March 2016

The Representation of Probabilistic Phonological Patterns: Neurological, Behavioral, and Computational Evidence from the English Stress System

Claire Moore-Cantwell
University of Massachusetts Amherst

THE REPRESENTATION OF PROBABILISTIC PHONOLOGICAL PATTERNS: NEUROLOGICAL, BEHAVIORAL, AND COMPUTATIONAL EVIDENCE FROM THE ENGLISH STRESS SYSTEM

A Dissertation Presented

by

CLAIRE MOORE-CANTWELL

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2016

Linguistics

THE REPRESENTATION OF PROBABILISTIC PHONOLOGICAL PATTERNS: NEUROLOGICAL, BEHAVIORAL, AND COMPUTATIONAL EVIDENCE FROM THE ENGLISH STRESS SYSTEM

A Dissertation Presented

by

CLAIRE MOORE-CANTWELL

Approved as to style and content by:

_____

Lyn Frazier, Co-chair

_____

Joe Pater, Co-chair

_____

John Kingston, Member

_____

Lisa Sanders, Member

_____

John Kingston, Department Chair
Linguistics

For Katie, Wes, and the newest Butterfield
Ask until you get an answer

ACKNOWLEDGMENTS

Looking back over the course of this dissertation research, I find that a truly staggering number of people helped me in crucial ways throughout the process. It will be impossible to thank all of them, or to thank any of them adequately, but I will do my best.

First of all, I must thank Lyn and Joe for being willing to co-chair this dissertation. It would have been impossible to do the research herein without both of them on board. I always looked forward to meetings with them, even in those inevitable moments when everything was terrible, because I always came out with a renewed love for the work I was doing - or if not that, I at least came out with a plan. One reason I came to grad school was because I wanted to be challenged in a way that would force me to develop as a person - that would help me to further develop the tools I needed to tackle the supremely difficult problem of finding truth, and knowing whether or not you've found it. Joe and Lyn provided that challenge. I believe that both this dissertation and my entire scientific career have greatly benefited from the empirical and theoretical rigor with which they conduct their own research, and which they have graciously, and at no small cost to themselves in time and energy, passed on to me.

Joe's help at every stage of my graduate career has been invaluable. He helped me become a grown-up linguist by treating me like I already was one, almost as soon as I got to UMass. He was always going to be part of this dissertation - it is directly because of Joe that I work on variation, that I work on English stress, and that I've been able to blend psycholinguistic work with phonological theory. He was frank and level-headed in giving advice, and both supportive and critical of my

work. He took my off-the-wall ideas seriously, and even turned some of them into actual good ideas. One of my favorite things about working with Joe was how much fun our meetings always were - you can tell you're having an excellent meeting when the guy in the office next door has to come tell you to stop laughing so loud (sorry Kyle). I have endless respect for Joe's boldness and enthusiasm in entertaining new ideas (nonetheless always tempered by solid logic), which helped this dissertation immensely, and which I will always try to emulate.

I am very proud to be part of the long tradition of students who have been asked over and over again 'what's your hypothesis?' Lyn's patience with my underdeveloped ideas over the years boggles the mind. With her unwaveringly high expectations for my work, she constantly pushed me to be more than I was. Absent her encyclopedic knowledge of the field, her clarity of thinking, and her relentless push toward answering the big questions, I might have been able to write a dissertation, but it would not have been this one. I would have had much less fun without Lyn. Besides constantly reminding me why I was even studying linguistics in the first place, she somehow made things fun that really shouldn't have been, like fixing terrible writing, or preparing for an interview. And she was always generous with advice, from overcoming writer's block to organizing a bibliography. Of the depth of Lyn's influence on every aspect of my academic life, I will only say this: I never used to wear so many blazers and scarves as I do now.

A brief experience with John's polymathic and infectious intellectual curiosity was enough to convince me to come to UMass for graduate school in the first place, and it has been a joy to have him on my committee. John constantly pushed me to look at things from new angles, and to stretch my understanding of the problems I was working on. Perhaps even more importantly, he was someone I could ask the difficult questions, and always get a reasoned, and usually encouraging, response. John was

always there when I needed him, and meetings and conversations with him got me through a lot.

To Lisa, I owe a debt of thanks for taking me in to her lab, even though I had no training or really any right to be there. She has a seemingly endless capacity for immediately understanding linguistic theory, as well as a depth of knowledge about the brain and the techniques we use to understand it that I can only hope to someday achieve a mere fraction of. She answered all my questions with utmost patience, and in pushing me to become a better psychologist also made me a better linguist.

Other members of the Umass linguistics faculty also provided invaluable help and support along the way. John McCarthy deserves specific thanks for teaching me how to work OT, teaching me how to write for people like John McCarthy, and for excellent feedback on the early stages of this dissertation work. Kyle Johnson not only made a valiant effort to get me to understand syntax (sorry Kyle), and also laughed at my jokes a lot. Kristine Yu and Brian Dillon provided lots of advice and encouragement, both about specifics of this dissertation and about my career in general. To others, I am grateful both for teaching and advising, and for generally making UMass a really nice department to grow up in, especially: Rajesh Bhatt, Seth Cable, Ilaria Frana, Lisa Green, Alice Harris, Angelika Kratzer, Magda Oiry, Barbara Partee, Anne Pycha, Tom Roeper, Lisa Selkirk, Peggy Speas, and Ellen Woolford.

No acknowledgements could possibly be complete without mention of Kathy Adamczyk, Tom Maxfield, Michelle McBride, and Sarah Vega-Liros. Kathy and Tom in particular listened to no small amount of frustrations, kept me company on multiple occasions while people were deliberating about me, and always told me that my knitting looked nice. All four collectively made sure all my i's were dotted, all my t's were crossed, and that it all happened on time.

The UMass Psychology department as a whole was gracious enough to let me hang around for several years. I am particularly grateful to Chuck Clifton, Andrew

I would certainly have been lost in my first year without the companionship of my cohort - Elizabeth Bogal-Allbritten, Minta Elsman, Seda Kan, Jason Overfelt, Presley Pizzo, Robert Staubs, and Andrew Weir. I was fortunate enough to get to share South College 311 with them throughout the majority of our graduate careers. It was so much fun working together, laughing,arguing, and engaging in general shenanigans. I could not have asked for a better cohort, and I already miss all of them. I'd especially like to thank Robert, Minta, and Andrew, who I lived with at 40 Grant my first year, and who instantly made being 3000 miles away from everyone I knew more

bearable. Robert deserves one additional mention, since it is because of him that I know anything about computational linguistics, a subfield which I found somewhat difficult to break into. He took lots of time out of his own work over the years to answer questions and give advice, and has in general been an awesome collaborator and friend. Jason Overfelt also deserves special thanks for letting me use his voice for basically every experiment I conducted at UMass.

