browser[6]. Nonce words are presented in 3ms perfective (dictionary citation) form which is CəCCəCə, where the 2nd consonant is geminated. Although gemination is not conveyed in the Amharic writing system, this is not problematic since subjects know which form of the verb the data are presented in. The perfective form does not distinguish between Type A and Type B, and all nonce forms will therefore be uniform in their prosodic shape.

The experiment is divided into 5 parts:

- The first screens are in English and are used to obtain informed consent for the experiment and to collect some information on the participant's language background (where the participant learned to speak Amharic, what other languages are spoken, and countries where the participant has lived). This

[6]http://code.google.com/p/webspr/.

section also collects a password that participants use to claim their payment. This is the only section presented in English. All subsequent screens are in Amharic.

- To ensure that participants do know Amharic, they are presented with 3 real words of Amharic and 3 invented words in a randomized sequence and are asked to indicate which ones are real words by pressing 'yes' and 'no' buttons.

- Instructions: There are several screens explaining that the goal is to assign a grade to invented verbs presented in 3ms perfective (dictionary citation) form. Participants are instructed to assign the grade based on whether they think that it would be a good word for Amharic.

- Training: Training consists of a list of 19 nonce verb roots that are representative of the conditions and controls of experimental items. The goal is to allow the participants to become familiar with the use of the 9 point scale. Each training item appears on the screen above square buttons numbered 1-9. To the left of the '1' button, is the Amharic term for 'good'. To the right of the '9' button appears the Amharic expression for 'not good'. Below the button scale appears the Amharic expression for 'use the number keys to respond'. Each item remains on the screen until the participant presses a number key. The item is replaced by a screen that displays a large red star. The star remains on the screen for 500ms before a new item appears. This was done to ensure that subjects noticed the change to the next test item. The training items are pseudo-randomized but are presented in the same order across participants.

- Main task: The presentation of the 200 experimental items is similar to that of the training task. The items are pseudo-randomized between subjects such that a specific experimental condition does not appear more than twice in a row. The task is divided into 4 blocks with a rest period between each. The participant determines the length of the rest by pressing a button when he or she is ready to continue to the next block).

A sample rating screen and part of the Amharic experiment instructions are shown in Appendices A and B respectively.

5.2.3 Participants

The targeted participants are native speakers of Amharic aged 18 years or older who have access to a computer and internet connection. Given the nature of the task and the necessity of giving limited instructions (to avoid biasing the results), we expected that some participants would misunderstand the nature of the task. Feedback (via email) from 5 participants showed that this was indeed the case. For example, one participant thought that there was some connection between acceptability and the possibility that the word might be used in a poem and others thought that the task was to determine if the experimental items were real words. These misunderstandings produced unary or binary ratings. Unary ratings do not have a clear interpretation given the experiment instructions and, although Frisch et al. (2004) showed that averaged binary ratings are strongly correlated with gradient ratings, our subjects were instructed and trained to use the full rating scale. As subjects are not given a means of explaining their response strategies, we cannot know on what basis binary ratings were assigned. For this reason, unary and binary ratings are considered uninterpretable and are not included in the analysis. 3-ratings were considered to be an attempt to use the scale as instructed and were included in the analysis. Participants were primarily from the USA and Canada, but two were from Ethiopia, contacted by family in the USA. Participants receive payment of $25 paid by check or Western Union money transfer for those outside the USA.

5.3 Model predictions for Experiment I

The choice of a reference model to use for predicting speaker ratings is not straightforward. The representational system used in simulation II poses problems for modelling words with identical consonants. In the training data for that simulation, occurrences of a consonant in a word after the first one (starting from the left edge of the word) are replaced by a single place marker 'X'. This representation effectively means that the model sees no examples of words with identical consonants in any pattern. Presumably, as identical consonants do not occur in the learning data, experimental items with identical consonants in surface realization would receive a heavy penalty. Alternatively, we could encode the experimental items with 'X' in the place of identical consonants in the same way as for the learning data. Presumably, the simulation II model learned that 'X' in C_2 is rare (AAB forms) but 'X' in the case of 'X' in C_3 is ambiguous because 'X' as a copy of C_1 (ABA) is rare but 'X' as copy of C_2 is frequent (ABB). For this reason, we compare speaker ratings to predictions of the automatic model created in simulation I where the identical consonants in the training data are encoded in surface realization. We determined that the most logical grammar size to use to model the judgements is the one with the highest log-likelihood (360 constraints, see figure 4.1 on page 73) based our assumption that speaker judgements are a function of $L_\theta(D)$.

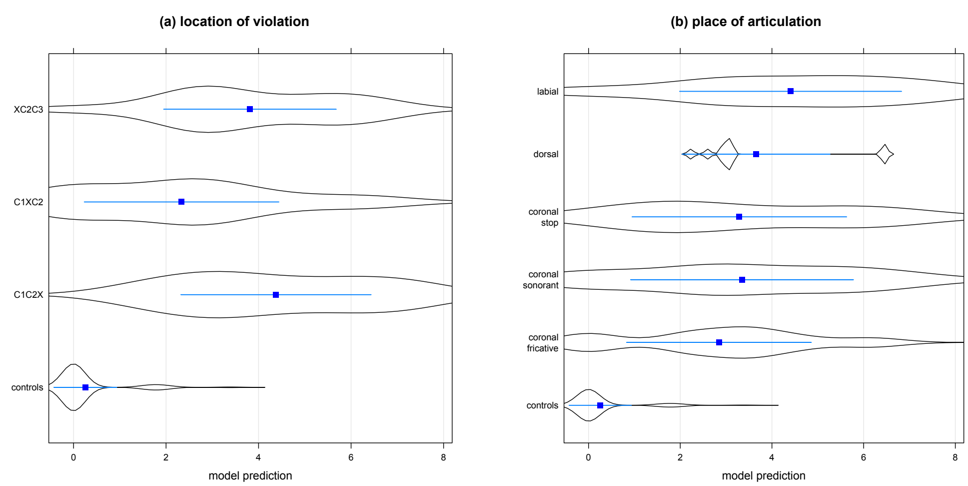


Figure 5.3: Model predictions for OCP-Place violations

OCP-Place Figure 5.3 shows the predicted weights for the stimuli with OCP-Place violations. (a) shows that all locations of violation are predicted to be rated as less wordlike than controls, though non-adjacent violations are predicted to be mildly less so than adjacent violations at either word edge. (b) shows that OCP-Place violations are predicted to be rated worse than controls for all POA, though slightly less so for some coronals.

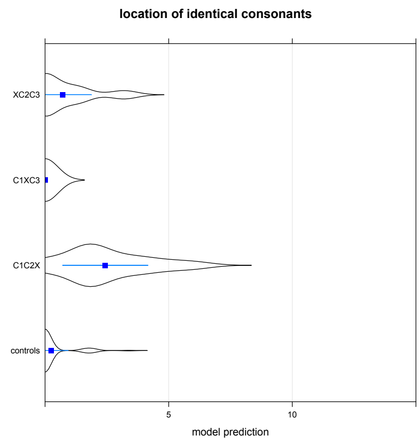


Figure 5.4: Model predictions for identical consonants

Identical consonants Figure 5.4 shows the reference model predictions for patterns of identical consonants according to location. Left edge (C1C2X) patterns, which are rare in the lexicon, are correctly predicted to be worse than controls but right edge patterns (XC2C3) are predicted to be slightly worse than controls although, given their frequency, we might expect them to be more acceptable than controls. Finally, non-adjacent (C1XC3) patterns are predicted to be rated as similar to controls, which is surprising given how rare they are in the lexicon. This may be the combined result of the generalization algorithm which selects, all else being equal, constraints over natural classes with few features over those with many,[7] and short constraints over long ones.

Non-homorganic fricative sequences Regarding non-homorganic fricative sequences, the statistical analysis (chapter 4) indicated that only adjacent non-homorganic fricatives where the first segment is voiceless are under-represented.

[7]As there is no mechanism for generalizing over identical consonants, a constraint restricting the co-occurrence of identical consonants would have to use enough features to fully specify each consonant.

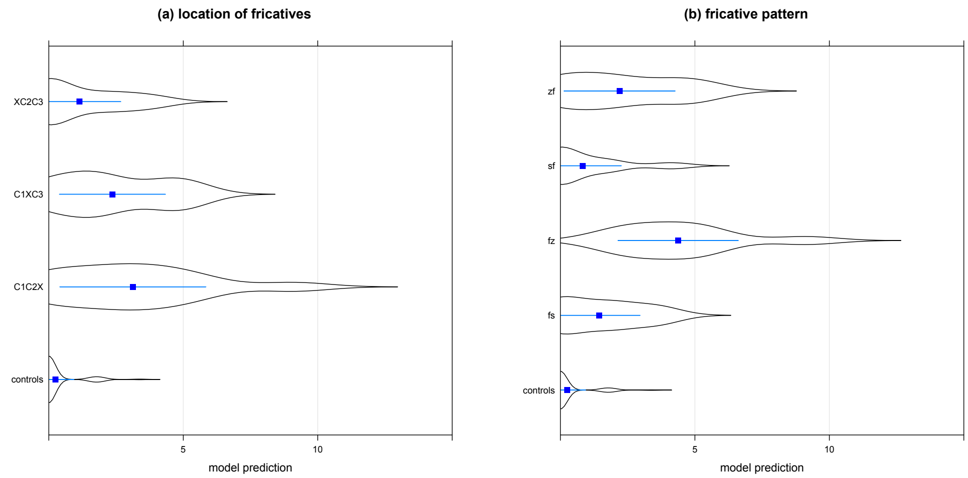


Figure 5.5: Model predictions for non-homorganic fricative sequences

However, figure 5.5(a) shows that all non-homorganic f, s, z sequences regardless of pattern and location are predicted to be dis-preferred compared to controls with a stronger dis-preference for sequences containing [z].

5.4 Experimental results

Based on Baayen et al. (2008),we use maximal LMEMs (linear mixed effect models) with by-subject random slope for the fixed effect, and by-item random intercepts for each nonce word, using the lmer function of the statistical software R (package lme4). To estimate the p-value of a fixed effect, we compute the likelihood ratio between two models identical with regard to random effects but only one of which contains the fixed effect of interest. For a p-value $< .05$, factors with $|t|$-values > 2 are assumed to be significant contributors to the effect. To clarify the procedure, we explicitly detail the analysis for OCP-Place violations according to location of violation.

OCP-Place Using a subset of the experimental data that contains only controls and stimuli with OCP-Place violations, we create two models. :

```
>m1 <- lmer(rating ~ location + (location |subject_number) +
```

```
(1 | stimuli_number), ocp_and_controls,REML=F)
>m0 <- lmer(rating ~ 1 + (place | subject_number) +
(1 | stimuli_number), ocp_and_controls,REML=F)
```

The models m0 and m1 are identical except that m1 has a fixed effect for location with "controls" as the baseline level of the factor. We compare the models with the anova function of R.

```
>ocp_and_controls_anova<-anova(m1,m0)
>ocp_and_controls_anova
Data: ocp_and_controls
Models:
m0: rating ~ 1 + (location | subject_number) + (1 | stimuli_number)
m1: rating ~ location + (location | subject_number) + (1 | stimuli_number)
   Df   AIC   BIC  logLik  Chisq Chi Df Pr(>Chisq)
m0 13 11399 11475 -5686.6
m1 16 11378 11472 -5673.2 26.788      3  6.522e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see from this that the difference between the models is significant with $p < .001$. Based on this, we examine the t-values (which refer to the comparison with the baseline factor) for the factors of m1, the model with the fixed effect.

```
> m1
Linear mixed model fit by maximum likelihood
...
Fixed effects:
              Estimate Std. Error t value
(Intercept)     4.1988     0.2964  14.166
locationleft    1.4912     0.3092   4.823
locationnon     0.8112     0.2567   3.160
locationright   1.8712     0.2869   6.522
...
```

In this example, the baseline level is the control condition and the other conditions are evaluated in comparison to that baseline. For the non-baseline levels, the $|t|$-values > 2 show that all locations of violation are significant contributors to the effect (left edge: 4.823, right edge: 6.52, non-adjacent: 3.160). This is shown graphically in figure 5.6(a).

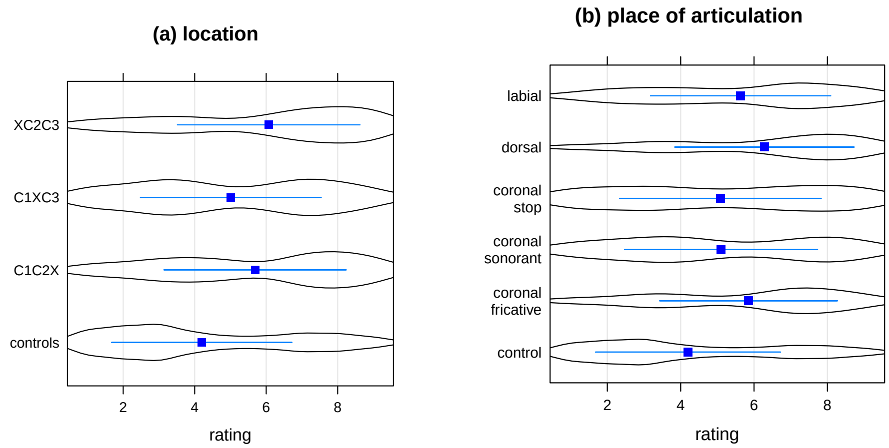


Figure 5.6: Speaker ratings for OCP-Place

The analysis of OCP-Place violations in isolation (so without the controls) with non-adjacent violations as the baseline factor shows that left edge and right edge violations are rated significantly worse than non-adjacent nonces ($p < .01$, t-values: C_1C_2X -2.16, XC_2C_3 3.3).

The analysis of OCP-Place violations and controls according to place of articulation, figure 5.6(b), shows that place of articulation is significant ($p < .0001$) and that all places of articulation are contributors to the effect (t-values: dorsals 5.144, coronal fricatives 4.936, labials 4.632, coronal sonorants 2.761, coronal stops 2.827). For OCP-Place violations in isolation with coronal sonorants as the baseline factor (because these received the lowest ratings), place of articulation is not a significant fixed effect ($p = .07$). However, the t-values show that the difference between coronal sonorant violations and dorsal violations (which receive the highest ratings) is significant (t-values: coronal fricatives 1.67, coronal stops -0.04, labials 1.25, dorsals 2.69).

Identical consonants The analysis of identical consonants and controls (figure 5.7) shows that location is a significant fixed effect ($p < .01$) but only left edge C_1C_2X (AAB) and non-adjacent C_1XC_3 (ABA) are significant contributors to the model (t-values: C_1C_2X (AAB) 4.097, C_1XC_3 (ABA) 2.627, XC_2C_3 (ABB)

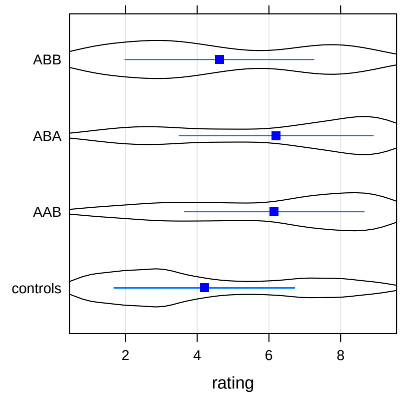


Figure 5.7: Speaker ratings for identical consonants

1.237). The analysis of ratings for stimuli with identical consonants (without controls) according to location with the XC_2C_3 (ABB) location as the baseline factor (because these are rated lowest) shows that location is significant ($p < .01$) and t-values show that XC_2C_3 (ABB) patterns are significantly better than C_1C_2X (AAB) and C_1XC_3 (ABA) (t-values: C_1C_2X (AAB) = 3.7; C_1XC_3 (ABA) = 3.3).

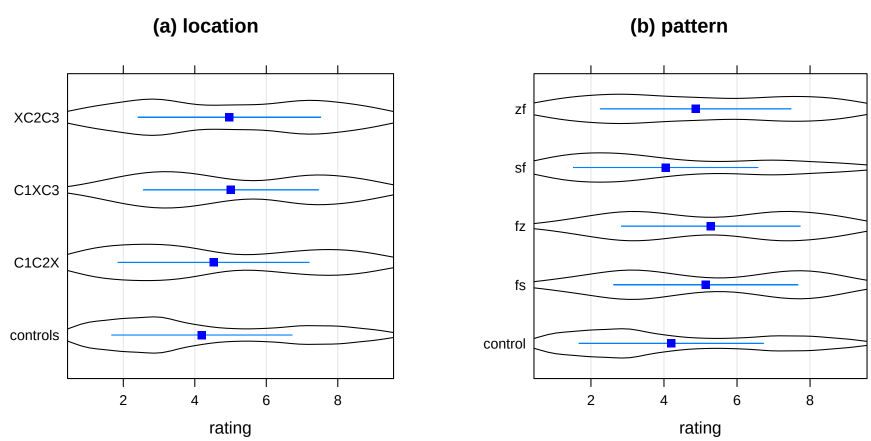


Figure 5.8: Speaker ratings for non-homorganic fricatives

Non-homorganic fricative sequences The location of the pattern is a significant contributor to the model ($p < .001$) (figure 5.8(a)). Right edge and non-adjacent sequences are rated significantly worse than controls (t-values: C_1C_2X 3.4, C_1XC_3 3.5, respectively) but this not the case for left edge sequences (t-value: XC_2C_3 1.4). Model comparison for non-homorganic fricatives without the controls according to location (and the left edge location as the baseline factor), shows that location is not a significant effect (p = .2) and the t-values show that neither non-adjacent nor right edge patterns are significantly different to left edge patterns (t-values: XC_2C_3 1.3, C_1XC_3 1.4).

Figure 5.8(b) shows the results for non-homorganic fricatives according to pattern. The comparison of a model with controls and non-homorganic fricatives according to pattern shows that the pattern is a significant effect ($p < .001$) but only 'fs', 'fz' and 'zf' sequences are significantly different to controls (t-values: 'fs' 3.8, 'fz' 4.1, 'zf' 2.6). Ratings for 'sf' sequences are not rated significantly differently to controls (t = -0.6). Model comparison for non-homorganic fricatives without the controls according to pattern (and the 'sf' pattern as the baseline factor), shows that pattern is a significant effect ($p < .01$). t-values show that all other patterns are significantly different to 'sf' patterns (t-values: 'fs' 3.1, 'fz' 3.7, 'zf' 2.4).

5.5 Discussion

5.5.1 Speaker ratings

The results of the judgement task show that the model predictions for under-represented sequences generally confirm the hypothesis that low-frequency patterns are dis-preferred by speakers. However, the relationship between lexical frequency and speaker judgements for ABB nonces, if such a relationship exists, is not straightforward.

OCP-Place violations (regardless of POA and location of violation) are always dis-preferred compared to controls. OCP-Place is generally described as gradient, where the strength of the restriction is a function of POA and location of violation[8] such that:

POA	bad	$coronal \rightarrow labial \rightarrow dorsal$	worst
location	bad	$CXC \rightarrow CCX \rightarrow XCC$	worst

Our results show that 1) adjacent violations are rated worse than non-adjacent 2) there is no difference between the two kinds of adjacent (C_1C_2X and XC_2C_3) and 3) that coronal-sonorant violations are less dis-preferred than dorsal violations. It may be that our experimental task is not sensitive enough to capture gradiency but these results also raise interesting questions regarding the frequency differential required for speakers to rate two conditions as significantly different.

For identical consonants, the model predicts AAB patterns to be worse than controls (and this is in line with the dictionary analysis), ABA forms (which are rare) to be rated on a par with controls and ABB forms (which are very frequent) to be slightly worse than controls. Speakers rate both ABA and AAB nonces as significantly less acceptable than controls , but ABB sequences are rated no differently to controls. In fact, the trend seems to be that they are slightly less acceptable than controls. This result indicates that speakers are not judging ABB sequences on frequency alone (see the high O/E values for ABB patterns shown in table 3.14). It has been noted that there may be methodological issues in the elicitation of speaker judgements for both high and low frequency patterns within the same task as the subtle differences in acceptablity between normally and over-represented patterns are likely to be compressed by the presence of strongly un-acceptable forms (Albright, 2011b). On the other hand, these ABB forms are not judged to be significantly less acceptable than controls as was the case for in a very similar study for Hebrew (Berent and Shimron, 2003) where speakers were asked to rate (on a 5 point scale) the acceptability of nonce roots with either right edge identical consonants, right edge homorganic (non-identical) consonants or neither. The results of that study could be accounted for by assuming that ABB (underlying AB) roots are rated lower because of their morphological complexity (Bat-El, 2006) base on the assumption that, all else being equal, morphological complexity

[8]This is confirmed for Amharic by the dictionary study in Chapter 3.

influences speakers to judge a word as less acceptable. However, morphological complexity is presumably similar in both Hebrew and Amharic.

Table 5.3: Comparison of Hebrew and Amharic triliteral roots with identical consonants

	Hebrew	Amharic
triliterals	1449	1872
AAB	4 (.27%)	57 (2.5%)
ABA	21 (1.44%)	16 (0.7%)
ABB	140 (9.66%)	320 (14.12%)

A possible analysis is that the shape of the lexicon distinguishes the two languages. The Hebrew database used by Berent et al. (2012) and Berent and Shimron (1997) contains 1449 triliteral roots and includes verb roots with identical consonants in all locations. With regard to weak roots, they are either absent from the database, or the consonant missing from the surface realisation has been replaced. We compared the Hebrew database to the set of triliteral Amharic verb roots including those with identical consonants and without weak roots[9] Table 5.3 shows that ABA and AAB patterns are rare in both databases and that the percentage of ABB forms is larger for Amharic than for Hebrew[10] (14.12% and 9.66% respectively). Such a trend could explain the difference in acceptability of these roots in the two languages, but a comparison of the full lexicon of verb roots for both languages would be desirable.

As the model predicts, non-homorganic fricative sequences are generally dis-preferred compared to controls. However, the statistical analysis reported in Chapter 4 predicted that only adjacent sequences that start with a voiceless fricative would be dispreferred whereas the speaker ratings show only ‘sf’ patterns

[9]The construction of the Hebrew database implies a preference for surface realization, so we assume that weak roots have been removed.

[10]It would be good to have the equivalent data for Arabic, but verb roots with identical consonants, in any location, are not included in the database made available to me by Stefan Frisch

are not rated as significantly worse than controls and that both right edge and non-adjacent are dis-preferred compared to controls.

In fact, the dictionary analysis reported in Chapter 4 was based on the set of triliteral verb roots and evaluated only the fricatives ‘f’, ‘s’ and ‘z’. We made this choice because the analysis of the entire database that included all non-homorganic fricatives did not reveal any evidence that non-homorganic fricative sequences are generally under-represented (O/E values were generally close to 1). The alternate database was chosen because our nonce forms are all triliterals and contain no fricatives other than ‘f’, ‘s’ and ‘z’.

Given the experimental results, a possible hypothesis is that the restriction is motivated by statistical patterns present in the larger lexicon. The hypothesis that speakers generalize across locations and patterns and cannot distinguish between them appears unlikely in this case because there is no apparent reason that ‘sf’ sequences would be less dis-preferred than other patterns. Note that as ‘sf’ is rated as acceptable but ‘fs’ is not, it does not appear that there is a type of laryngeal agreement operating among fricatives, in the same way as laryngeal agreement operates among stops.

5.5.2 Simulations

We will now compare the reference model predictions to speaker judgements on an item by item level. In the simulations using the King and Rose (2003) test data, the best correlation between the speaker ratings and the model predictions was r = .45 for the best hand-written grammar and r = .39 for the best automatic model. The correlation between the speakers ratings of Experiment I and the reference model (see discussion in section 5.3) predictions is r = .49. This improvement indicates that our experimental methodology, in spite of the shortcomings due to the online presentation, is successful.

Figure 5.9 shows the relationship between speaker ratings and the reference model (r = .49). Within each category (figure 5.10), the model performs best on OCP-Place violations (r = .48) and identical consonants (r = .40), but less well on non-homorganic fricatives (r = .317) and controls (r = .27). Note that although

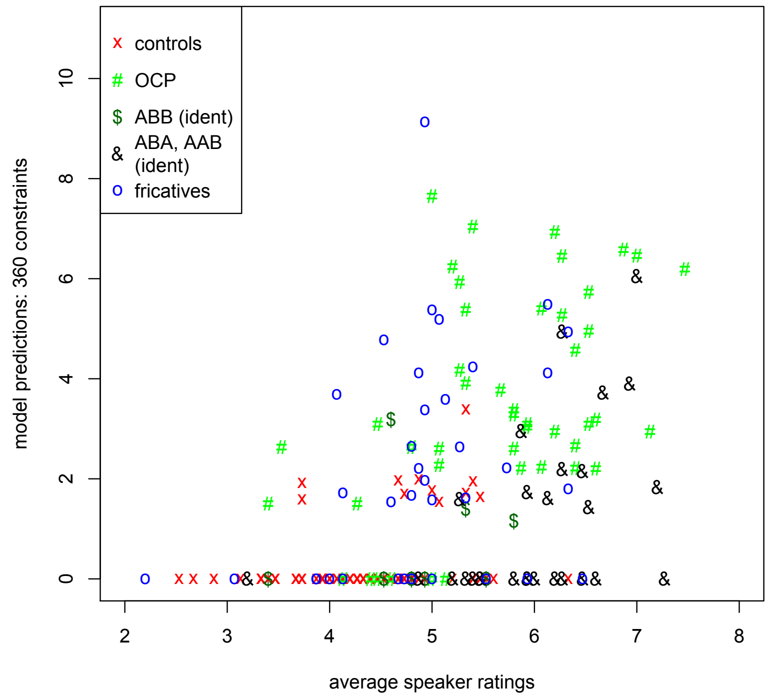