The other students and affiliates at UMass, both older and younger than me, provided immensely useful and necessary discussion and support, both intellectual and social. In particular, I'd like to thank Caroline Andrews for getting me outdoors sometimes when I was deep in writing mode, Hannah Greene for listening a lot (and just generally being amazing), Amanda Rysling for some really useful discussion surrounding the data in Chapter 4, the Slogging-Ryslett (terrible to behold) for lots wine and gossip, and Brian Smith for getting me out of bed early for breakfast on countless occasions - many hours of useful work time on this dissertation are directly attributable to him. I'd also like to thank Michael Becker, Sakshi Bhatia, Michael Clauss, Tracy Conner, Noah Constant, Chris Davis, Emily Elfner, David Erschler, Lena Fainleib, Meg Grant, Ivy Hauser, Clint Hartzell, Coral Hughto, Jyoti Iyer, Karen Jesney, Stefan Keine, Mike Key, Wendell Kimper, Leland Kusmer, Nick LaCara, Josh Levy, Suzi Lima, Jon Ander Mendia, Kevin Mullin, Alexei Nazarov, Ali Neyers, Deniz Ozyildiz, Yangsook Park, Jérémy Pasquereau, Ethan Poole, Kathryn Pruitt, Aynat Rubinstein, Shayne Sloggett, Megan Somerday, Guillermo Vales-Kennedy, Alicia Wolfe, Alexander Wolfe-Walkow, and Martin Walkow.

I'd also like to thank the Schardl-Hine household: Anisa, Snowflake, Gizmo, Callie, Geordie, and Alistair. They took care of me throughout the past several years in many different ways, including giving me a place to live, helping me take breaks when I needed to, and even giving me a bike when I needed one. Anisa also helped me figure out how to teach, and provided a great deal of insight and discussion about

ABSTRACT

THE REPRESENTATION OF PROBABILISTIC PHONOLOGICAL PATTERNS: NEUROLOGICAL, BEHAVIORAL, AND COMPUTATIONAL EVIDENCE FROM THE ENGLISH STRESS SYSTEM

FEBRUARY 2016

CLAIRE MOORE-CANTWELL

B.A., UNIVERSITY OF CALIFORNIA LOS ANGELES

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lyn Frazier and Professor Joe Pater

This dissertation investigates the cognitive mechanism underlying language users' ability to generalize probabilistic phonological patterns in their lexicon to novel words. Specifically, do speakers represent probabilistic patterns using abstract grammatical constraints? If so, this system of constraints would, like categorical phonological generalizations (a) be limited in the space of possible generalizations it can represent, and (b) apply to known and novel words alike without reference to specific known words. I examine these two predictions, comparing them to the predictions of alternative models. Analogical models are specifically considered.

In chapter 3 I examine speakers' productions of novel words without near lexical neighbors. Speakers' productions of these novel words are compared to actual (relatively distant) words which could serve as an analogical base. Participants successfully extended a probabilistic trend in the lexicon to novel words, and did not use

the analogical bases to do so: the contents of an analogical base for a given nonword did not predict participants' behavior on that nonword.

In chapter 4 I discuss a case of mismatch with the lexicon - participants extend a near-categorical trend in the lexicon to novel words, but they undermatch the distribution found in the lexicon. This undermatching would not be predicted if learners could induce arbitrarily complex constraints. I argue instead that the trend is represented grammatically, and that the mismatch arises because of a bias for simpler constraints either in learning or in the structure of the grammar itself.

If probabilistic phonological generalizations are represented abstractly, how do they interact with the lexicon of stored word forms? I address this issue in chapter 2 by looking at the perception of known and novel forms. ERP data demonstrates that a productive probabilistic trend influences the early stages of the lexical access process, specifically in known words. I consider two possible mechanisms for this: (1) that the lexical entries of known exceptional forms differ from known trend-observing forms, or (2) that the process of accessing an exceptional form involves a violation of expectations imposed by the grammar, and thus requires more processing power than the process of accessing a trend-observing form.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Speakers and listeners extend regularities in the lexicon of their native language
to novel forms. For example, in English, sequences of stops and fricatives (e.g. [st]
or [sf]) at the ends of words always agree in voicing: there are words like [mæst],
but no words like *[mæsd]. In new words, for example new plurals, English speakers
obey this generalization categorically. [wʌg]$_{pl}$ is pronounced as [wʌgz], where the [g]
and [z] are both voiced, while [wʌk]$_{pl}$ is pronounced as [wʌks], with [k] and [s] both
voiceless (Berko, 1958).

What is the cognitive mechanism that underlies speakers' ability to generalize
forms in their lexicon to novel words? This ability proceeds from a system of linguistic
knowledge containing at least a lexicon and a production and perception mechanism.
An important question for phonological research is whether or not this system also
contains a grammatical component - a system of abstract rules or constraints. If so,
what is its structure, and what are its limitations? In this dissertation I specifically
investigate the question of whether probabilistic generalizations are encoded in the
phonological grammar. In this introduction, I start by defining what a grammatical
encoding of a generalization is, and by contrasting it with other ways that general-
ization might occur, in particular through analogy to existing lexical items.

A grammatical system of abstract generalizations would be an instance of what
Russell (1910) calls 'knowledge by description': knowledge of a set of propositions
about an object or concept without direct experience or acquaintance with that ob-

1

ject or concept. He gives the example of an election in which it is possible to know that (P) "the candidate who gets the most votes will be elected". One can know this proposition about the candidate who will get the most votes without knowing anything else about the candidate. In particular, we do not need to link 'the candidate' in (P) to any actual candidate in the real world of which we have specific knowledge. Even though one may also happen to have specific knowledge of an actual candidate, call her X, who will happen to be the candidate who wins the election, it is not necessary to know that X will be the candidate who gets the most votes in order to know the proposition (P). In the same way, if a language user has as part of her linguistic system the rule 'word-final obstruents agree in voicing', this rule does not refer to any particular instances of word-final obstruents or words that contain them. A person having this generalization about the phonology of his or her language technically need not know any words with word-final obstruents, or even any words at all, to know that the generalization is true of the language. In actual linguistic systems, there would be a close relationship between the content of the generalizations in the phonological grammar and the statistics of the lexicon. A speaker would know not only a generalization, but many words which observe it, and possibly also some exceptions which violate it. The point is that knowledge of the generalization, and knowledge of the words which observe or violate it, do not depend on each other.