Figure 5.9: Comparison of speaker ratings and model predictions for OCP-Place patterns (homorganic, non-identical consonants), rare patterns of identical consonants (ABA and AAB), frequent patterns of identical consonants (ABB) and non-homorganic fricatives

the correlation between model predictions and controls is low, including them into the model always improves the correlation. For example, the correlation between model predictions and speaker judgements for OCP-Place violations *and* controls is r = .65. This is compatible with Albright's (2009) analysis that the Maxent learner's excellent performance on English onsets is in part based on the binary separation between under-represented and non-under-represented patterns.

5.6 Conclusion

The results of this first experiment show that the frequency∼acceptability hypothesis holds for restrictions over the co-occurrence of consonants, such as

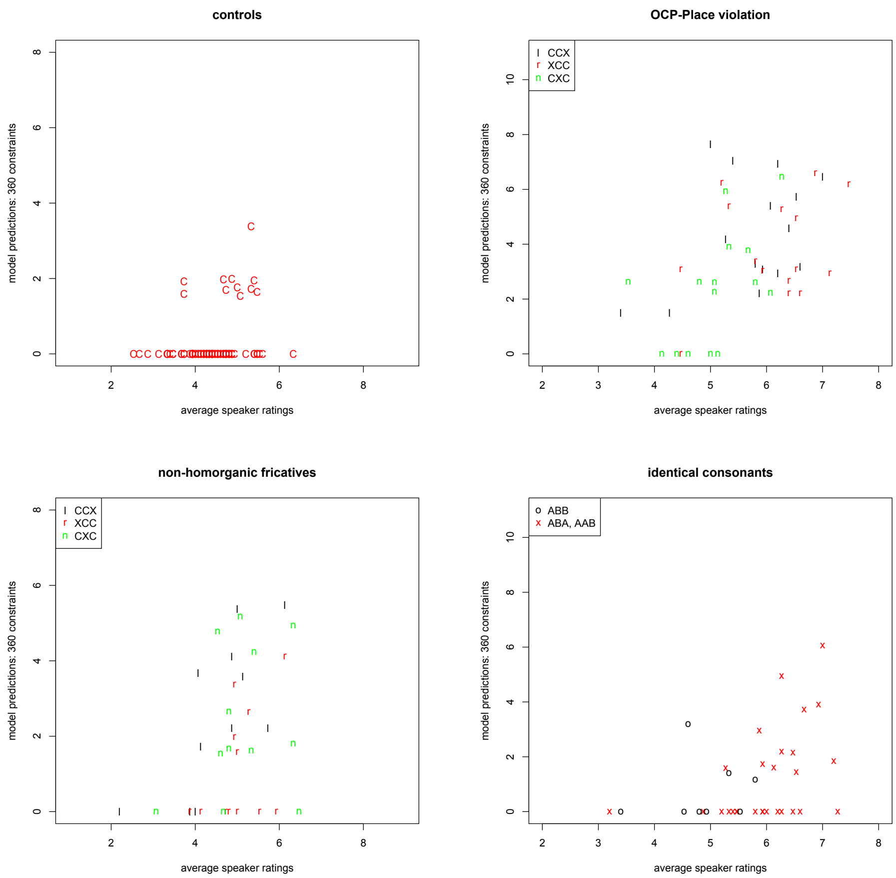


Figure 5.10: Comparison of speaker ratings and model predictions

OCP-Place and identical consonants in rare patterns (AAB, ABA). Speakers also showed a dis-preference for non-homorganic fricative sequences and this is a first step in showing that the model can be used to discover previously unknown constraints. Although speakers did not show a preference for nonce words with ABB patterns of identical consonants over controls, this may not be a failure of the frequency˜acceptability hypothesis for over-represented patterns because the role of morphology in these judgements is undetermined.

One of the goals of this study is to determine whether the Maxent learner (or equivalent models) can be used to approximate speaker judgements in phonological research. Experiment I shows that the model performs acceptably in predicting judgements for nonce words with under-represented patterns, but although the model performs acceptably in predicting ratings for nonce words with identical consonants in ABB patterns relative to rarer patterns, both the experimental results and model predictions are unexpected and require further investigation. The simulations also predict penalties for rare and distributionally irregular segments and the the goal of the second experiment will examine the well-foundedness of those predictions.

The second experiment, which elicits judgements for words with under-represented segments, over-represented segments and irregularly distributed segments (segments which have an affinity or avoidance for a particular location) is designed specifically to investigate a range of frequencies, including nonce forms with [r], the most frequent (and over-represented) segment in Amharic verb roots. The stimuli for [r] are designed to test the frequency˜acceptability hypothesis for the over-representation condition where morphology is not a confound.

Phonologists are also interested in the relative strength of different restrictions and the discussion of the model performance in predicting relative strength (rather than a determination of whether a sequence is generally rated as more or less wordlike than controls) is reserved for the full analysis of the results from both experiments.

6 Experiment II: Segmental irregularities

6.1 Overview

The goal of experiment I was to investigate speaker ratings for patterns of consonantal co-occurrence. The results of that experiment show that speakers assign lower acceptability ratings to nonce words with patterns of under-represented consonants such as OCP-Place violations, non-homorganic fricative sequences and identical consonants in rare patterns (ABA, AAB). The results for identical consonants in the frequent ABB pattern did not show a speaker preference for over-represented patterns but, given the possible confound of morphological complexity, it is not clear whether this is effectively a failure of the frequency $\sim$ acceptability hypothesis for over-represented patterns.

The goal of experiment II is to investigate speaker ratings of nonce stimuli with segments of differing levels of over- and under-representation, allowing us to determine 1) if, absent the confound of morphological complexity, nonce forms with sound patterns that are over-represented are rated as more wordlike than controls and 2) if it is possible to model speaker judgements of non-under-represented items with a model such as the Maxent learner than assigns only positive weights, and 3) if there is a statistical threshold for speaker sensitivity to phonotactic patterns.

The decision to attempt to find evidence for speaker preference of over-represented sequences has two consequences for the experimental design:

- The most frequent segments in the lexicon of Amharic verb roots are [r] and

[t]. These segments were part of the set of 14 most frequent and evenly distributed segments used to create nonce forms for experiment I. For experiment II, [r] and [t] are part of the experimental conditions so the set of segments used to create controls is reduced to twelve: (m, l, f, t', d, s, z, n, l, k, k', g).

- Neighbourhood density may be a confound in rating tasks that include nonce words with under-represented segments in the same word acceptability task as nonce words with over-represented segments. The argument, developed by Shademan (2007), is based on the comparison of results from rating tasks containing both real and nonce words. Shademan hypothesizes that the presence of real words (which tend to inhabit denser neighbourhoods than nonce words) activates analogical processing (evaluation by comparison with words in the lexicon) rather than probabilistic processing. Similarly, nonce words with high frequency patterns tend to have more neighbours than those with rare patterns and, in rating tasks containing both, probabilistic processing might be masked for nonce words with high frequency patterns. By evaluating speaker judgements for nonce words with high-frequency patterns in a separate task, the neighbourhood density effects are kept constant and the effects of phonotactic probability, though possibly muted, should be discernible because speakers can express them over the full range of the scale. For this reason, experiment II is divided into two parts. Part A investigates judgements for nonce words with segments that are not generally under-represented, and part B explores judgements for segments representing a broad range of segmental frequencies, from strongly under-represented to mildly over-represented. The rationale for including non-under-represented segments in part B is that it allows us, at no cost, to evaluate the hypothesis that nonce words from sparse neighbourhoods should not be evaluated in the same task as nonce words from dense neighbourhoods.

In experiment I, the choice of experimental conditions was motivated by O/E values. Recall that O/E is a measure of how frequently two segments (or natural classes) co-occur, given their individual frequencies. We determined that, according to the metric, non-identical homorganic consonants, non-homorganic fricative sequences and identical consonants in rare patterns AAB and ABA are under-represented in the lexicon of verb roots but that ABB patterns of identical consonants are over-represented. For experiment II, we use three metrics (equation 6.1), OES_{word}, evaluates the general lexical frequency of a segment type, given the number of segment types. For example, [s'] is generally under-represented in verb roots (OES_{word}(s') = .15). $OES_{location}$ evaluates the frequency of a segment type in a specific location, given the number of segment types possible in that location. $OES_{location}(\text{s'}_{C1}) = .2$, so [s'] is also under-represented in the C_1 location. $OES_{relative}$ evaluates the frequency of a segment type in a specific location, given the frequency of that segment type. $OES_{relative}(\text{s'}_{C1}) = 1.27$, so although [s'] is both generally under-represented and under-represented in position C_1, it is not under-represented in C_1 relative to its frequency in other word locations. Note that for each of these metrics, as for O/E, values below 1 indicate under-representation and values above 1 indicate over-representation.

$$
\begin{aligned}
&\text{OES}_{word}(x) = \frac{\text{count of segment x in corpus}}{\frac{\text{total segments in corpus}}{\text{size of segment inventory}}}\\
&\textit{and}\\
&\text{OES}_{location}(x_m) = \frac{\text{count of segments x in location m}}{\frac{\text{total segments in position m}}{\text{size of segment inventory}}}\\
&\textit{and}\\
&\text{OES}_{relative}(x_m) = \frac{\text{count of segments x in location m}}{\frac{\text{total number of segments x in corpus}}{\text{number of positions in word}}}
\end{aligned}
\tag{6.1}
$$

As the goal of experiment II is to investigate ratings for triliteral nonce verb roots containing one segment that has an unusual distribution, we began by evaluating all three metrics for the C_1, C_2 and C_3 positions that are represented in the experimental items. However, it was not obvious how to evaluate O/E values for the C_3 position. One solution would be to assume that the C_3 position corresponds to the final segment of roots, regardless of length. In that case, the C_2 consonant could be equated either to second root consonant (the actual C_2

of triliterals) or the penultimate consonant ($C_{root\_length-1}$) for longer roots. For the co-occurrence patterns in Chapter 5, we computed O/E values over several versions of the lexicon of verb roots because OCP and OCP-Place restrictions are assumed to be active primarily over verb roots. However, there is no particular reason to assume that judgements for nonce roots with segmental irregularities are not influenced by the statistics of the non-verb root lexicon. For this reason, we based the statistical analysis for experiment II on the set of all triliteral non-reduplicative roots[1], assuming that, with regards to segmental frequencies, non-reduplicative verb roots are statistically representative of words generally. Given this broad picture, we first selected segments that are over-represented, normally-represented and under-represented. Within those broad categories, we attempted to find segments that were unevenly distributed over the three possible locations in the word (C_1, C_2, C_3). For example, we chose [w] as a condition because its OES_{word} value is close to 1 (.94), indicating that it appears in verb roots as often as would be expected, but it is rare in position C_3 both compared to other segments ($OES_{location}$ = .01) and compared to how often it occurs in other locations ($OES_{relative}$ = .02).

Table 6.1 shows the distribution of the segments chosen as experimental conditions and the statistics that motivated their selection. The conditions that are of special interest to us are highlighted in gray. Both [s'] and [ʤ] are globally under-represented but [ʤ] is particularly rare in C_3 ($OES_{location}$ = .04, $OES_{relative}$ = .25). [w] and [t] are not generally under-represented but [w] is very under-represented in C_3 position and [t] is moderately under-represented in C_1 and C_2. [r], although generally over-represented, is moderately rare in C_1 position.

Table 6.1: Segmental frequencies

condition	segment	count (OES$_{word}$)	positional count (OES$_{location}$) (OES$_{relative}$)		
			C1	C2	C3
globally rare in all positions	[s’]	33 (.15)	15 (.2) (1.27)	10 (.14) (1.1)	8 (.11) (.72)
globally rare; positionally restricted	[ʤ]	36 (.17)	16 (.22) (1.33)	17 (.23) (1.42)	3 (.04) (0.25)
not rare; positionally restricted	[w]	200 (.94)	106 (1.4) (1.59)	93 (1.3) (1.4)	1 (.01) (.02)
	[t]	247 (1.16)	53 (.74) (.64)	63 (.88) (.77)	131 (1.8) (1.59)
over-represented; positionally restricted	[r]	485 (2.27)	52 (.7) (.32)	191 (2.69) (1.18)	242 (3.4) (1.5)

6.2 Methodology

We have discussed the rationale for dividing the experiment into two parts, Part A containing stimuli with segments that are not strongly under-represented (as judgements for these might mask the more nuanced differences between over- and moderately under-represented conditions), and Part B containing a broader range of segmental frequencies. For consistency, it would be desirable to repeat all of the conditions of Part A in Part B but the space of possible triliteral nonce words with [r] is extremely narrow so stimuli with [r] are not included in Part B.

[1]Consonants assumed to be missing from weak roots are replaced by a place marker to maintain edge information.

Table 6.2: Experimental items for part A

Experiment II: part A			
segment	locations	repetitions	total
w	2	3	6 items
t	3	3	9 items
r	3	3	9 items
controls			27 items
Total			54 items

Table 6.3: Experimental items for part B

Experiment II: part B			
segment	locations	repetitions	total
s’	3	3	9 items
ʤ	3	3	9 items
w	3	3	9 items
t	3	3	9 items
controls			36 items
Total			72 items

The division of experimental items for parts A and B is summarized in tables 6.2 (Part A) and 6.3 (Part B). Note that Part A includes only conditions for [r] and [t], and [w] in C_1 and C_2 which range between strong over-representation and mild-under-representation (but not [w] in C_3 which is strongly under-represented) and that Part B includes a range of frequencies such as [t] in C_3, which is mildly over-represented, [w] in C_1 which has an $\text{OES}_{location}$ value close to 1 (so occurs about as often as would be expected) and strongly under-represented segments such as [w] in C_3 and [s’] in all locations.

As for experiment I, each sub-part of experiment II is preceded by a training sequence. For part A, this contains 15 items: 3 locations (C_1, C_2, C_3) * 2 segments ([r],[t]), 2 locations (C_1, C_2) for [w]) + 6 controls = 14. At the end of part A, a screen appears explaining that the second part of the experiment requires extra training. For part B, the training sequence contains 18 items: 3 locations (C_1, C_2, C_3) * 4 segments ([s'], [ʤ],[t],[w])+ 6 controls = 18. In all other aspects, Part A and Part B are analogous to experiment I.

6.2.1 Nonce word selection

For consistency, we used the same methodologies for nonce word selection (controls and stimuli) that was used in experiment I. Even the modified procedures had been sometimes difficult for experiment I, 1) because the TP, PP and ND values for some co-occurrence conditions tend to be heavily skewed and 2) because the space of nonce words using only the most frequent and evenly distributed segments is very restricted. For experiment II, the space of nonce words available with [r], [t] and [w] (in C_1 and C_2) is limited. In the case of [r] and [t] nonces it was not possible to follow our usual procedure and we were forced to use almost all the available nonces, without reference to the metrics and, for [r] nonces, there were only enough available for part A. For the rare segment conditions of part B, the space of possible nonce words is much larger and selection posed no problems. The full list of experimental items is listed in Appendix B.

Controls For control selection, we used the same methodology that we developed for experiment I. We created the set of all possible bland[2] triliteral nonce words using the 12 most frequent and evenly distributed segments (rather than the 14 used in experiment I) and selected controls from the subset of nonces having TP (transitional probability), PP (positional unigram probability) and ND (neighborhood density) values in the inter-quartile range of the values for the set of similar bland triliteral *real* words. In practice, the space of nonce triliteral roots with the 12 most frequent and evenly distributed consonants is very restricted so 26 of the controls used in experiment II are also used in experiment I.

[2]Where 'bland' denotes forms that have no known phonotactic irregularities such as OCP-Place violations or identical consonants, etc.

Stimuli We created all possible triliteral nonce forms containing 2 members of the set of 12 most frequent and evenly distributed segments and 1 segment from the set of experimental condition segments. We then selected stimuli such that their PP, TP and ND values fell within the inter-quartile range of *nonce* words for that segment and location. For example, the stimuli for [s'] in C_1 were selected from the set of nonces for which the ND, TP and PP values fall within the interquartile range for the set of nonces with [s'] in C_1 (and no other known phonotactic violations).

6.2.2 Presentation and speakers

The online presentation and recruitment of speakers was analogous to experiment I. There were 32 participants and 20 were usable according to our previously defined criteria (subjects were eliminated for unary or binary usage of the rating scale). In experiment I, only one participant was currently living in Ethiopia, but for experiment II, a linguistics graduate student in Addis Ababa kindly forwarded the link to the experiment to colleagues and 8 of them participated in the study. However, the language background of subjects living in Ethiopia is similar to that of North American participants; their first language is Amharic, but they are fluent in English and sometimes in other regional languages.

6.3 Model predictions

In this section, we evaluate the predictions of our reference model (see discussion in section 5.3)

6.3.1 Part A

In the general predictions for Part A, shown in figure 6.1 (All), we can see that the predictions for [r] are slightly lower than for controls (indicating a predicted preference for nonce words with [r] compared to controls). The median weights assigned [t] and [w] nonces are slightly higher than those assigned to

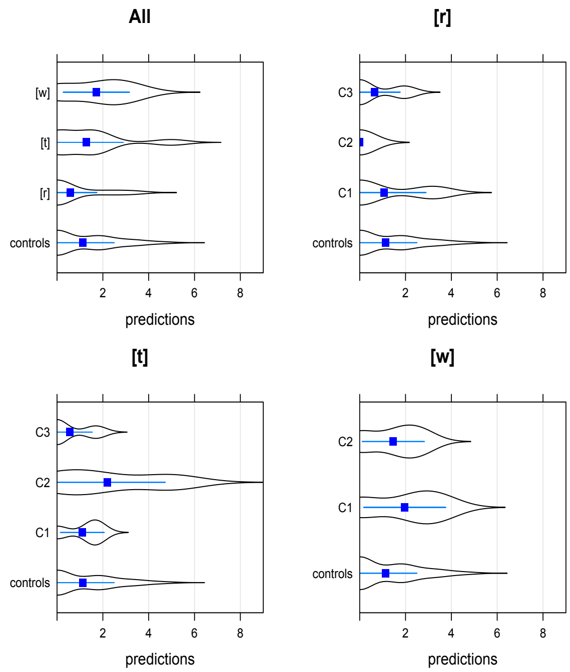


Figure 6.1: Model predictions for Part A

controls.

For [r] nonces according to location, our metrics indicate that [r] is slightly under-represented in C_1 but over-represented in C_2 and C_3, with the highest occurrence in C_3. The predicted ratings are somewhat similar to our expectations. The model predicts [r] in C_2 to be slightly better than controls but both C_1 and C_3 occurrences are predicted to be similar to controls.

The statistical analysis of the database shows that [t] is mildly under-represented in C_1 and C_2 and mildly over-represented in C_3. The model effectively predicts C_3 occurrences to be better than controls and C_2 to be worse but C_1 occurrences are close to controls.

The dictionary study shows that [w] is mildly over-represented in C_1 and C_2 but the model predicts [w] in C_1 and C_2 to be slightly worse than controls.

6.3.2 Part B

The predictions for all the conditions of part B, are shown in figure 6.2. The reference model predicts a general dispreference for [s’],[w] and [ʤ] (though least so for [w]) over controls and a mild dispreference for [t] nonces.

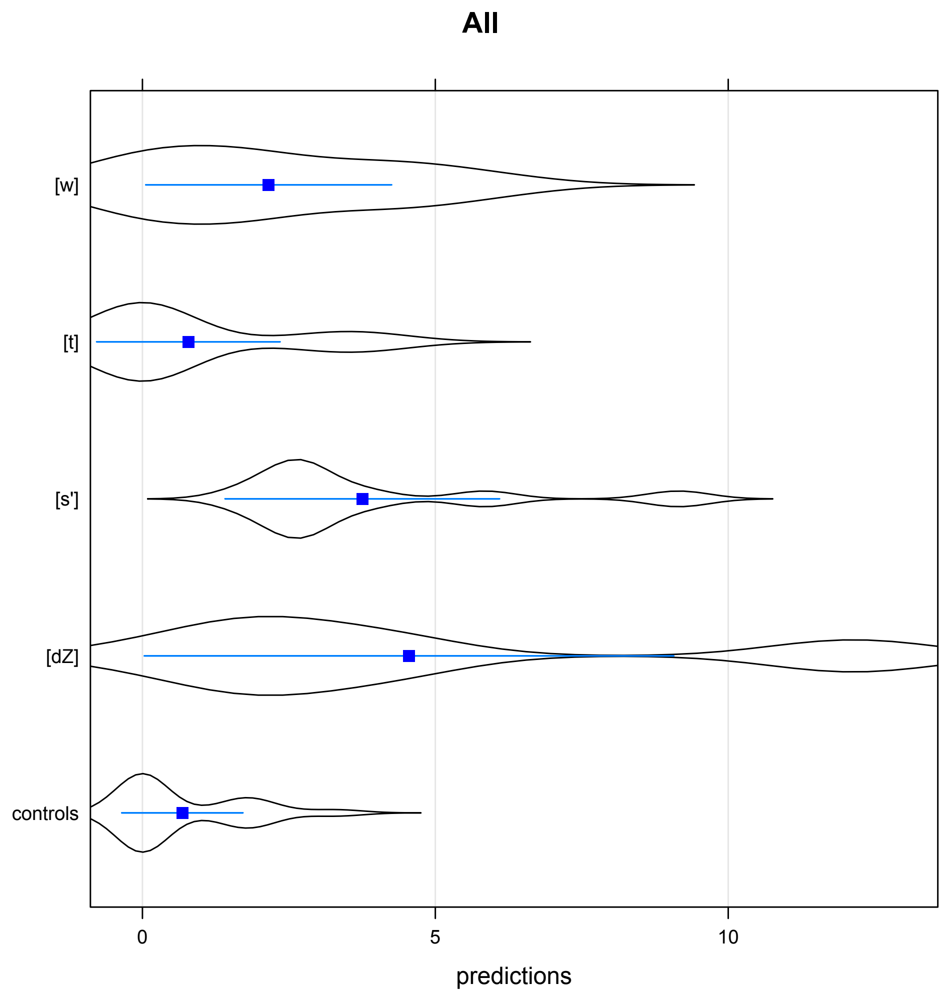


Figure 6.2: Model predictions for Part B

Regarding the predictions broken down according to the location of occurrence (figure 6.3), the predictions for [t] nonces are similar to those for the stimuli of Part A and correspond to the predictions of the dictionary analysis, with C_3 preferred to control) and C_2 and C_1 dispreferred to controls.

[w] in C_3 nonces are predicted to be strongly dispreferred (where it is very rare, according to the dictionary analysis) and on a par or mildly dispreferred to controls in the other positions.

Both [ʤ] and [s'] (which are both rare in all positions), are predicted to be dispreferred in all positions.

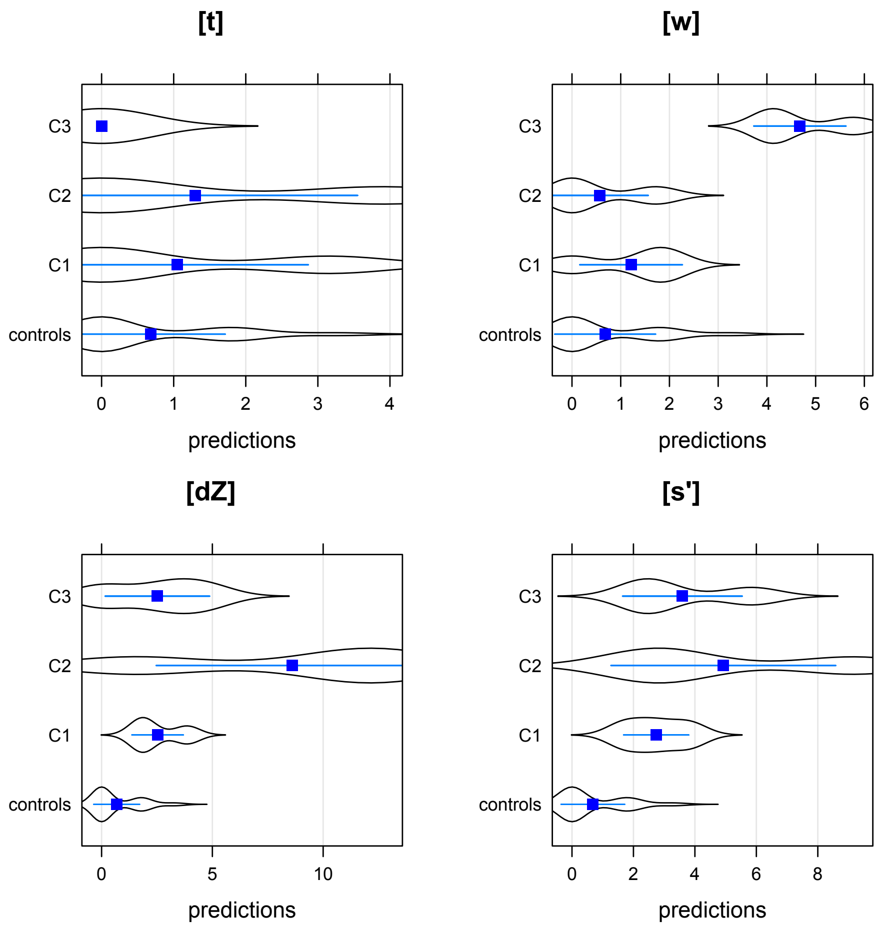


Figure 6.3: Model predictions for Part B (by segment)

6.4 Results and discussion

All results are analysed as for Experiment I; to reiterate the procedures, we use maximal LMEMs with by-subject random slope for the fixed effect, and by-item random intercepts for each nonce word using the lmer function of the statistical software R (package lme4). To estimate the p-value of a fixed effect, we compute the likelihood ratio between two models identical with regard to random effects but only one of which contains the fixed effect of interest. For a p-value $< .05$, factors with t-values $> |2|$ are assumed to be significant contributors to the effect.

6.4.1 Part A

Table 6.4: Segmental frequencies ([r])

segment	count (OES_{word})	positional count ($\mathrm{OES}_{location}$) ($\mathrm{OES}_{relative}$)		
		C1	C2	C3
[r]	485 (2.27)	52 (.7) (.32)	191 (2.69) (1.18)	242 (3.4) (1.5)

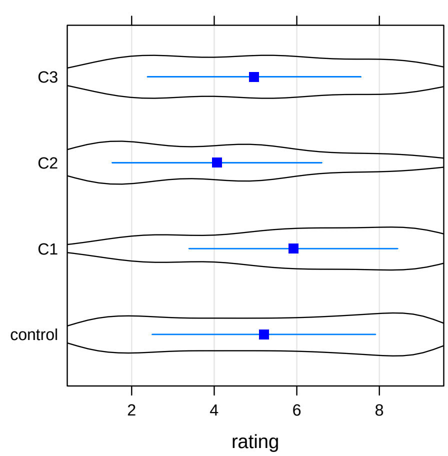


Figure 6.4: Over-represented segment [r]

[r] is over-represented both globally and in C_2 and C_3 positions but mildly under-represented in C_1 (table 6.4). Figure 6.4 (a) shows the distribution of speaker ratings for nonce forms with [r] compared to controls. There is no significant difference between controls and the set of all nonces with [r] ($p = .45$, t-value=.75), but the model with a factor for location (figure 6.4(b)) is significantly better than the model without ($p < .03$). However, only [r] in position C_1 and C_2 are significant contributors (t-values: $C_1 : 1.68, C_2 : -2.57, C_3 : -.56$).

To summarise the results for [r], in the case of [r] in C_2 there is a significant effect in the expected direction -a preference compared to controls, and for C_1, a predicted and significant dis-preference.

Table 6.5: Segmental frequencies ([t])

segment	count (OES$_{word}$)	positional count (OES$_{location}$) (OES$_{relative}$)		
		C1	C2	C3
[t]	247 (1.16)	53 (.74) (.64)	63 (.88) (.77)	131 (1.8) (1.59)

Table 6.5 is a summary of the dictionary analysis for [t] and figure 6.5 shows the results for the experimental results according to location. A model with factors for [t] stimuli as a group and controls is not significantly better than one without ($p = .6$). The model with controls and stimuli with [t] by location is not significantly better than one without ($p = .2$). However, t-values show that the dispreference for [t] in position C_2 is very close to significance ($C_1 : .1$, $C_2 : 1.92$, $C_3 : -.9$). This is contrary to the frequency prediction according to our database as the frequency (according to our metrics) of [t] is lower in C_1 than in C_2, though not by much.

To summarise the results for [t], the trend in figure 6.5 is in line with our expectations based on the dictionary analysis, but the results are close to significance only for the [t] in C_2 (trend towards dispreference compared to controls).

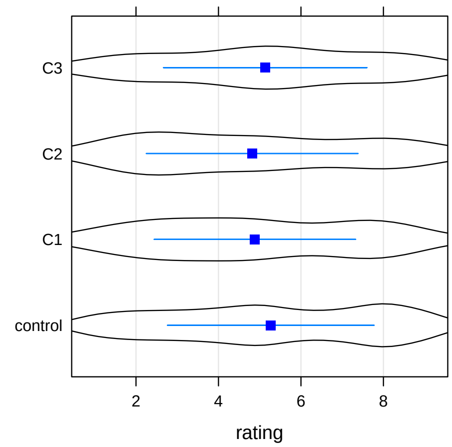


Figure 6.5: Normally-represented segment [t]

Table 6.6: Segmental frequencies ([w])

<table>
<tr><td rowspan="2">segment</td><td rowspan="2">count (OES$_{word}$)</td><td colspan="2">positional count
(OES$_{location}$)
(OES$_{relative}$)</td></tr>
<tr><td>C1</td><td>C2</td></tr>
<tr><td>[w]</td><td>200 (.94)</td><td>106