Russell contrasts knowledge by description with 'knowledge by acquaintance', a concept closer related to a speakers' knowledge of the individual words of his or her lexicon. He characterizes knowledge by acquaintance as direct cognitive relation, or direct awareness of an object. It is the converse of presentation: If A is presented to B, then B is acquainted with A. This is the kind of knowledge with which we know or are familiar with a particular instance of a thing, such as a particular chair with which we have direct experience, or a particular person whose behavior we have directly observed. He points out that we can in fact only be acquainted with

2

cognitive objects or sense data since we do not experience the world outside ourselves directly but rather through the filter of our senses. Knowledge of a word of course proceeds from individual sensory experiences with that word, and ultimately may consist of exactly a collection of particular experiences (as in exemplar theory, e.g. Pierrehumbert, 2001). E.g. Chomsky and Halle (1968) take the view that knowledge of a word consists of an abstract sequence of phonemes or feature bundles, rather than a collection of individual experiences. Under this view, knowledge of a word would be more akin to knowledge of a person. An understanding of the behavior of a particular person is based on many separate (sensory) experiences with that person, but those experiences all relate to the same cognitive object - a knowledge by acquaintance with an individual person in the real world. Similarly, a collection of sensory experiences with a particular word all relate to one's understanding (or representation) of that word - a cognitive object.

If speakers do in fact form grammatical generalizations separate from knowledge of any words of their lexicon, it is not straightforward to show that they exist, or to discover what exactly they are. This is chiefly because these generalizations are implicit in the sense that they are not consciously accessible the way knowledge of a proposition (like P) would be. Speakers are not consciously aware that they know generalizations about the phonology of their language or what those generalizations are. The content of the generalizations can only be inferred through observations of speakers' judgments of individual instances of words, or else through observation of the production or perception process of individual instances of words. Figure 1.1 illustrates a potential underlying structure for an individual's phonological system. The individual's lexicon - a set of words known by acquaintance - and grammar - a set of abstract generalizations - are not accessible to direct observation. Rather, we suppose that an individual's production and perception mechanisms make reference to the lexicon, and are influenced by the contents of the grammar. By more or less

3

directly observing production processed we can make hypotheses about a potential underlying grammatical system.

Figure 1.1. Boxological model of the phonological system
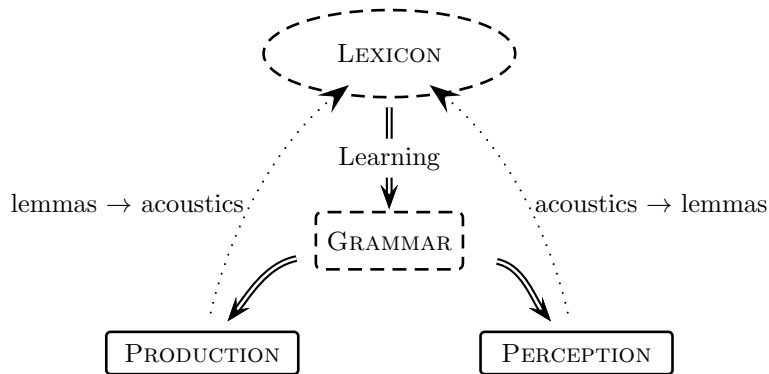
The grammar and the lexicon here are related to each other only through a learning process. Because of this process, the contents of a language's grammar will closely mirror patterns found in the lexicon. Additionally, constraints on the learning process can result in a grammar that diverges from the patterns found in the lexicon[1]. Once learning is complete the grammar does not rely on the lexicon to function. Conceptually, if the lexicon were to change drastically, this would not necessarily affect the contents of the grammar after learning is complete. In the real world, one can think of the early stages of second language acquisition as a case of drastic change in the lexicon without much change in the grammar. In learning a second language as an adult, one learns a new lexicon, whose sound patterns are different from the patterns of the learner's first language. The early stages of second language acquisition might be characterized as the use of the first language grammar and a second language lexicon.

---

[1]I am ignoring here the very real possibility that the acquisition of the lexicon is also influenced by the contents of the grammar.

Figure 1.1 also indicates that production and perception processes necessarily access the lexicon in different ways. Speech production is a process that flows from concepts to motor processes (and ultimately to acoustics) (Levelt, 1993). One step of this process is accessing the appropriate phonological content for a given lemma. On the other hand, perception flows from acoustic input to concept. The corresponding crucial step is accessing a word's lemma given some acoustics. This characterization of both processes is grossly simplified, and in particular leaves out issues of morphological composition and potential interaction between phonological processing and other levels of structure, such as syntax or semantics. The point here is that perception and production processes share some structure, but do not necessarily use that structure in the same way.

Speakers' behavior on novel words provides crucial data for determining the contents of speakers' abstract generalizations about their lexicon, since a speaker has had no previous direct experience with a novel word, and can therefore have no knowledge by acquaintance of that word. If speakers knew a set of words but had no mechanism for extending patterns in those words to novel situations, we would expect that speakers' behavior on novel words would be unpredictable, and would not relate to the contents of the lexicon in any clear way. In fact, in experiment after experiment, phonologists find close matches between the contents of a speaker's lexicon and their behavior on novel words. Speakers judge novel words which exhibit sounds or patterns not found in the words of their lexicon to be less like the words of their language, less 'pronounceable', or less likely to be a word than novel words which share sounds or patterns with words in their lexicon. When speakers are asked to pronounce a novel word, or to inflect it, they do so in ways that follow the patterns found in their lexicon, avoiding structures which do not occur. This ability to generalize patterns over known words to novel words must proceed either from a system of abstract generalizations (knowledge by description), or from some strategy for extending knowledge

by acquaintance with existing words directly to that novel word. A variety of models instantiating each strategy appear in the phonological literature, as well as models incorporating both strategies.

Perhaps the most straightforward class of models designed to extend knowledge of the lexicon to novel words without abstract generalization is analogy. The basic mechanism of an analogical model is that a speaker chooses the behavior of a novel word based on analogy to the behavior to a single actual word, or possibly to a group of actual words. A word is chosen as the 'analogical base' based on its similarity to the novel word. For example, in choosing the plural for a novel form [wʌk], a speaker could note that a similar actual word 'duck' has the plural [dʌks] with a voiceless [s], and use the voiceless [s] as the plural for [wʌk]. If a real word with a different plural were chosen as the analogical base, then that word's plural would be used. For example, if the word 'mug' were chosen, whose plural is [mʌgz] with a voiced [z], he novel word would be inflected with [z] - [wʌkz]. The accuracy of these models as predictors of human behavior on novel words is entirely dependent on the mechanism used to determine which analogical base is chosen. Novel words would behave like the actual words to which they are the most similar according to the model, and in aggregate patterns in the lexicon would also hold of novel words. Notable models of this type include Skousen (1989); Daelemans et al. (1994); Nakisa et al. (2001) - see section 3.4 for further discussion of each type of model.