(1.4)
(1.59)</td><td>93
(1.3)
(1.4)</td></tr>
</table>

Table 6.6 shows that [w] is mildly over-represented in C_1 and C_2 but speaker ratings with [w] in C_1 and C_2 locations are not significantly different to controls (p = .4) (figure 6.4.1) and the t-values are not close to significance (C_1 : 1.2, C_2 : .12).

The results for [w] are very close to our predictions based on the dictionary analysis. These patterns are mildly over-represented but are not rated as significantly different to controls.

The experimental results for part A are generally in line with frequency based predictions, providing some evidence that nonce words with over-represented

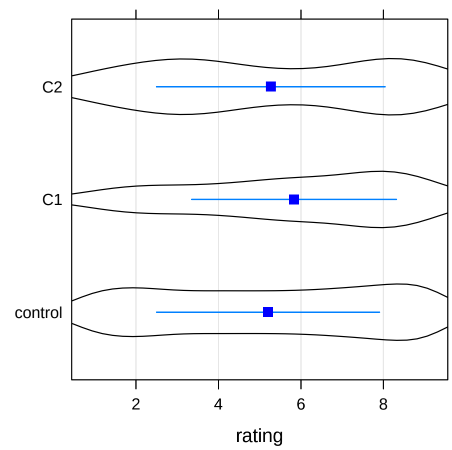


Figure 6.6: [w] in C_1 and C_2

segments may be rated better than controls. However, the ratings do not precisely line up with predictions for specific word locations. This may be because speakers are influenced by non-verb root lexical effects that are not accounted for by our database. It is also possible that judgements for a segment in a given location are not influenced only by the frequency of the segment in that location, but also by the frequency of that segment in other locations.

The role of the C_2 location is thought provoking. [r] is rated significantly better than controls only in that position, although it is actually more frequent in position C_3, and [t] is rated significantly worse than controls only in position C_2 although it is less frequent in position C_1. These results seem to indicate that the C_2 location is more salient, even though word onsets are generally considered the most salient part of the word (Cole and Jakimik, 1980; Grosjean, 1980). In the 3ms perfective verb forms used in all our experiment, ($C_1əC_2C_2əC_3ə$) the C_2 consonant is geminated. This is not a part of our model, since it is part of the perfective template rather than the root. However, it could serve to promote the saliency of segments in that position.

Part A: simulations The correlation between the average speaker judgements for the items of Part A and the reference model predictions is very weak

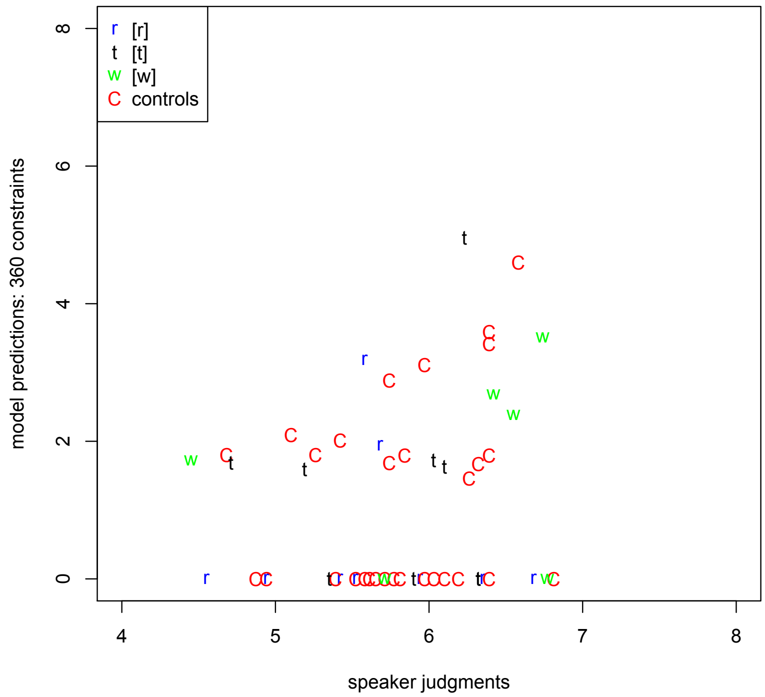


Figure 6.7: Predictions and ratings for Part A

(r = .19). Figure 6.7 shows the relationship between the reference model (360 constraints) predictions and speaker ratings.

The relationship between the ratings and model predictions according to condition and location in figure 6.8 show that the model predictions are weak for both the stimuli conditions and the controls.

6.4.2 Part B

The [t] condition was repeated in part B (with different nonce forms). As in Part A, the model with factors for controls and [t] according to location was not better than the one without ($p = .59$). The t-values ($C_1 : -.8$, $C_2 : -1.1$, $C_3 : -.3$) show that the trend towards significance for location C_2 (t-value: 1.92) in part A is no longer present.

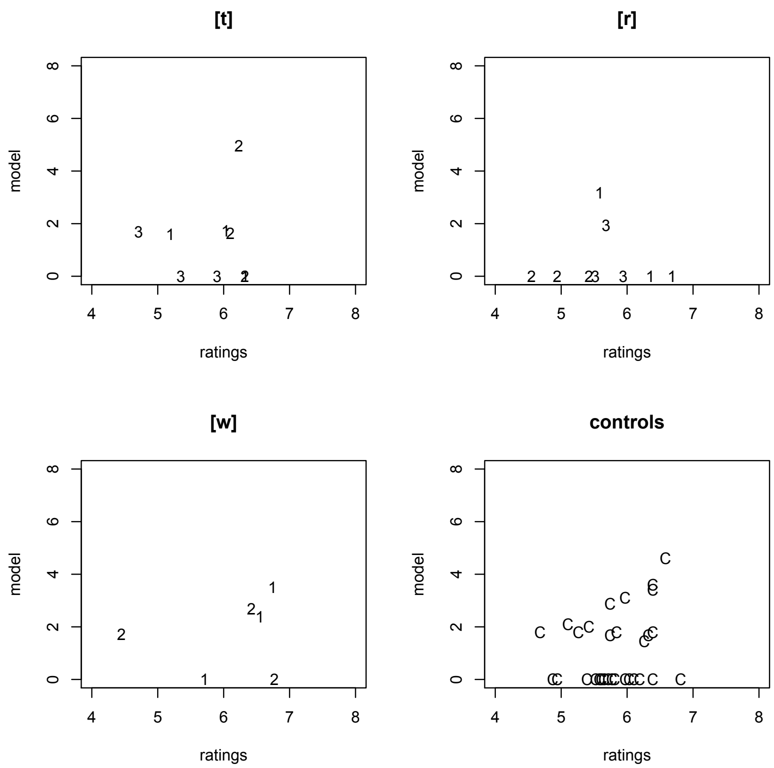


Figure 6.8: Predictions and ratings for Part A (detail)

Part B includes all three locations for the segment [w]. The dictionary analysis (table 6.7) shows that [w] is very under-represented in C_3. However, for the speaker ratings (figure 6.10), a model with a factor for location is not significantly better than the one without ($p = .36$). The t-values ($C_1 : -.36$, $C_2 : .84$, $C_3 : 1.58$) indicate only mild trend for [w] in C_3 to be dispreferred compared to controls, even though [w], at least in our database, is extremely rare in that position. It is worth nothing that the only triliteral that has [w] in C_3 is an ABA form. There are, however, quite a number of quadriliterals with [w] in C_3. It may be the case that people are processing the [w] in C_3 as the third consonant from the beginning of the word, regardless or root length, rather than as the final consonant.

For [ʤ], the dictionary analysis (table 6.8) shows that this segment in under-represented in all locations. For the speaker ratings (figure 6.11), there is a significant difference between controls and nonce words with [ʤ] ($p < .001$). Location is

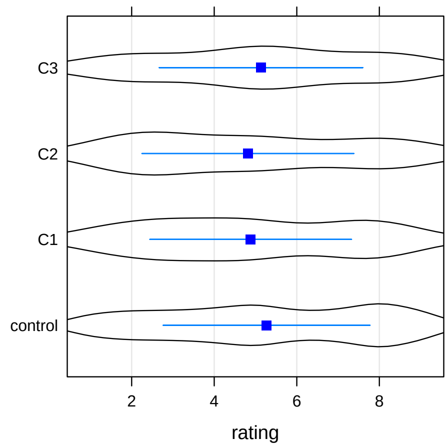


Figure 6.9: Nonce words with [t] (Part B)

Table 6.7: Segmental frequencies ([w])

segment	count (OES_{word})	positional count ($OES_{location}$) ($OES_{relative}$)		
		C1	C2	C3
[w]	200 (.94)	106 (1.4) (1.59)	93 (1.3) (1.4)	1 (.01) (.02)

significant ($p < .0011$) for C_2 and C_3 (t-values: C_2: 4.7, C_3: 3.7) and approaching significance for C_1 (t-value: 1.9). For [ʤ], the experimental results are in line with our predictions, but we note that the effect (in this case a dispreference) is more salient in C_2 (relative to other locations) than we would have predicted based on the dictionary analysis.

There is a significant difference between controls and [s'] (all locations) ($p < .05$) but the model comparison with a factor for location although significant ($p < .0001$), shows that only the C_2 and C_3 locations are significantly differnt to controls (t-values: C_1: .26, C_2: 3.4, C_3: 3.3). This result is unexpected because

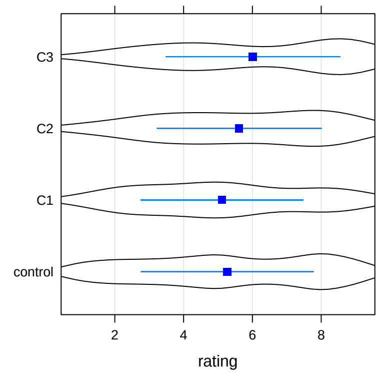


Figure 6.10: Nonce words with [w]

Table 6.8: Segmental frequencies ([ʤ])

segment	count (OES$_{word}$)	positional count (OES$_{location}$) (OES$_{relative}$)		
		C1	C2	C3
[ʤ]	36(.17)	16 (.22) (1.33)	17 (.33) (1.42)	3 (.04) (.025)

table 6.9 indicates that [s'] in C_1 is only slightly more frequent than in any other locations. Furthermore, although our dictionary analysis considers only triliterals, there are also very few quadri- or quinquiliterals with [s'] in initial position. [s'] is rare in Amharic through a diachronic process of fortition to [t'], but is preserved as [s'] more frequently in word initial position in words associated with religion and culture (Leslau, 1995) and if these are token frequent, this might account for the speaker preference for this location over others. [s'] is also the only condition where there is a significant dispreference for a rare segment in C_3 compared to controls.

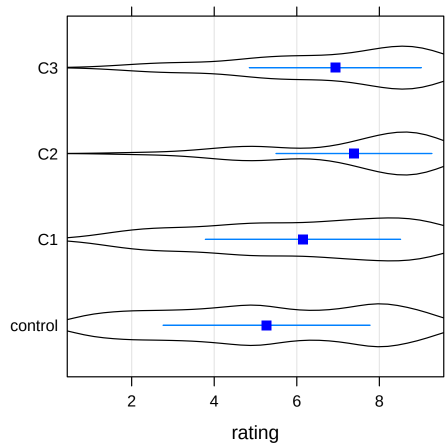


Figure 6.11: Nonce words with [ʤ]

Table 6.9: Segmental frequencies ([s'])

segment	count (OES_{word})	positional count ($OES_{location}$) ($OES_{relative}$)		
		C1	C2	C3
[s]	33(.15)	15 (.2) (1.27)	10 (.14) (1.1)	8 (.11) (.72)

Part B: simulations Figure 6.13 shows the predictions of the reference model and speaker ratings for part B. The correlation between the reference model predictions and speaker ratings is moderately high (r = .58).

Figure 6.14 shows the correlation between the speaker judgements and the reference model predictions for each Part B condition as a function of location. Note that the model tends to assign 0 or low weights to items that are not strongly under-represented ([t] in C_3, [w] in C_1 and C_2 locations). Overall, the model performs best on nonce words with strongly under-represented segments.

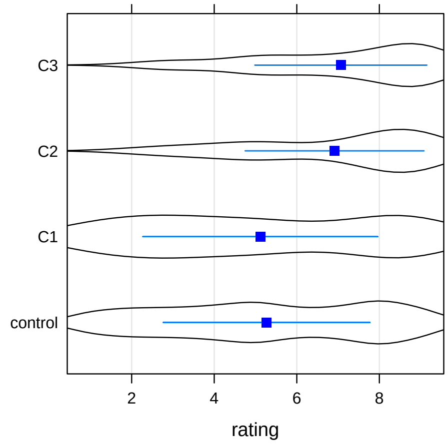


Figure 6.12: Nonce words with [s']

6.5 General discussion

Although the ratings for [s'] in C_1 and [w] in C_3 are unexpected, the results of Part B show that low-frequency segments are generally rated as less acceptable than controls. The metric that best predicts speaker ratings is OES$_{location}$ and it is most reliable for the C_2 location. OES$_{word}$ is only a good predictor of speaker ratings for [s'] and and [ʤ] which are strongly under-represented in all locations. We can eliminate OES$_{relative}$ as a useful metric because it is not predictive of speaker ratings. For example, there is no case where the predictions of OES$_{location}$ and OES$_{relative}$ conflict and OES$_{relative}$ makes the correct prediction. For example, for [s'] in C_3, OES$_{relative}$ = 1.1 and OES$_{location}$ =.11 but speakers disprefer nonce forms with [s'] in C_3, indicating that OES$_{location}$ is a better metric. Furthermore, OES$_{relative}$ > 1 does not predict a speaker preference. For example, OES$_{relative}$(ʤ) is highest for C_2 (1.42), yet nonce forms with [ʤ] in that location are preferred neither to controls nor to nonce forms with [ʤ] in other locations.

With regards to a threshold for speaker sensitivity to segmental irregularities, we must distinguish the case of C_2 from other locations. We can say that for [t] (part A), which is moderately under-represented, OES$_{location}$ < .88 in C_2 is sufficient to find a significant effect on speaker ratings compared to controls but that in

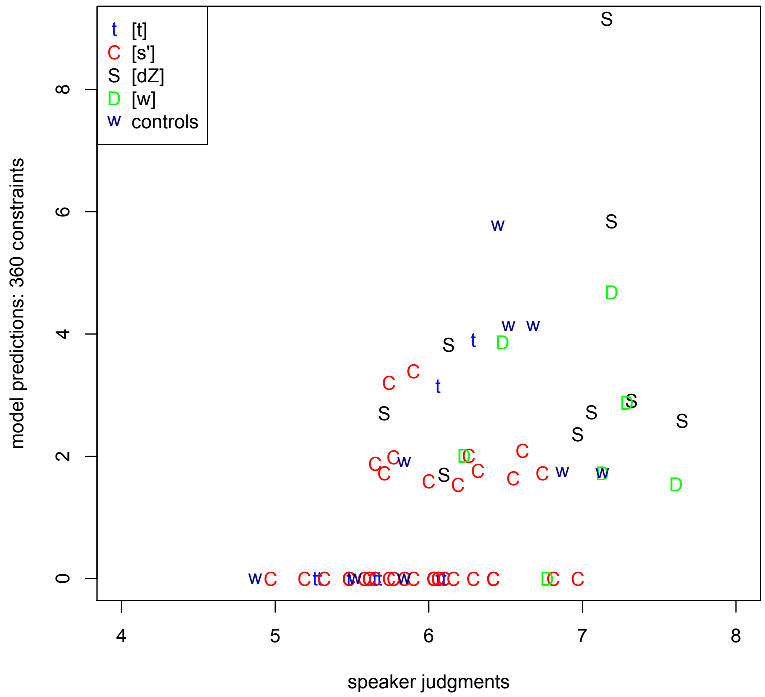


Figure 6.13: Predictions and ratings for Part B

other locations, the threshold is undetermined, but lower than .74 (OES$_{location}$(t) in C_1 = .74). For over-represented segment [r], OES$_{location}$ = 2.69 in C_2 is sufficient to find an effect, but in C_3, even OES$_{location}$=3.4 is not sufficient.

For strongly under-represented segment [ʤ], OES$_{location}$ = .33 in C_2 is sufficient to find a statistical difference with controls, but in other locations, that threshold is lower. For example, OES$_{location}$(ʤ) = .22 in C_1 but the difference between [ʤ] in C_1 and controls does not quite reach significance.

Finally, the results for nonce words with [t] modestly support the hypothesis that analogical processing plays a role in speaker judgements for nonce words from dense neighbourhoods because an effect that was close to significance in Part A was absent in Part B, indicating that stimuli from dense neighbourhoods and stimuli from sparse neighbourhoods should not be included in the same task.

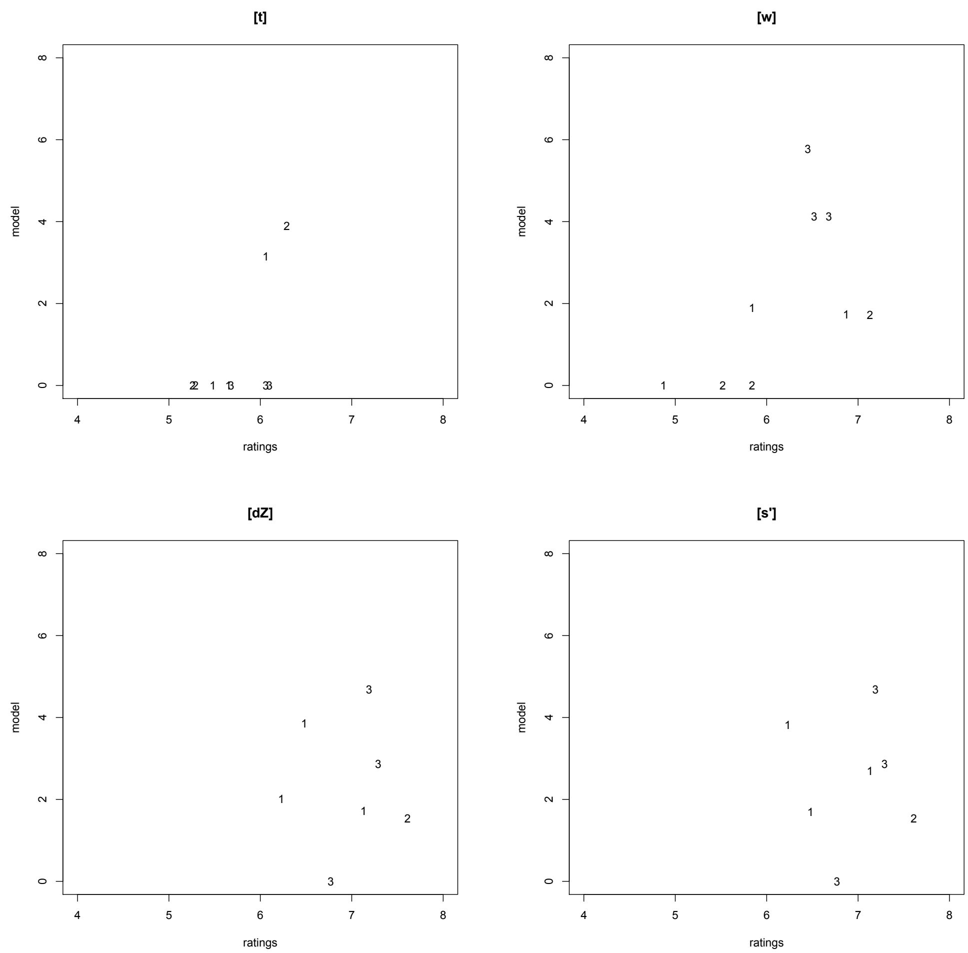


Figure 6.14: Predictions and ratings for Part B (according to location)

6.6 Conclusion

One goal of experiment II is a proof of concept; to show that speakers may, in optimum circumstances, assign ratings to nonce worlds with over-represented and moderately under-represented segments that are significantly different to controls. The results of part A has shown some success in this area. Our second goal was to determine whether the Maxent learner, a model that assigns only penalty weights, can be used to model the full range of phonotactic probability. The model consistently performs quite poorly in predicting speaker judgements for nonce words with patterns or segments that are not strongly under-represented, such as controls in experiment I, and the experimental conditions of Part A. At the same time, the experimental results themselves are quite noisy and it may be the that the experimental methodology is insufficiently sensitive to capture such fine distinctions of acceptability.

Finally, there is some evidence that the C_2 position may play a greater role in judgements for verb roots than has previously been discussed. This finding will need to be confirmed in a task designed to test that specific condition, but if it is the case, there is currently no model that can make predictions accounting for that type of saliency. For example, we could include consonant length as part of the representational system of the Maxent learner, but the model would learn that long consonants are frequent in C_2 position, not that irregularities occurring in that location are more important than elsewhere.

In the final chapter, we will present the general discussion of our simulations and experiments, and indicate the many avenues for further research.

7 Conclusion

The main goals of this dissertation are 1) to investigate the frequency-acceptability hypothesis for an under-studied language, 2) to evaluate the representational system necessary to model a language with long distance constraints over homorganic consonants and 3) to determine whether a probabilistic model such as the Maxent learner that assigns only positive weights can be used to model the full range of phonotactic probability.

With regards to the first goal, experimental results confirm that Amharic speakers are sensitive to under-represented patterns in nonce verb-roots, both in terms of co-occurrence restrictions (experiment I) and rare segments (experiment II, Part B). The results of experiment II, Part A, also provide modest evidence that speakers are sensitive to the presence of over-represented segments in the data. It is clearly more difficult to show gradience of speaker judgements of non-under-represented patterns as a function of phonotactic probablity than for under-represented conditions.

Our simulations (chapter 4) show that the representation system of the Maxent learner needs to be modified to account for the fact that identical consonants in frequent patterns do not add to speaker knowledge of harmony. In a rating task, Berent et al. (2012) collected Hebrew speaker judgements for nonce words with identical consonants and showed that speakers can extend their knowledge of identical consonant pattern frequency to consonants that do not occur in their language. They propose enriching the model with two new representations, 1) a natural class that applies to all segments and 2) an indexed representation allowing successive occurrences of a natural class matric to be represented as void of distinctive features. Although we do not have access to this modified version of the Maxent learner, given our results, we suspect that it would perform well on the Amharic data.

Regarding the performance of the Maxent learner in predicting speaker judgements, we model speaker judgements for whole words based on the consonants of a subset of the lexicon, and without frequency information (both in terms of the token frequency of verb roots and of the number of words that are derived from a given root). Given these limitations, the model performs surprisingly well in predicting speaker judgements for under-represented conditions.

The model performs poorly in predicting speaker judgements for non-under-represented conditions. The correlation between stimuli and model predictions in experiment II, Part A, is very low. There are several (possibly interacting) explanations for this failure. First, it is possible that model is correctly predicting speaker judgements based on the frequency of patterns in the lexicon, but that 1) the role of analogical processing masks probabilistic processing and 2) the speaker rating task is insufficiently sensitive. It might also be the case that the model is too noisy to correctly predict speaker judgements for non-under-represented items. Although it is theoretically possible to model the full range of phonotactic probability with a model like the Maxent learner that assigns only penalty weights, it requires the acquisition of an exhaustive grammar. Albright (2010) attempted to model English onsets with (some approximation) of an exhaustive grammar and found that model predictiveness deteriorated as a function of grammar size. This problem might be solved by modifying the constraint selection algorithm such that useless constraints are eliminated when a later, better generalization is found. However, given that our experimental evidence indicates that phonotactic probability plays a role in speaker judgements for all conditions, it would be desirable that the model acquire both positive and negative weights.

There are many avenues for future work. We cannot rule out the role of the non-verb-root lexicon word token and frequency effects in determining speaker judgements. We are currently in the process of creating a database of non-verb-roots and we are hopeful that a dictionary analysis of the data will be informative in our analysis of speaker judgements for verbs. Regarding frequency information, we suspect that the advent of Amharic corpora and online text will make this information accessible in the medium term.

The results of our experimental work provide an unusually broad perspective on the phonotactics of an under-studied language and this is because we did not select the experimental conditions based solely on phonological insights and previous work for related languages. Our choices were guided by the predictions of the Maxent learner and, although it might be said that we collected data for idiosyncratic phenomenon that are not of general interest to phonologists, the results provide information beyond the use that we have made here. For example, such results might enrich the debate on the difference in importance between phonologically natural and unnatural constraints (Hayes et al., 2009).

There is still much work to be done in the area of probabilistic phonotactics, but this thesis constitutes a contribution to this research area. However, one of the the main challenges of the field is to find appropriate and diverse experimental data to test models. Our hope is that our use of the Maxent learner to investigate the phonotactics of a language, and the success of our online rating task using a non-latin script will provide researchers with tools that will motivate them to investigate other under-studied languages.

Appendix A: Experimental items I

stimuli	root	condition	location	POA	ND	PP	TP
ለረመ	l r m	ocp	left	son	19	-8.15	-50.3
ረለቀ	r l k’	ocp	left	son	22	-8.76	-50.64
ነለሰ	n l s	ocp	left	son	22	-8.4	-50.28
ተደለ	t d l	ocp	left	stop	21	-9.39	-12.52
ተደረ	t d r	ocp	left	stop	27	-9.21	-12.29
ደጠዘ	d t’ z	ocp	left	stop	7	-9.44	-12.91
ከገለ	k g l	ocp	left	dor	27	-8.23	-48.49
ገከሰ	g k s	ocp	left	dor	19	-8.29	-48.66
ገቀነ	g k’ n	ocp	left	dor	23	-8.31	-48.6
ሰዘበ	s z b	ocp	left	fric	13	-9.31	-49.06
ዘሰመ	z s m	ocp	left	fric	11	-9.4	-49.22
ዘሰተ	z s t	ocp	left	fric	13	-9.32	-49.66
መፈደ	m f d	ocp	left	lab	15	-8.54	-12.85
ፈመለ	f m l	ocp	left	lab	18	-8.42	-49.04
ፈበቀ	f b k’	ocp	left	lab	17	-8.58	-12.96
በለረ	b l r	ocp	right	son	33	-7.43	-49.17
ሰነለ	s n l	ocp	right	son	24	-7.58	-49.25
ከነረ	k n r	ocp	right	son	23	-7.44	-10.65
ለመፈ	l m f	ocp	right	lab	14	-8.94	-12.97
ሰፈበ	s f b	ocp	right	lab	13	-8.56	-13.35

stimuli	root	condition	location	POA	ND	PP	TP
ጠመፈ	t’ m f	ocp	right	lab	18	-8.7	-12.91
በሰዘ	b s z	ocp	right	fric	12	-9.61	-49.29
ነዘሰ	n z s	ocp	right	fric	17	-10.02	-49.52
ጠዘሰ	t’ z s	ocp	right	fric	11	-9.78	-50.64
መተደ	m t d	ocp	right	stop	14	-9.02	-12.3
ለደጠ	l d t’	ocp	right	stop	14	-9.25	-12.62
ሰተደ	s t d	ocp	right	stop	10	-9.27	-12.76
ፈከገ	f k g	ocp	right	dor	13	-9.14	-50.09
ለገቀ	l g k’	ocp	right	dor	18	-8.83	-50.12
ነገከ	n g k	ocp	right	dor	15	-9.2	-49.67
ከለገ	k l g	ocp	non	dor	21	-7.97	-8.3
ገመቀ	g m k’	ocp	non	dor	23	-7.74	-8.37
ገፈከ	g f k	ocp	non	dor	16	-8.41	-8.9
ለጠነ	l t’ n	ocp	non	son	23	-9.41	-9.32
ረገነ	r g n	ocp	non	son	23	-9.53	-9.24
ረቀለ	r k’ l	ocp	non	son	17	-9.55	-9.09
ተበደ	t b d	ocp	non	stop	15	-9.09	-9.64
ተለደ	t l d	ocp	non	stop	18	-8.55	-9.99
ደሰጠ	d s t’	ocp	non	stop	19	-8.71	-9.99
ሰመዘ	s m z	ocp	non	fric	13	-9	-9.34
ሰከዘ	s k z	ocp	non	fric	13	-9.36	-11.02
ዘጠሰ	z t’ s	ocp	non	fric	9	-9.3	-11.05
መዘበ	m z b	ocp	non	lab	17	-9.07	-8.77
ፈገበ	f g b	ocp	non	lab	21	-8.6	-9.31
በከፈ	b k f	ocp	non	lab	16	-8.88	-9.12
በበነ	b b n	redup	left	lab	22	-8.49	-9.93
ደደፈ	d d f	redup	left	stop	18	-9.07	-8.39
ፈፈጠ	f f t’	redup	left	lab	19	-8.75	-8.8
ገገዘ	g g z	redup	left	dor	30	-8.47	-8.98

stimuli	root	condition	location	POA	ND	PP	TP
ከከፈ	k k f	redup	left	dor	25	-8.88	-8.67
ቀቀነ	k’ k’ n	redup	left	dor	33	-8.37	-7.85
ለለቀ	l l k’	redup	left	son	27	-8.22	-8.56
መመሰ	m m s	redup	left	lab	22	-8.27	-8.89
ነነገ	n n g	redup	left	son	10	-8.39	-8.96
ረረደ	r r d	redup	left	son	16	-8.65	-10.53
ሰሰመ	s s m	redup	left	fric	17	-9.1	-8.83
ተተፈ	t t f	redup	left	stop	15	-9.89	-9.62
ጠጠለ	t’ t’ l	redup	left	stop	18	-8.83	-8.25
ዘዘገ	z z g	redup	left	fric	11	-9.86	-8.44
በጠበ	b t’ b	redup	non	lab	24	-8.77	-8.6
ደመደ	d m d	redup	non	stop	23	-8.45	-8.48
ፈቀፈ	f k’ f	redup	non	lab	13	-9.17	-9.96
ገሰገ	g s g	redup	non	dor	18	-8.49	-9.03
ከነከ	k n k	redup	non	dor	10	-8.17	-8.38
ቀፈቀ	k’ f k’	redup	non	stop	22	-8.11	-9.17
ለበለ	l b l	redup	non	son	26	-8.58	-8.63
መተመ	m t m	redup	non	lab	14	-9.06	-8.97
ነዘነ	n z n	redup	non	son	20	-9.98	-8.97
ረበረ	r b r	redup	non	son	28	-8.94	-8.85
ሰደሰ	s d s	redup	non	fric	13	-9.14	-10.94
ተመተ	t m t	redup	non	stop	24	-9.07	-9.84
ጠደጠ	t’ d t’	redup	non	stop	8	-9.01	-13.15
ዘጠዘ	z t’ z	redup	non	fric	7	-9.78	-11.95
ረደደ	r d d	redup	right	stop	38	-10.11	-10.15
ነቀቀ	n k’ k’	redup	right	dor	43	-9.18	-8.9
ደፈፈ	d f f	redup	right	lab	35	-8.75	-8.59
ነገገ	n g g	redup	right	dor	36	-9	-9.03
ዘከከ	z k k	redup	right	dor	20	-9.37	-10.99

stimuli	root	condition	location	POA	ND	PP	TP
ተሰሰ	t s s	redup	right	fric	29	-9.72	-10.61
ረተተ	r t t	redup	right	stop	33	-10.22	-10.27
ከዘዘ	k z z	redup	right	fric	15	-10.08	-10.81
ፈሰደ	f s d	frics	fs	left	11	-9.34	-11.14
ፈሰቀ	f s k’	frics	fs	left	16	-9.18	-9.78
ፈሰለ	f s l	frics	fs	left	24	-9	-9.11
ሰፈደ	s f d	frics	sf	left	14	-8.78	-10.09
ሰፈገ	s f g	frics	sf	left	15	-8.8	-10.13
ሰፈጠ	s f t’	frics	sf	left	29	-8.47	-9.5
ፈገሰ	f g s	frics	fs	non	27	-8.85	-9.93
ፈነሰ	f n s	frics	fs	non	25	-8.24	-9.43
ፈተሰ	f t s	frics	fs	non	17	-9.58	-10.45
ሰደፈ	s d f	frics	sf	non	24	-9.11	-10.29
ሰገፈ	s g f	frics	sf	non	23	-8.54	-9.32
ሰቀፈ	s k’ f	frics	sf	non	20	-8.9	-9.59
ደፈሰ	d f s	frics	fs	right	28	-8.78	-9.15
ገሰፈ	g s f	frics	sf	right	26	-8.47	-9.52
ከሰፈ	k s f	frics	sf	right	24	-9.09	-9.35
ለሰፈ	l s f	frics	sf	right	12	-9.51	-10.45
ረፈሰ	r f s	frics	fs	right	16	-9.82	-10.29
ተፈሰ	t f s	frics	fs	right	10	-9.44	-10.16
ፈዘደ	f z d	frics	fz	left	5	-9.81	-14.32
ፈዘነ	f z n	frics	fz	left	17	-9.81	-10.6
ፈዘጠ	f z t’	frics	fz	left	9	-9.5	-12.37
ፈገዘ	f g z	frics	fz	non	18	-9.33	-10.12
ፈከዘ	f k z	frics	fz	non	11	-9.64	-11.86
ፈለዘ	f l z	frics	fz	non	23	-8.72	-9.76
ደፈዘ	d f z	frics	fz	right	12	-9.26	-10.72
ቀፈዘ	k’ f z	frics	fz	right	13	-8.78	-10.89

stimuli	root	condition	location	POA	ND	PP	TP
ረፈዘ	r f z	frics	fz	right	8	-10.3	-11.86
ዘፈከ	z f k	frics	zf	left	6	-9.3	-9.73
ዘፈለ	z f l	frics	zf	left	24	-8.75	-8.87
ዘፈጠ	z f t’	frics	zf	left	31	-8.78	-8.97
ዘገፈ	z g f	frics	zf	non	21	-8.85	-8.81
ዘከፈ	z k f	frics	zf	non	14	-9.15	-10.67
ዘጠፈ	z t’ f	frics	zf	non	16	-9.26	-10.65
ደዘፈ	d z f	frics	zf	right	7	-9.5	-11.37
ረዘፈ	r z f	frics	zf	right	10	-10.53	-10.85
ተዘፈ	t z f	frics	zf	right	8	-10.16	-12.22
መሰተ	m s t	control			34	-8.77	-8.99
መዘለ	m z l	control			29	-8.95	-8.6
መተለ	m t l	control			29	-8.68	-8.61
መደሰ	m d s	control			25	-8.9	-8.98
መደቀ	m d k’	control			30	-8.7	-8.81
መጠሰ	m t’ s	control			25	-8.75	-8.82
መቀለ	m k’ l	control			35	-8.31	-8.53
ፈለከ	f l k	control			27	-8.42	-8.77
ፈረከ	f r k	control			31	-8.15	-8.43
ፈደረ	f d r	control			35	-8.87	-8.9
ፈከለ	f k l	control			29	-8.78	-9.06
ፈገነ	f g n	control			26	-8.81	-9.21
ፈቀለ	f k’ l	control			29	-8.83	-8.78
በሰተ	b s t	control			26	-9.05	-9.31
በደቀ	b d k’	control			35	-8.98	-9.13
በጠነ	b t’ n	control			33	-8.98	-8.55
በከሰ	b k s	control			30	-8.91	-9
በገለ	b g l	control			30	-8.23	-9.01
በቀነ	b k’ n	control			32	-8.92	-8.63

stimuli	root	condition	location	POA	ND	PP	TP
በቀጠ	b k’ t’	control			27	-8.61	-8.8
ለመተ	l m t	control			27	-8.9	-9.39
ለበተ	l b t	control			29	-8.88	-9.28
ነበሰ	n b s	control			35	-8.95	-8.77
ነበደ	n b d	control			30	-8.91	-8.74
ነገፈ	n g f	control			29	-8.99	-8.98