Models employing abstract generalizations come in two main flavors: rules (Chomsky and Halle, 1968), and constraints (Prince and Smolensky, 1993/2004). The chief difference is that rules specify an environment, and a change ('t becomes d after a voiced obstruent'), while constraints specify a marked structure ('obstruent clusters disagreeing in voicing are prohibited'). Both contain abstractions over the lexicon, and do not reference any particular actual words. Two crucial properties of a system with abstract generalizations can immediately be understood. First, since the

6

abstractions do not reference any particular actual words, they apply inexorably to known and novel words alike. Pinker and Prince (1988) point out that at least some phonological generalizations (they discuss the regular past tense in English) are broadly applicable, in that speakers can generate a legal output for any input regardless of the details of its representation, or similarity to existing words (Prasada and Pinker, 1993). Some generalizations can exert a strong influence on the perception and production systems - listeners can perceptually 'repair' forms that violate a generalization in their native language (Dupoux et al., 1999; Berent et al., 2007; Moreton, 2002), and generalizations can carry over into a second language (Finn and Kam, 2008; Whalen and Dell, 2006), in some cases even after a learner has acquired a large number of foreign lexical items.

Second, restrictions on the form and substance of rules and constraints will lead to restrictions on what can be represented by the grammar, so not all regularities in the lexicon will actually be captured by the language's grammar. Which regularities do get represented by the grammar is constrained both by the structure of the grammar itself, and by the learning mechanism used to induce the grammar from the lexicon. Becker et al. (2011); Hayes et al. (2009) both describe cases of mismatch between participants' responses and the lexical statistics, citing the complexity of the constraints involved as well as their phonetic naturalness as underlying factors leading to the mismatch.

Argumentation surrounding the the existence or non-existence of implicit abstract grammatical generalizations has typically focused on (a) the fit of models of either type to empirical data or (b) the simplicity of one overall type of model compared to the other. Typically this latter argument is used to favor models without a grammatical component (because they have one less component than grammatical models they are 'simpler'). In the next sections I will discuss each of these metrics, arguing that both are flawed. Both types of models can produce very good fits to experimental

data, and it is not clear how to directly compare the complexity of two very different types of models, since both need a mechanism beyond just the lexicon for generalizing to novel words.

In this dissertation, I focus on probabilistic phonological generalizations rather than categorical ones. Probabilistic generalization present an interesting challenge to models of phonology: Models with abstract generalizations excel at capturing speakers' knowledge of categorical phonological generalizations, and are often assumed to be only categorical. Unlike categorical phonology, variable phonological patterns are typically conditioned by a wide array of factors (Bayley, 2002). Models which use analogy or spreading activation can easily capture the effects of many different factors on the outcome of a production or judgment. Because such models refer only to specific instances of lexical items and not to any features or classes of lexical items, it is actually *more* difficult for these models to capture a pattern the less different factors affect it. In contrast, for models with abstract generalizations, the addition of more generalizations, or the addition of more features to the generalizations, makes the system more complex.

1.2 Probability-Matching behavior

Importantly, it is not just categorical phonological patterns which are productive. Probabilistic trends can also be generalized to novel words. Ernestus and Baayen (2003) examined the distribution of voicing in Dutch root-final obstruents (stops and fricatives), finding that participants in a production task mimicked the lexical distributions in their choices of voicing for experimental items.

In Dutch, all word-final consonants (excluding sonorants) are voiceless, but in some words the addition of a suffix causes that final consonant to become voiced. For example, the words 'widen' (spelled *verwijd*, pronounced [vɛrʋɛit]), and 'reproach' (*verwijt*, [vɛrʋɛit]) are homophones in isolation, but when suffixed, the final consonant

of *verwijd* becomes voiced ([vɛrʋɛidən]), while the final consonant of *verwijt* does not ([vɛrʋɛitən]).

Obstruents differ in how often they are voiced. Labial stops are very rarely voiced (only 9% of labial stops are [b], the rest are [p]), while velar fricatives are almost always voiced (97% velar fricatives are [ɣ], while only 3% are [x]). Ernestus and Baayen tested the productivity of these statistical generalizations by giving subjects novel verbs in the first-person singular present, in which all final obstruents are realized as voiceless (e.g. [ɪk dɑu̲p̲]), and asking them to produce the past tense form, in which the root-final obstruent can vary in voicing (e.g. [ɪk dɑu̲p̲-te] or [ɪk dɑu̲b̲-de]). Subjects voiced consonants in proportion to how often those consonants are voiced in the lexicon.

Table 1.1. Lexical probabilities vs. production probabilities, from Ernestus and Baayen, 2003, pp. 9 and 15

	% voiced in lexicon	% voiced in production
p/b	9%	4%
t/d	25%	9%
s/z	33%	23%
f/v	70%	49%
x/ɣ	97%	80%

Subjects in this experiment did not choose one output option or the other (voiced or voiceless consonants) categorically, and they also did not choose each option equally often. Instead, they probabilistically followed the trends in the lexicon. The percentage of time participants voiced a particular type of consonant roughly matches the percentage of time that type of consonant is actually voiced in the Dutch lexicon. The probabilities are not identical, but crucially the order of preference for voicing is preserved: in the lexicon x/ɣ is voiced more often than f/v, which is voiced more often than s/z, etc. The same order obtains in the production data. In this sense, participants have 'matched' the voicing probabilities present in their lexicon. Similar probability-matching behavior has been observed repeatedly for probabilistic generalizations in languages' lexica, both in productions and in acceptability ratings for

nonce words (notably: Zuraw, 2000; Hayes et al., 2009; Becker et al., 2011; Albright and Hayes, 2003).

1.3 Abstract Knowledge

Probability-matching behavior presents an interesting challenge to models of phonological knowledge which rely on abstract generalizations. Models which account for phonological productivity by way of abstract generalizations include systems of rewrite rules ("[s]→[z] when next to a voiced obstruent")(Chomsky and Halle, 1968) and systems of constraints ("adjacent obstruents must agree in voicing") which can potentially conflict, and whose conflicts are resolved by a ranking relationship between constraints (Prince and Smolensky, 1993/2004). Both types of models are typically used to model categorical grammatical knowledge only. If there is variation in surface forms, this can be modeled as a categorical generalization with specific exceptions, in which case the pattern should generalize categorically to new forms and the exceptions should not generalize. Alternatively, both variants can be produced as acceptable by the grammar, in which case neither variant should be preferred in a novel word. Any preference for one form over the other must be due to a qualitatively different, extra-grammatical mechanism, such as the production or perception system. Dual-route models (Pinker, 1999) divide categorical and probabilistic generalizations in a similar way: categorical generalizations are captured by a system of abstract rules, while probabilistic trends are captured by an analogical mechanism operating over the lexicon.