ነገጠ	n g t’	control			41	-8.68	-9.05
ሰመተ	s m t	control			29	-8.45	-9.19
ሰመደ	s m d	control			25	-8.49	-8.75
ሰመቀ	s m k’	control			26	-8.33	-9.48
ሰበደ	s b d	control			26	-8.46	-8.77
ሰበጠ	s b t’	control			37	-8.15	-8.54
ሰለገ	s l g	control			36	-7.94	-8.41
ሰተለ	s t l	control			29	-8.92	-9.07
ሰጠለ	s t’ l	control			25	-8.62	-9.45
ሰገመ	s g m	control			32	-8.58	-8.76
ሰገለ	s g l	control			38	-8.2	-8.41
ሰገነ	s g n	control			30	-8.54	-8.9
ሰቀነ	s k’ n	control			25	-8.89	-8.6
ዘመጠ	z m t’	control			31	-8.48	-8.57
ዘበቀ	z b k’	control			25	-8.61	-9.09
ዘለመ	z l m	control			30	-8.27	-8.51
ዘለበ	z l b	control			32	-8.01	-8.67
ዘለተ	z l t	control			33	-8.19	-9.11
ዘተረ	z t r	control			34	-9.05	-9.27
ዘገተ	z g t	control			30	-8.8	-8.97
ዘገደ	z g d	control			27	-8.85	-8.35
ዘቀለ	z k’ l	control			26	-8.86	-8.96
ዘቀረ	z k’ r	control			32	-8.68	-8.67

stimuli	root	condition	location	POA	ND	PP	TP
ተመረ	t m r	control			35	-8.59	-9.16
ተፈረ	t f r	control			27	-8.89	-9.05
ተበለ	t b l	control			25	-8.75	-9.11
ተበረ	t b r	control			34	-8.57	-8.85
ተለፈ	t l f	control			30	-8.55	-9.19
ተለበ	t l b	control			35	-8.33	-9.16
ተረመ	t r m	control			26	-8.32	-8.57
ተረሰ	t r s	control			36	-8.31	-9.23
ተረዘ	t r z	control			26	-8.8	-9.47
ደመሰ	d m s	control			35	-8.48	-8.51
ደበዘ	d b z	control			32	-8.94	-8.76
ደበከ	d b k	control			26	-8.64	-8.83
ደሰለ	d s l	control			26	-8.68	-8.87
ደሰረ	d s r	control			32	-8.51	-9.18
ደከሰ	d k s	control			36	-8.84	-9.16
ጠመለ	t’ m l	control			26	-8.36	-8.48
ጠመሰ	t’ m s	control			32	-8.73	-8.93
ጠለገ	t’ l g	control			25	-8.15	-9.05
ጠነቀ	t’ n k’	control			33	-7.98	-8.7
ጠቀለ	t’ k’ l	control			30	-8.76	-8.65
ከመለ	k m l	control			28	-8.17	-8.32
ከመሰ	k m s	control			26	-8.55	-8.76
ከመደ	k m d	control			25	-8.52	-8.72
ከረዘ	k r z	control			26	-8.2	-8.64
ከሰበ	k s b	control			27	-8.87	-8.52
ገፈደ	g f d	control			34	-8.2	-8.73
ገሰነ	g s n	control			32	-8.47	-8.36
ገሰተ	g s t	control			29	-8.43	-9.16
ገዘበ	g z b	control			31	-8.73	-8.57

stimuli	root	condition	location	POA	ND	PP	TP
ገደሰ	g d s	control			37	-8.56	-8.44
ቀፈጠ	k’ f t’	control			36	-7.95	-8.65
ቀሰበ	k’ s b	control			29	-8.32	-8.57
ቀዘለ	k’ z l	control			37	-8.67	-8.75
ቀደፈ	k’ d f	control			31	-8.59	-9.29
ቀደነ	k’ d n	control			26	-8.58	-9.09

Appendix B: Experimental items II

item	root	condition	location	Part	ND	PP	TP
ከደበ	k d b	control	control	A	22	-8.91	-9.57
በከሰ	b k s	control	control	A	30	-8.91	-9
ደቀመ	d k' m	control	control	A	20	-8.9	-9.59
መደሰ	m d s	control	control	A	25	-8.9	-8.98
ሰቀነ	s k' n	control	control	A	25	-8.89	-8.6
ገጠዘ	g t' z	control	control	A	23	-8.89	-9.99
ጠቀበ	t' k' b	control	control	A	18	-8.89	-9.33
ፈቀጠ	f k' t'	control	control	A	23	-8.86	-9.14
ዘቀለ	z k' l	control	control	A	26	-8.86	-8.96
ፈገደ	f g d	control	control	A	24	-8.82	-9.18
ፈገነ	f g n	control	control	A	26	-8.81	-9.21
ለቀበ	l k' b	control	control	A	25	-9.13	-9.06
ነበከ	n b k	control	control	A	17	-9.13	-9.35
ፈከነ	f k n	control	control	A	22	-9.12	-9.38
ጠቀፈ	t' k' f	control	control	A	17	-9.11	-9.84
ለፈቀ	l f k'	control	control	A	23	-9.08	-9.6
ለከበ	l k b	control	control	A	20	-9.08	-9.48
ነከበ	n k b	control	control	A	23	-9.07	-9.3
ደሰነ	d s n	control	control	A	23	-9.02	-9.05
በጠገ	b t' g	control	control	A	20	-9.01	-9.75

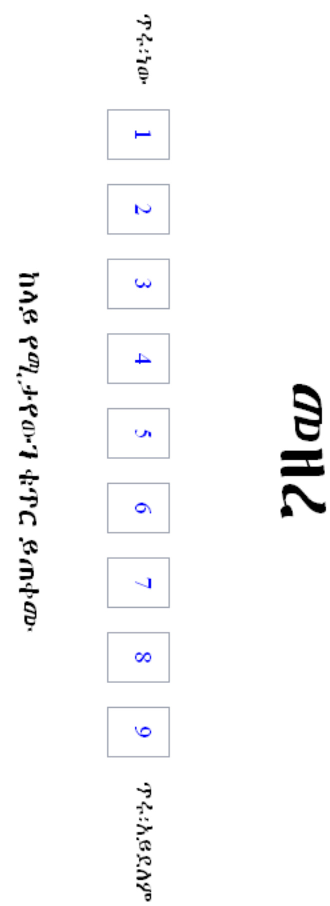
item	root	condition	location	Part	ND	PP	TP
ለገፈ	l g f	control	control	A	22	-8.99	-9.43
ነገፈ	n g f	control	control	A	29	-8.99	-8.98
ቀጠዘ	k' t' z	control	control	A	24	-8.96	-9.61
ለመገ	l m g	control	control	A	20	-8.96	-9.34
ሰጠነ	s t' n	control	control	A	22	-8.96	-9.5
መዘለ	m z l	control	control	A	29	-8.95	-8.6
ሰቀመ	s k' m	control	control	A	20	-8.93	-9.46
በሰቀ	b s k'	control	control	A	28	-8.93	-9.01
ጠሰለ	t' s l	control	control	A	17	-8.93	-9.26
ረቀደ	r k' d	r	C1	A	15	-9.89	-10.81
ረጠሰ	r t' s	r	C1	A	15	-9.99	-10.65
ረሰበ	r s b	r	C1	A	13	-9.84	-10.6
ዘረበ	z r b	r	C2	A	34	-7.74	-8.34
ዘረቀ	z r k'	r	C2	A	33	-7.8	-8.32
ደረመ	d r m	r	C2	A	34	-7.66	-8.28
ደሰረ	d s r	r	C3	A	32	-8.51	-9.18
ፈደረ	f d r	r	C3	A	35	-8.87	-8.9
ዘቀረ	z k' r	r	C3	A	32	-8.68	-8.67
ተበከ	t b k	t	C1	A	13	-9.31	-10.26
ተሰለ	t s l	t	C1	A	17	-9.34	-10.22
ተበዘ	t b z	t	C1	A	15	-9.61	-10.19
ዘተመ	z t m	t	C2	A	16	-9.61	-10.32
በተሰ	b t s	t	C2	A	16	-9.33	-10.1
ለተገ	l t g	t	C2	A	13	-9.74	-10.02
ዘከተ	z k t	t	C3	A	27	-9.11	-10.42
ለመተ	l m t	t	C3	A	27	-8.9	-9.39
በሰተ	b s t	t	C3	A	26	-9.05	-9.31
ወዘጠ	w z t'	w	C1	A	18	-9.31	-9.98
ወጠሰ	w t' s	w	C1	A	18	-9.08	-9.63

item	root	condition	location	Part	ND	PP	TP
ወሰገ	w s g	w	C1	A	15	-9.17	-8.91
ደወከ	d w k	w	C2	A	15	-9.48	-10.53
ዘወጠ	z w t'	w	C2	A	20	-9.29	-10.14
ጠወገ	t' w g	w	C2	A	18	-9.53	-10.01
ከፈደ	k f d	control	control	B	23	-8.81	-8.98
ደከፈ	d k f	control	control	B	18	-8.81	-9.28
ዘመገ	z m g	control	control	B	16	-8.81	-9.26
ደሰበ	d s b	control	control	B	19	-8.8	-9.37
ጠገመ	t' g m	control	control	B	26	-8.79	-9.79
ፈከለ	f k l	control	control	B	29	-8.78	-9.06
ለመቀ	l m k'	control	control	B	23	-8.78	-9.68
በሰጠ	b s t'	control	control	B	17	-8.78	-9.45
ጠቀለ	t' k' l	control	control	B	30	-8.76	-8.65
መጠሰ	m t' s	control	control	B	25	-8.75	-8.82
ሰጠበ	s t' b	control	control	B	23	-8.74	-9.54
መደቀ	m d k'	control	control	B	30	-8.7	-8.81
ሰመከ	s m k	control	control	B	18	-8.7	-9.41
ደሰለ	d s l	control	control	B	26	-8.68	-8.87
ደመከ	d m k	control	control	B	16	-8.67	-9.13
ጠፈለ	t' f l	control	control	B	22	-8.65	-9.01
ደበከ	d b k	control	control	B	26	-8.64	-8.83
ደቀበ	d k' b	control	control	B	23	-8.64	-9.21
ዘመቀ	z m k'	control	control	B	21	-8.63	-9.6
ሰከበ	s k b	control	control	B	20	-8.62	-9.04
ሰጠለ	s t' l	control	control	B	25	-8.62	-9.45
በቀጠ	b k' t'	control	control	B	27	-8.61	-8.8
ዘበቀ	z b k'	control	control	B	25	-8.61	-9.09
ደከበ	d k b	control	control	B	22	-8.59	-9.76
ቀደነ	k' d n	control	control	B	26	-8.58	-9.09

item	root	condition	location	Part	ND	PP	TP
በገደ	b g d	control	control	B	23	-8.57	-9.47
ከመሰ	k m s	control	control	B	26	-8.55	-8.76
ሰገነ	s g n	control	control	B	30	-8.54	-8.9
ከመደ	k m d	control	control	B	25	-8.52	-8.72
ሰመደ	s m d	control	control	B	25	-8.49	-8.75
ቀመዘ	k' m z	control	control	B	20	-8.48	-9.18
ሰበገ	s b g	control	control	B	20	-8.48	-9.83
ከለዘ	k l z	control	control	B	19	-8.47	-9.23
ደመገ	d m g	control	control	B	16	-8.47	-8.86
ሰበደ	s b d	control	control	B	26	-8.46	-8.77
ሰበነ	s b n	control	control	B	26	-8.46	-9.83
ጀነዘ	ʤ n z	ʤ	C1	B	6	-10.24	-11.05
ጀለዘ	ʤ l z	ʤ	C1	B	11	-10.23	-10.63
ጀፈገ	ʤ f g	ʤ	C1	B	8	-10.6	-10.56
ሰጀገ	s ʤ g	ʤ	C2	B	8	-10.84	-49.18
ነጀቀ	n ʤ k'	ʤ	C2	B	9	-11.11	-49.08
ሰጀበ	s ʤ b	ʤ	C2	B	10	-10.6	-48.3
ቀፈጀ	k' f ʤ	ʤ	C3	B	7	-9.93	-11.4
ሰመጀ	s m ʤ	ʤ	C3	B	8	-10.15	-11.16
ዘለጀ	z l ʤ	ʤ	C3	B	9	-9.89	-10.86
ጸበነ	s' b n	s’	C1	B	8	-10.75	-12.52
ጸመቀ	s' m k'	s’	C1	B	9	-10.62	-12.15
ጸመለ	s' m l	s’	C1	B	11	-10.44	-11.01
ደጸመ	d s' m	s’	C2	B	8	-11.19	-49.71
ነጸጠ	n s' t'	s’	C2	B	10	-11.32	-48.92
ከጸደ	k s' d	s’	C2	B	7	-11.22	-48.82
መደጸ	m d s'	s’	C3	B	8	-11.5	-48.91
መጠጸ	m t' s'	s’	C3	B	11	-11.35	-48.75
ጠለጸ	t' l s'	s’	C3	B	11	-10.77	-12.01

item	root	condition	location	Part	ND	PP	TP
ተገፈ	t g f	t	C1	B	14	-9.17	-9.89
ተመገ	t m g	t	C1	B	12	-9.13	-9.79
ተበሰ	t b s	t	C1	B	18	-9.13	-9.68
ለተበ	l t b	t	C2	B	16	-9.5	-9.98
ነተገ	n t g	t	C2	B	12	-9.73	-9.89
ሰተገ	s t g	t	C2	B	15	-9.28	-9.67
ለበተ	l b t	t	C3	B	29	-8.88	-9.28
ነፈተ	n f t	t	C3	B	27	-9.19	-9.23
ሰመተ	s m t	t	C3	B	29	-8.45	-9.19
ወጠለ	w t' l	w	C1	B	22	-8.71	-8.91
ወለዘ	w l z	w	C1	B	23	-8.53	-8.78
ወሰቀ	w s k'	w	C1	B	21	-8.99	-8.73
ደወገ	d w g	w	C2	B	18	-9.28	-9.7
ጠወነ	t' w n	w	C2	B	20	-9.51	-9.54
ዘወነ	z w n	w	C2	B	21	-9.6	-9.47
ቀደወ	k' d w	w	C3	B	7	-10.6	-10.46
ቀዘወ	k' z w	w	C3	B	8	-11.02	-10.16
ሰከወ	s k w	w	C3	B	8	-10.85	-10.01

Appendix C: Word rating screen



Appendix D: Experiment instructions

የዚህ፡ ጥናት፡bአላማ፡ የኣማርኛ፡ ቃላት፡ ከሌሎች፡ ቋንቋዎች፡ ቃላት፡ በምን፡ እንደሚለያዩ፡ ለመረዳት፡ነው። ቃላቶቹም፡ አዲስ፡ የተፈጠሩ፡ ሲሆኑ፡ የሚሰጡትም፡ አስተየየት፡ እያንዳንዱ፡ ቃላት፡ ምን፡ ያክል፡ የአማርኛ፡ ቃል፡ ሊሆን፡ እንደሚችል፡ ሀሳቦን፡ እንዲሠጡ፡ ነው። ለቋንቋ፡ ጥናት፡ የተፈጠሩት፡ ቃላት፡ ሁሉ፡ በቃለ፡ መዝገባት፡ ውስጥ፡ የግሥ፡ ክፍል፡ ናቸው። ለምሳሌ፤ ሰበረ ("he broke")

በያንዳንዱ፡ ገፅ፡ አንድ፡ ቃልና፡ ከአንድ፡ እስከ፡ ዘጠኝ፡ የሚያሣይ፡ ሚዛን፡ ይኖራል።

ቃሉ፡ ጥሩ፡ የአማርኛ፡ ቃል፡ መሆን፡ የሚችል፡ ከሆነ፡ አንድ(1)፡ ቁጥር፡ ይመርጣሉ። ቃሉ፡ ፈፅሞ፡ የአማርኛ፡ ቃል፡ሊሆን፡ የማይችል፡ ከሆነ፡ ዘጠኘ (9)፡ ቁጥር፡ ይመርጣሉ። ቃሉ፡ ጥሩ፡ የአማርኛ፡ ቃል፡ ባይመስልም፡ የአማርኛ፡ ቃል፡ ሊሆን፡ የሚችል፡ ከሆነ፡ ከአንድ፡ ቁጥር፡ የቀረበ፡ ቁጥር፡ ይስጡት። ቃሉ፡ የአማርኛ፡ ቃል፡ባይመስልም ፡ምናልባት፡ የአማርኛ፡ ቃል፡ የሚመስል፡ ካልሆነ፡ ወደ፡ ዘጠኝ (9)፡ ቀጥር፡ የቀረበ፡ቁጥር፡ ይስጡት።

የመጀመርያ፡ ሀያ (twenty) ፡ ቃላት፡ የመለማመጃ፡ ቃላቶች፡ ናቸው። በያንዳንዱ፡ ቃላት፡ ብዙ፡ጊዜ፡ አያሳልፉ። እያንዳንዱ፡ ቃል፡ ምን፡ ያህል፡ ጥሩ፡ የአማርኛ፡ ቃል፡ እንደሚሆን፡ በቁጥር፡ከገመቱ፡ በኃላ፡ ይቀጥሉ፡ የሚል፡ ምልክት፡ ይጫኑ። የዚህ፡ ጥናት፡ አላማ፡ የኣማርኛ፡ ቃላት፡ከሌሎች፡ ቋንቋዎች፡ ቃላት፡ በምን፡ እንደሚለያዩ፡ ለመረዳት፡ ነው። ቃላቶቹም፡ አዲስ፡ የተፈጠሩ፡ ሲሆኑ፡ የሚሰጡትም፡ አስተየየት፡ እያንዳንዱ፡ ቃላት፡ ምን፡ ያክል፡ የአማርኛ፡ ቃል፡ሊሆን ፡እንደሚችል፡ ሀሳቦን፡ እንዲሠጡ፡ ነው። ለቋንቋ ፡ጥናት፡ የተፈጠሩት፡ ቃላት፡ ሁሉ፡ በቃለ፡ መዝገባት፡ ውስጥ፡ የግሥ፡ ክፍል፡ ናቸው። ለምሳሌ፤ ሰበረ ("he broke")

በያንዳንዱ፡ ገፅ፡ አንድ፡ ቃልና፡ ከአንድ፡ እስከ፡ ዘጠኝ፡ የሚያሣይ፡ ሚዛን፡ ይኖራል።

ቃሉ፡ ጥሩ፡ የአማርኛ፡ ቃል፡መሆን፡ የሚችል፡ ከሆነ፡ አንድ (1)፡ ቁጥር፡ ይመርጣሉ። ቃሉ፡ፈፅሞ፡ የአማርኛ፡ ቃል፡ ሊሆን፡ የማይችል፡ ከሆነ፡ ዘጠኘ (9)፡ ቁጥር፡ ይመርጣሉ። ቃሉ፡ ጥሩ፡የአማርኛ፡ ቃል፡ ባይመስልም፡ የአማርኛ፡ ቃል፡ ሊሆን፡ የሚችል፡ ከሆነ፡ ከአንድ፡ ቁጥር፡

የቀረበ፥ቁጥር፡ ይስጡት። ቃሉ፡ የአማርኛ፡ ቃል፡ ባይመስልም፡ ምናልባት፡ የአማርኛ፡ ቃል፡ የሚመስል፡ ካልሆነ፡ ወደ፡ ዘጠኝ (9)፡ ቀጥር፡ የቀረበ፡ ቁጥር፡ ይስጡት።

የመጀመርያ ፡ሀያ (twenty) ፡ቃላት፡ የመለማመጃ፡ ቃላቶች፡ ናቸው። በያንዳንዱ፡ ቃላት፡ ብዙ፥ጊዜ፡ አያሳልፉ። እያንዳንዱ፡ ቃል፡ ምን፡ ያህል፡ ጥሩ፡ የአማርኛ፡ ቃል፡ እንደሚሆን፡ በቁጥር፡ ከገመቱ፡ በኃላ፡ ይቀጥሉ፡ የሚል፡ ምልክት፡ ይጫኑ።

Bibliography

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26:9–41.

Albright, A. (2010). Personal communication.

Albright, A. (2011a). *Modeling analogy as probabilistic grammar*. Oxford University Press. (in press).

Albright, A. (2011b). Natural classes are not enough: Biased generalization in novel onset clusters. draft.

Albright, A. and Hayes, B. (2003). Rules vs. analogy in english past tenses: a computational/experimental study. *Cognition*, 90:119–161.

Armbruster, C. H. (1908). *Initia Amharica*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Bachra, B. (2001). *The phonological structure of the verbal roots in Arabic and Hebrew*. Studies in Semitic languages and linguistics. Brill, Leiden.

Bailey, T. M. and Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44:568–591.

Banksira, D. P. (2000). *Sound Mutations: The morphophonology of Chaha*. John Benjamins, Amsterdam.

Bat-El, O. (1994). Stem modification and cluster transfer in modern Hebrew. *Natural Language and Linguistic Theory*, 12:571–596. 10.1007/BF00992928.

Bat-El, O. (2003). *Semitic verb structure with a universal perspective*, pages 29–59. John Benjamins, Amsterdam.

Bat-El, O. (2006). Consonant copying and consonant identity: The segmental and prosodic structure of Hebrew reduplication. *Linguistic Inquiry*, 37:179–210.

Becker, M., Ketrez, N., and Nevins, A. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in turkish laryngeal alternations. *Language*, 87(1):84–125.

Beckman, J. N. (1997). Positional faithfulness, positional neutralisation and shona vowel harmony. *Published by: Cambridge University Press*, Vol. 14(No. 1):1–46.

Bender, L. and Fulass, H. (1978). *Amharic Verb Morphology.* The African Studies Center, Michigan State University, East Lansing, Michigan.

Benmamoun, E. (1999). Arabic morphology: The central role of the imperfective. *Lingua*, 108:175–201.

Berent, I. (2008). Are phonological representations of printed and spoken language isomorphic? Evidence from the restrictions on unattested onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 34.

Berent, I. and Shimron, J. (1997). The representation of Hebrew words: Evidence from the Obligatory Contour Principle. *Cognition*, 64:39–72.

Berent, I. and Shimron, J. (2003). Co-occurrence restrictions on identical consonants in the Hebrew lexicon: Are they due to similarity? *Journal of Linguistics*, 39:31–55.

Berent, I., Steriade, D., Lennertz, T., and Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104:591–630.