More recently, researchers have sought to use models with abstract generalizations to provide a unified explanation of categorical and probabilistic behavior. Coetzee and Pater (2009) provide a comprehensive overview of this literature. Models adapted to match probabilistic knowledge include variable rules (Labov, 1969; Cedergren and Sankoff, 1974; Guy, 1991)which include in their structural description a statement

about how often they should apply under what circumstances, and various adaptations of Optimality Theory. Partially Ordered Constraints (Anttila, 1997 et. seq.) is a framework which predicts the rate of occurrence of a particular class of output based on the number of different constraint rankings which can produce it. Stochastic OT (Boersma and Hayes, 2001), Noisy Harmonic Grammar (Coetzee and Pater, 2009), and Maximum Entropy (MaxEnt) Grammar (Goldwater and Johnson, 2003) are all OT-based models which use numerical weights on constraints to predict exact rates of application of a process. For example, Coetzee and Kawahara (2013) use a MaxEnt model to predict English speakers' rates of t/d deletion. Constraints are each assigned weights, and when constraints compete their relative weight determines the probability that the outcome will observe one constraint or the other. In this way, they can use abstract constraints to predict particular probabilities for particular outcomes. Their model could predict, for example, that t/d will delete 30% of the time in a particular context.

Ernest and Baayen use Stochastic OT to predict the probability-matching behavior in their data. They employ constraints such as "A velar fricative should be voiced", "A labial stop should be voiceless" which interact with opposing constraints like "A velar fricative should be voiceless" to predict probability distributions over voicing alternatives for each nonce word in their experiment. These predicted distributions correlate well with the distributions actually produced by participants.

1.4   Epiphenomenal Knowledge

A lexicon and some lexical access mechanism are necessary components of the linguistic system. Therefore, if patterns can be generalized from the lexicon without the use of any additional mechanism, then abstract representations of those patterns would be redundant. An example of this economy argument comes from the TRACE model of lexical access (McClelland and Elman, 1986). The model is designed to

account for how listeners use acoustic information about features to resolve phonemes, and ultimately words. However, without the addition of any extra machinery, the model can mimic the effects of abstract phonotactic generalizations. Nonwords which violate a phonotactic generalization will be perceived differently than legal nonwords, since legal nonwords will be phonologically similar to more actual words. For example, no actual English words begin with 'sr', but many words begin with 'sl', so a nonword beginning with 'sl' will generate activation on several real words, while a nonword beginning with 'sr' will not.

TRACE and other abstraction-free models do require extra machinery to capture generalizations more complex than a simple string of phonemes like '*sr'. For some phonotactic generalizations, prosodic structure such as syllables and feet is required to capture the lexical trends and/or productivity data (Kager and Pater, 2012 is one notable example). In other cases, models require special generalizations about which features of the word are the most important to analogize to (for example: the beginning of the word is more important than the end, consonants are more important than vowels, etc.)

Ernestus and Baayen fit a spreading activation model to their data. The spreading activation model contains three layers: an output layer with two units, voiced and voiceless; a second layer which contains the lexical exemplars, and an initial layer with three units specifying particular features of the input: the place of articulation of the word's final consonant, the quality of the word's final vowel, and the type of segment that precedes the final consonant (vowel, obstruent, or sonorant). Activation flows from each feature to the lexical items which match that feature, and from there activation flows from each lexical item to the voicing specification of its final consonant. The features are abstractions which serve to privilege some aspects of the existing lexical entries - namely their rhyme features - over others.

1.5 Economy

In the domain of syntax, it is necessary to posit the existence of abstract rules in order to account for the extreme generative power of natural language. Chomsky's (1957) famous example sentence "Colorless green ideas sleep furiously." illustrates this. The sentence is a well-formed sentence of English even though the individual words that comprise it virtually never occur together. Although there are certain specific sentences that occur over and over again in natural speech (such as greetings) sentences which are completely novel abound.

In the domain of word-level phonology, the same is not true. Instead, the majority of uttered words are forms which could potentially be lexical forms, and truly novel lexical items - items that the speaker has never previously been exposed to - only occur in exceptional situations. For this reason, it is tempting to posit a phonological system that is simplified compared to the syntactic system: in which phonological productivity is the result of direct reference to known lexical items, and there is no implicit 'knowledge by description' in the form of abstract grammatical generalizations. Instead of a grammatical component, independently necessary cognitive mechanisms would account for phonological productivity. For example, McClelland and Elman (1986) develop a model of the lexical access process which also predicts, essentially as a side effect, that generalizations in the lexicon should affect speakers' perception of novel words. Spreading-activation models (Rumelhart and McClelland, 1986) and analogical models (Skousen, 1989; Daelemans et al., 1994; Nakisa et al., 2001) do posit some extra mechanism beyond the lexicon, but one without explicit representation of any generalizations.

Although epiphenomenal knowledge is often argued for in place of abstract generalizations on the grounds that it requires less extra machinery (McClelland and Elman, 1986; Eddington, 2000; Skousen, 1989), it is actually somewhat difficult to compare the complexity of the machinery used in abstract models with that used in epiphenom-

enal models. Statistical tests which incorporate the complexity of two models into the calculations typically require that the two models be in a subset-superset relationship, so that the comparison is between a 'null' model and an alternative model which is identical to the null model except that it has an extra parameter. Epiphenomenal models and abstract models are incommensurable in this system since they are not in a subset-superset relationship. Ernestus and Baayen compare the complexity of their OT-style model and their spreading activation model by counting the number of parameters: 10 or 20 constraints in the OT-style model, and 3 input features in the spreading activation model. However, constraints and input features are quite different types of parameters - it is not clear that the addition of one constraint adds the same amount of complexity to the OT model as the addition of one feature adds to the spreading activation model.

Ernestus and Baayen point out that their parameter counting metric is flawed on other grounds, namely that it does not take into account other types of complexity such as the complexity of the learning process or the computational time required to produce a given output (which can be quite long for certain analogical models). Since a model of phonological knowledge is obliged to account for all human language data, rather than just a single pattern, it is important to consider a model's ability to model many different kinds of actually existing patterns. If model A captures some patterns more simply than model B but requires a lot of additional machinery for other patterns, which one is simpler? Since both epiphenomenal models and abstract models can each capture certain patterns very simply and require more complexity for other patterns, it is difficult to discern their relative complexity based on a single pattern.

1.6 Model Fit

Epiphenomenal models and abstract models account for probability matching behavior in qualitatively different ways, but the degree to which each type of model can actually fit the data is very similar. While epiphenomenal models typically have an easier time capturing the variety of factors which can influence a form's well-formedness, abstract models have an easier time capturing generalizations selectively - meaning that they can capture some but not all trends in the lexicon. In Ernestus and Baayen's data, the place of articulation of the word's final consonant is only one of the factors that appears to condition the choice of voicing for the final consonant (both in the lexicon and in subjects' responses). Other factors are (1) The segment immediately preceding the final consonant: words with immediately preceding vowels and sonorants are more likely to voice than words with a final obstruent cluster. (2) The quality of the word's final vowel: words with long vowels are less likely to voice. A model capturing these preferences via abstract generalizations would require a separate abstraction for each conditioning factor. In general, models with abstract generalizations increase in complexity as the number of conditioning factors increases. An epiphenomenal model, on the other hand, typically does not increase in complexity between the case with a single conditioning factor (like word-final voicing agreement in English) and the case with many conditioning factors.

In fact, epiphenomenal models typically cannot help but generalize any preference present in the lexicon (stronger preferences generally being expressed more strongly). Pinker and Prince (1988) criticize connectionist models of phonological productivity on exactly these grounds, claiming that at least in some cases these models' attention to detail is too fine-grained. They discuss the case of categorical phonological generalizations such as the generalization mentioned above that in English word-final consonant clusters must agree in voicing (cats [kæts but dogs dɔgz). They point out that this generalization is observed 100% of the time, in nonce words as well as real

words, but the connectionist model would predict occasional violations in plural nonce words if other properties of the singular besides the voicing of the final consonant were strong predictors of the voicing of the plural suffix.

Despite these rather different qualitative predictions, the two types of models are both fairly good at fitting a given set of production data. Ernestus and Baayen look at the correlation of each model's predictions with their observed data, and find that for both the OT-style model and the spreading activation model, the correlation was fairly high, and also that the two correlations were very similar (between 0.80 and 0.85).

1.7 Distinguishing between Abstract and Epiphenomenal knowledge

In the following chapters, I explore alternatives to using model fit or economy as arguments to distinguish between models using abstract vs. epiphenomenal knowledge. I compare the predictions of a model in which probabilistic generalizations are part of a grammatical component, as in Figure 1, with the predictions of a model in which they are not (either in which there is no grammatical component, or in which that component contains only abstract generalizations). If probabilistic phonological generalizations are grammatically encoded, then (1) those generalizations should influence not just judgments of novel words but also production and perception systems, (2) speakers should be able to use probabilistic generalizations in judgment, production, and perception even when an alternative strategy, such as analogy, is not available, and (3) the probabilistic trends produced by speakers may mismatch the lexical statistics because of restrictions on the form or substance of the rules or constraints involved, or because of biases in the learning process.

In chapter 2, I investigate the influence of a probabilistic trend on the perception of both known and novel words. I find that speakers process known words which are

16

violators of the trend differently from known words which observe the trend. This difference in processing happens in the early stages of lexical access (280-380ms), and may be attributable to differences in the lexical representations of trend observers vs. trend violators. Alternatively, the processing difference may arise from a mismatch between the expectations of the grammar and the contents of a lexical entry. In chapter 3, I investigate speakers' knowledge of a trend within the English stress system, testing prediction (2) above. Novel words which had no near lexical neighbors were used. Additionally, participants provided potential analogical bases fr each nonword. Speakers extend the trend in the lexicon to novel words, and do not use the properties of the analogical bases to do so.

In chapter 4, I investigate a case of mismatch between the lexical statistics and participants' productions. Participants extend a near-categorical trend in the lexicon to nonwords, but they undermatch it, producing significantly more exceptional forms in the experiment than are found in the lexicon. Using a MaxEnt model of this pattern, I argue that this mismatch can be explained as participants' using a constraint set that is simplified relative to the constraint set which produces a close match with the lexical data. In order to perfectly match the statistical trends in the lexicon, English speakers would have to induce a complex constraint. I argue that a simpler and more general version of this constraint also provides a close enough match to the lexicon, and therefore that is the constraint which is part of speakers' grammars. A bias towards learning or using simpler constraints - constraints referencing fewer features - has been argued for in the domain of phonology by Moreton and Pater (2012). The analysis in chapter 4 provides additional evidence for such a bias, and provides specifically evidence from a natural language rather than from artificial grammar experiments.

CHAPTER 2

EFFECTS OF PROBABILISTIC PHONOLOGY ON THE PERCEPTION OF WORDS AND NONWORDS

2.1 Introduction

The process of perceiving speech sounds is strongly conditioned by the listener's experience with his or her language - sounds and sequences of sounds which do not occur in a language may be difficult or impossible for listeners to viridically perceive (Berent et al., 2009; Dupoux et al., 1999; Dehaene-Lambertz et al., 2000; Moreton, 2002; Breen et al., 2013). Likewise, when inflecting or producing novel words, speakers avoid structures which are not part of their language (which do not occur in their lexicon). Speakers extend to nonwords not just categorical generalizations, but also probabilistic ones. Ernestus and Baayen (2003) show that speakers 'probability match', mimicking the probability distribution in their lexicon over certain output types in their productions of nonwords. Although probabilistic generalizations are part of speakers' knowledge about their language, to date no evidence has turned up suggesting that perceptual illusions are induced by probabilistic constraints. One could easily imagine that this is because listeners have to be able to accurately perceive violations of them in order to correctly perceive all the words of their language.

I investigate a probabilistic generalization in the English stress system, showing that speakers' knowledge of this generalization does affect their perception of real English words. In the early stages of lexical access, listeners appear to have more difficulty accessing the stress pattern of exceptional words than the stress pattern of generalization-observant words. This result demonstrates that listeners' knowledge of

probabilistic generalizations impacts their perception even of known words for which there is a lexically stored form.

Berent et al., Breen et al., and Moreton argue that the perceptual illusions they observe are the result of the action of a grammatical system rather than a result of listeners' knowledge of the statistics of their lexicons. Berent et al. (2009) found that listeners perceive a schwa intervening between two word-initial consonants which form an illegal cluster in English (e.g. [lbif] is perceived as [ləbif]). Moreton (2002); Breen et al. (2013) found that English listeners perceptually repaired illegal [dl/tl] clusters to legal [kl/gl] clusters. Each study compares listeners' behavior on an ungrammatical structure (e.g. dl/tl) to listeners' behavior on a structure which is unattested or nearly unattested in the lexicon, but not disallowed (e.g. bw/pw clusters at the beginning of words). All three studies find different rates of perceptual repair for different unattested structures, indicating that their results are not due only to speakers' knowledge of their lexicons.

Investigations into the timecourse of these perceptual repairs indicate that they happen relatively early in processing. Using event-related potentials (ERPs), Dehaene-Lambertz et al. found no evidence for a perceptual stage before the repair had occurred. While Breen et al. did find evidence for two stages (also via ERPs), the repair occurred before about 400ms post-stimulus onset, a time window which is likely during the early stages of lexical access. Domahs et al. (2009); Pitkanen (2010) found effects of phonological grammar somewhat later in processing. Both found a late positivity (LPC) to phonologically ill-formed nonwords relative to phonologically well-formed nonwords. Domahs et al. investigated the ban on words like 'spup' in German, finding a difference in brain potentials between nonwords which violated this constraint vs. nonwords which observed it: a positive-going potential peaking at about 1200ms post-stimulus-onset. Pitkanen investigated vowel harmony in Finnish, and likewise

found a late positivity for nonwords with illegal vowel patterns relative to nonwords with legal vowel patterns. This negativity began at 600ms post-stimulus-onset.