Berent, I., Wilson, C., Marcus, G. F., and Bemis, D. K. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson. *Linguistic Inquiry*, 43(1):97–119.

Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Berkley, D. (1994). The OCP and gradient data. *Studies in the Linguistic Sciences*, 24:59–72.

Buckley, E. (1990). Edge-in association and ocp 'violations' in Tigrinya. In *Proceedings of the Ninth West Coast Conference on Formal Linguistics.*

Buckley, E. (1997). Tigrinya root consonants and the OCP. In *Penn Working Papers in Linguistics, 4*, pages 19–51.

Cantineau, J. (1946). Esquisse d'une phonologie de l'arabe classique. *Bulletin de la société linguistique*, 43:93–140.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, 2:pp. 225–252.

Clements, G. N. and Keyser, S. J. (1983). *CV Phonology: A Generative Theory of Syllable Structure*. The MIT Press.

Coetzee, A. W. and Pater, J. (2006). Lexically ranked OCP-place constraint in muna. Unpublished manuscript, Utrecht University, the Netherlands and University of Nantes, France.

Cohen, M. (1970). *Traité de langue amharique (Abyssinie)*. Institut d'Ethnologie, Musée de l'Homme, Université de Paris, Paris.

Cole, R. A. and Jakimik, J. S. (1980). *Perception and production of fluent speech*, chapter A model of speech perception, pages 133–163. Erlbaum, Hillsdale, NJ.

Coleman, J. S. (1996). The psychological reality of language-specific constraints. University of Manchester. Paper presented at the Fourth Phonology Meeting.

Coleman, J. S. and Pierrehumbert, J. B. (1997). Stochastic phonological grammars and acceptability. pages 49–56, Somerset, New Jersey. Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology, Association for Computational Linguistics.

Daland, R., Hayes, B., White, J., Gareliek, M., Davis, A., and Norrman, I. (2011). Explaining sonority projection effects. *Phonology*, (22):197–234.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997). Inducing features of random fields. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19, pages 380–393.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification*. Pattern Classification and Scene Analysis: Pattern Classification. Wiley.

Eisner, J. (2001). Expectation semirings: Flexible EM for finite-state transducers. In van Noord, G., editor, *Proceedings of the ESSLLI Workshop on Finite-State Methods in Natural Language Processing (FSMNLP)*. Extended abstract (5 pages).

Elmedlaoui, M. (1995). Géométrie des restrictions de cooccurrence de traits en sémitique et de berbère: synchronie et diachronie. *Revue canadienne de linguistique*, 40(1):39–79.

Even-Shoshan, A. (1993). *Ha'milon ha'xadash [The new dictionary]*. Kiryat Sefer, Jerusalem.

Frisch, S. (1996). *Similarity and Frequency in Phonology*. PhD thesis, Northwestern University.

Frisch, S. A., Large, N. R., and Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language*, (42):481–496.

Frisch, S. A., Pierrehumbert, J. B., and Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22:179–228.

Frisch, S. A. and Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, 77(1):91–106.

Gafos, A. I. (1999). *The articulatory basis of locality in phonology*. PhD thesis, Johns Hopkins University.

Gallagher, G. (2008). Total identity in co-occurrence restrictions. In *BLS 34*. Berkeley Linguistics Society, University of California, Berkeley.

Gallagher, G. (2010). Perceptual distinctness and laryngeal disharmony. *Phonology*, 27(3):435–480.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the workshop on variation within optimality theory*.

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, 6:162–81.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Attention, Perception, & Psychophysics*, 28(4).

Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies*, 4(2):1–24.

Hay, J., Pierrehumbert, J., and Beckman, M. (2003). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. O. and Temple, R., editors, *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge University Press, Cambridge.

Hay, J., Pierrehumbert, J., and Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39.3:379–440.

Hayes, B., Zuraw, K., Siptár, P., and Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85:822–863.

Hayward, K. and Hayward, R. J. (1992). Amharic. *Journal of the International Phonetic Association*, 22:48–52.

Hudson, G. (1979). Ethiopian B-type verbs. In Hess, R., editor, *Proceedings of the Fifth International Conference of Ethiopian Studies (Chicago)*, pages 101–110, Chicago. University of Illinois Press.

Kane, T. (1990). *Amharic-English dictionary.* Harrassowitz.

King, L. and Rose, S. (2003). Strong and weak constraints on Amharic roots: Evidence from a judgment task. Unpublished manuscript.

Kirby, J. P. and Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgements in Cantonese. In *ICPhS XVI*, pages 1389–1392.

Klingenheben, A. (1966). *Deutsch-Amharischer Sprachführer.* Wiesbaden, Harrassowitz.

Koskined, K. (1964). Kompatibilität in den dreikonsonantigen hebräischen wurzeln. *Zeitschrift der Deutschen Morgenländischen Gesellschaft*, 114:16–58.

Kucera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English.* Brown University Press, Providence, RI.

Kurylowicz, J. (1972). *Studies in Semitic Grammar and Metrics.* Warsaw: Polska Akademia Nauk.

Leben, W. (1973). *Suprasegmental Phonology.* PhD thesis, MIT.

Leslau, W. (1941). *Documents Tigrigna.* C. Klincksieck, Paris.

Leslau, W. (1957). *Une hypothèse sur la forme primitive du type B en amharique.* Linguistic Circle of New York.

Leslau, W. (1958). *The verb in Harari : (South Ethiopic).* University of California Press, Berkeley.

Leslau, W. (1968). *Amharic Textbook.* Harrassowitz.

Leslau, W. (1995). *Reference Grammar of Amharic.* Harrassowitz, Wiesbaden.

Levy, R. (2012). Probabilistic models of word order and syntactic discontinuity. Draft.

Lin, Y.-H. (2011). *Affricates*, chapter 16. Blackwell.

Lowenstamm, J. (1986). À propos d'une hypothèse sur la forme primitive du type B en amharique. *Revue québécoise de linguistique*, 16(1):157–179.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics*, 39(3):155–158.

Mackenzie, S. (2011). Contrast and the evaluation of similarity: evidence from consonant harmony. *Lingua*.

Maddieson, I. (1984). *Patterns of sounds. With a chapter contributed by Sandra Ferrari Disner*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge University Press.

Manning, C. D. and Klein, D. (2003). Optimization, maxent models, and conditional estimation without magic. In *HLT-NAACL*.

McCarthy, J. J. (1979). *Formal problems in Semitic phonology and morphology*. PhD thesis, MIT.

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12(3):373–418.

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17:207–263.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica*, 43:84–108.

McCarthy, J. J. (1994). The phonetics and phonology of Semitic pharyngeals. *Papers in Laboratory Phonology III: Phonological Structure and Phonetic Form*, pages 191–283.

Mester, R. A. (1986). *Studies in Tier Structure*. PhD thesis, U. Mass, Amherst.

Michael S. Vitevitch, Paul A Luce, D. B. P. and Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1-2):306–11.

Myers, J. and Tsay, J. (2005). The processing of phonological acceptability judgments. In *Proceedings of Symposium on 90-92 NSC Projects*, pages 26–45, Taipei, Taiwan.

Myers, S. (1997). OCP effects in optimality theory. *Natural Language and Linguistic Theory*, 15(4):847–892.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.

Ohala, J. J. and Ohala, M. (1986). *Experimental Phonology*, chapter Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. Academic Press.

Padgett, J. (1995). Partial class behavior and nasal place assimilation. In *Coyote Working Papers in Linguistics*. University of Arizona, Tucson.

Pierrehumbert, J. B. (1992). Dissimilarity in the Arabic verbal roots. In *Proceedings of the 23rd Meeting of the Northeastern Linguistic Society*, pages 367–381, U. Mass. Amherst. Graduate Student Association.

Prunet, J.-F. and Chamora, B. (2001). *New data and new methods in Afroasiatic linguistics: Robert Hetzron in memoriam*, chapter "The canonical shapes of Gurage verbs", pages 143–152. Harrassowitz.

Prunet, J.-F. and Petros, D. (1996). *Studies in Afroasiatic grammar*, chapter L'interaction entre schemes et racines en chaha. Holland Academic Graphics, The Hague.

Ratcliffe, R. (1997). Prosodic templates in a word-based morphological analysis of arabic. In Mushira. Eid, R. R., editor, *Perspectives on Arabic Linguistics X. Papers from the Tenth Annual Symposium on Arabic Linguistics*, pages 147–171, Amsterdam and Philadelphia. John Benjamins.

Raz, S. (1983). *Tigre grammar and texts*. Undena Publications, Malibu, California.

Riggle, J. (1999). *Relational markedness in Bantu vowel harmony*, volume 6. UC Santa Cruz.

Rose, S. (2000). Rethinking geminates, long-distance geminates and the OCP. *Linguistic Inquiry*, (31):85–122.

Rose, S. (2003a). *Language Processing and Acquisition in Languages of Semitic, Root-based Morphology*, chapter The formation of Ethio-Semitic internal reduplication. Joseph Shimron, Amsterdam.

Rose, S. (2003b). Triple take: Tigre and the case of internal reduplication. In *San Diego Linguistic Papers*, volume 1, pages 109–128.

Rose, S. and King, L. (2007). Speech error elicitation and co-occurrence restrictions in two Ethiopian languages. *Language and Speech*, 50 (4):451–504.

Rose, S. and Walker, R. (2004). A typology of consonant agreement as correspondence. *Language*, 80(3):475–531.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, 10(3):187–228.

Scholes, R. J. (1966). *Phonotactic grammaticality*. Mouton, The Hague.

Shademan, S. (2007). *Grammar and Analogy in Phonotactic Well-formedness Judgments*. PhD thesis, University of California, Los Angeles.

Steriade, D. (1995). Underspecification and markedness. In Goldsmith, J., editor, *The handbook of phonological theory*, pages 114–174. Blackwell, Oxford.

Sumner, C. and François, A. (1957). *'Etude expérimentale de l'amharique moderne: d'après la prononciation d'Abraha François*. Addis Ababa.

Treiman, R., Kessler, B., Kneewasser, S., Tincoff, R., and Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In Broe, M. and Pierrehumbert, J., editors, *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, pages 269–283. Cambridge University Press, Cambridge, UK.

Unseth, P. E. (2002). *Bi-consonantal reduplication in Amharic and Ethio-Semitic*. PhD thesis, University of Texas at Arlington.

Ussishkin, A. (2000). *The Emergence of Fixed Prosody*. PhD thesis, University of California, Santa Cruz.

Vitevitch, M. S. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, 9(4):325–.

Vitevitch, M. S. and Luce, P. A. (2004). A web-based interface to calculate phonotactic probabilities for words and nonwords in english. *Behavior Research Methods, Instruments, and Computers*, 36:481–487.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., and Kemmerer, D. (1997). Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech*, 40:47–62.

Wedekind, K. (1990). Glottalization constraints and Ethiopian counter-evidence. *Folia Linguistica*, 24:94–127.

Wehr, H. and Cowan, J. (1971). *A Dictionary of Modern Written Arabic*. O. Harrassowitz.

Yip, M. (1988). The Obligatory Contour Principle and phonological rules: a loss of identity. *Linguistic Inquiry*, 19:65–100.