Probabilistic grammatical generalizations have been less well investigated. It is unlikely that a generalization which has exceptions in the lexicon of a language would trigger a perceptual repair, since listeners must be able to correctly perceive all the lexical items of their language. Massaro and Cohen (1983) investigated a generalization with very few exceptions ('sr' cannot begin a word), and found that it could affect listeners' perception of ambiguous sounds. Listeners perceived a sound which was ambiguous between 'r' and 'l' as 'l' more often after 's', where 'r' is very rare, than after 'p' where both are equally common. Böcker et al. (1999) used ERPs to investigate the processing of actual words with common or rare stress patterns in Dutch. They found that words with the rare stress pattern elicited a negativity relative to words with the common stress pattern at about 350ms post-stimulus onset. They argue that their effect is related to early stages of lexical access.

I investigate the influence on perception of a probabilistic generalization in the English stress system. This trend is strongly attested in the lexicon in that it is supported by a large number of forms and has relatively few exceptions. I compare it to a second trend which is weakly attested in the lexicon, in that it is supported by a small number of forms and has many exceptions. The strongly attested trend is 'productive' in speakers' ratings of nonwords. Nonwords that violate this trend are rated as less likely to be English words than nonwords that observe it. The weakly attested trend is not productive - it does not influence speakers' ratings of nonwords. Despite the fact that the strongly attested trend is productive, there is no evidence in this experiment that it influences participants' processing of nonwords. In particular, it does not elicit an LPC of the type found by Pitkanen; Domahs et al.. However, the trend does influence the processing of actual words. Real-word exceptions to the trend elicit an early negativity, at 280-380ms relative to observers of the trend.

I propose that this early negativity is related to the negativity found by Böcker et al., and that it reflects the process of accessing a word's stress pattern. The negativity is larger for words with less common stress patterns because on those words the stress pattern is encoded more strongly.

In a system like the English stress system in which each word has a single correct pronunciation, but that pronunciation is not completely predictable from the probabilistic trends in the grammar, some mechanism must be in place for storing either the stress of all words or minimally of the exceptions. Strategies used in the English stress literature include lexically marked weight Hayes (1980), extrametricality Selkirk (1984), and exceptional segmental content Chomsky and Halle (1968); Burzio (1994). Other strategies include using lexically specific faithfulness constraints (Pater, 2005; Becker, 2009), so that each word is indexed to a faithfulness constraint, indicating that that word either obeys or violates a pattern encoded in the markedness constraints. Moore-Cantwell and Pater (2015) build on this approach by allowing each word's lexical representation to have some weight and to compete with grammatical constraints. Words with exceptional stress patterns have a higher weight on their lexical representation of stress in order to overcome the pressure of the phonological grammar.

The CMU pronouncing dictionary (Weide, 1994) together with the SUBTLEX$_{US}$ corpus (Brysbaert and New, 2009) was used to examine the distribution of stress in long words of English, beginning with trisyllabic all-light words. Words with heavy syllables were excluded because heavy syllables (especially heavy penultimate syllables, (Chomsky and Halle, 1968)) tend to attract stress in English (Guion et al., 2003). For purposes of this search, a light syllable is defined as one (a) with no coda consonant, and (b) whose vowel is a monophthong (one of [ɑ, æ, ɛ, ɔ, ɪ, ʊ, i, u, ə, ɚ]).

The maximal onset principle was used to determine syllabification[1], so e.g. 'palmistry' is treated as [pɔ.mɪ.stɹi], and included in the set of all-light words. However, a word like 'bandana' ([bæn.dæ.nə]) would not be included, since its first syllable must have a coda consonant - [nd] is not a legal onset in English.

Many trisyllabic all-light words are morphologically complex. Because particular affixes can influence the stress pattern of a word, either by attracting stress or by behaving as extrametrical (Chomsky and Halle, 1968; Burzio, 1994; Halle and Vergnaud, 1987), I attempted to exclude words containing productive morphology by excluding from our search words that ended orthographically in '-ly', '-er', '-ity', or '-ery' (the latter excluded words like 'gunnery'). Of the remaining words, many are morphologically complex, but contain morphemes that are neither productive nor easily split apart: e.g. 'circuitry', 'frequency', 'brevity'. Only words frequent enough to occur at least once in SUBTLEX were included in the search results, and log frequencies were recorded from there.

A total of 771 words met the criteria of being all-light, trisyllabic, and having a single stress. These ended in either [-i], [-ə], or a syllabic consonant such as a nasal ('positio<u>n</u>') or liquid ('ename<u>l</u>', 'vinega<u>r</u>'). The single stress was always antepenultimate or penultimate - never final[2]. How likely each stress pattern was depended on the final vowel. Words ending in [-i] were more likely to take antepenultimate stress than penultimate stress, but words ending in [-ə] were more likely to take penultimate stress than antepenultimate stress. Words ending in syllabic consonants will not be discussed here.

---

[1]Thanks to Robert Staubs for the use of his software implementation of the maximal onset principle

[2]There were four words which were coded as finally stressed in CMU. They were 'misconstrue', 'reoccur', 'honoree', and 'disagree'.

Figure 2.1. Counts of three syllable, monomorphemic, all-light words (one stress only), from the CMU pronouncing dictionary (North American English). Only words which end in either [i] or [ə] are included.
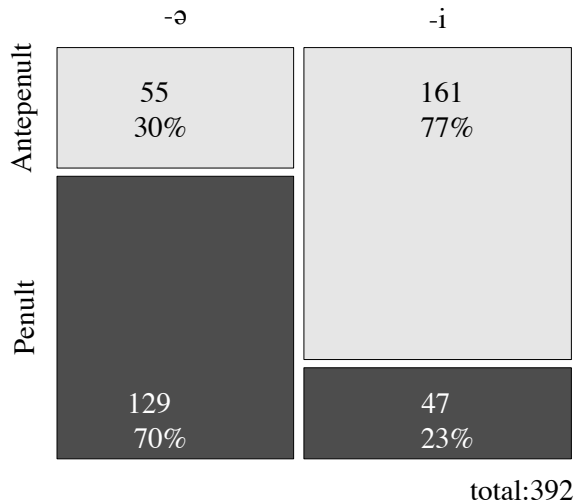
Each category of word above contains words of comparable token frequencies. Mean log frequencies of words in each category are reported below, gathered from SUBTLEX.

Table 2.1. Mean (standard deviation) log frequencies for three syllable, monomorphemic, all-light words. Log frequencies are taken from SUBTLEX and words themselves are gathered from the CMU pronouncing dictionary.

	Antepenult.	Penult.	Total	(example)
-[i]	1.62 (.73)	1.54 (0.74)	1.58	(canopy, spaghetti)
-schwa	1.37 (0.72)	1.21 (.61)	1.29	(cinema, vanilla)
all	1.51	1.45	1.48	

In this constrained search space, there is a trend for i-final words to take antepenultimate stress, and a comparably strong trend in the opposite direction, for ə-final words to take penultimate stress. In order to understand how these trends interact with the lexicon as a whole, the search space was expanded to include longer words, multimorphemic words, and finally words with heavy syllables. When the scope of

the search is expanded in any of these ways, the i-final trend becomes stronger while the ə-final trend becomes weaker.

Figure 2.2. Counts of all words three syllables long and longer, from the CMU pronouncing dictionary (North American English). Only words whose final vowel is either [i] or [ə] are included.
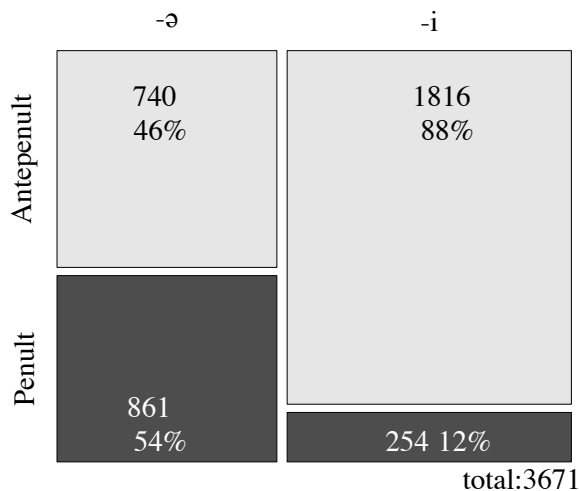
The i-final trend for antepenultimate stress is more strongly attested in the lexicon than the ə-final trend for penultimate stress. The i-final trend holds over a large portion of the lexicon, about 2000 words, is close to categorical (about 90% observance), and is not conditioned by morphological structure or syllable weight. The ə-final trend holds only over a small portion of the lexicon, about 200 words, is less categorical in the domain in which it holds (about 70% observance), and is conditioned by both morphological structure and syllable weight - words with derivational morphology and words with heavy syllables do not exhibit the trend. Any of these differences between the trends might bias an acquisition system to acquire the strong i-final trend and not the weak ə-final trend.

2.2 ERP background

Electroencephalogram (EEG) is obtained by using electrodes positioned on a participant's scalp to measure electrical potential at the scalp at the location of each electrode. The electrical potential at the scalp reflects the action of large groups of spatially aligned neurons firing together. Each neuron's post-synaptic potential contributes to an electrical dipole whose charge spreads out through the brain, skull, and scalp, and appears as at the scalp as a relatively diffuse region of positive or negative electrical potential, depending on the orientation of the dipole. Because of this diffusion, and because more than one dipole may be active (and almost certainly is) at any given time, it is very difficult to localize within the brain the actual group or groups of neurons contributing to a particular effect. However, because neurons' charge can change very fast (on the order of tens of milliseconds) and the charge moves from its origin to the scalp at the speed of light, EEG has very good time resolution.

Event-related potentials (ERPs) are obtained by averaging together segments of EEG collected from many individual trials of an experiment. Each segment begins at the onset of a particular stimulus (or potentially at some point of interest within that stimulus, say the onset of the second syllable). All segments from a particular condition of the experiment are averaged together to create an average wave, called an ERP. In the averaging process, brain activity that is not related to the event - the stimulus - will cancel out across many trials. ERPs to different conditions of an experiment can then be compared. A difference in potential between two conditions at a particular time after the onset of a stimulus suggests that the processing of the two types of stimuli differs at whatever stage of processing occurs at that time. Certain waves peaking at particular characteristic times after a stimulus onset have been associated through a series of many experiments with particular cognitive processes. An example relevant to the experiment presented below is the N400: a negative going potential peaking at about 400 ms post-stimulus-onset (Kutas and Hillyard, 1980).

It is associated with the perception of open-class words (content words), and is larger in amplitude when the word is semantically irregular, or low-probability in some way, such as being low frequency (Rugg, 1990). Other examples include the auditory potentials N1, and P2, which occur at about 100ms and 175ms post-stimulus onset respectively (Näätänen, 1992). They occur whenever a stimulus contains a sharp onset of sound, such as a beep or the beginning of a word.

2.3 Experiment 1 Methods

2.3.1 Participants

Participants were 22 right-handed native speakers of English, between the ages of 18 and 30. All had normal hearing, normal or corrected to normal vision, no known neurological conditions, and all were undergraduate or graduate students at the University of Massachusetts Amherst.

2.3.2 Materials

For each of the four types of word (-i final antepenultimate stress, -i final penultimate stress, -ə final antepenultimate stress, -ə final penultimate stress) given above 10 English words were selected, shown in Figure 2.2. All had three syllables, all of which were light: they had no coda consonants, and all the vowels were monophthongs. Some syllables did begin with consonant clusters (e.g. trave<u>sty</u>). Words that were morphologically complex or could be analyzed as morphologically complex (e.g. *basketry*) were avoided.

In the mis-stressed versions of these words, the stress was moved from the penultimate syllable to the first syllable, or from the first syllable to the penultimate syllable (no items had stress on the final syllable). Because vowels are reduced in stressless syllables in English, the syllable to which stress was moved always had a schwa as its vowel in the real word. Full vowels were chosen for each stressed syllable so that

Table 2.2. English words used in the experiment. The second and third columns are trend-violating words, and the first and last columns are trend-observing.

	-[i]		-[ə]	
	Antepenultimate	Penultimate	Antepenultimate	Penultimate
	elegy	salami	cinema	dilemma
	atrophy	finale	africa	eureka
	canopy	graffiti	nebula	lasagne
	ebony	jalopy	replica	manila
	fallacy	pastrami	retina	sierra
	cavity	bikini	spatula	sonata
	colony	safari	stamina	vanilla
	recipe	swahili	swastika	savanna
	remedy	spaghetti	taffeta	militia
	travesty	zucchini	canada	alaska
mean (sd) log(f)	*1.65(0.53)*	*1.76(0.32)*	*1.78(0.48)*	*1.72(0.46)*

(a) all full vowels were monophthongs (b) the distribution of vowel qualities in the mis-stressed words was similar to the distribution in real words, and (c) orthography was observed as much as possible. These mis-stressed words were produced by a native speaker of English; no acoustic manipulation was performed. Some example real word - mis-stressed word pairs are given in 2.3.

Table 2.3. Example mis-stressed items

Spelling	real word IPA	mis-stressed word IPA
elegy	έlədʒi	əlέdʒi
salami	səlámi	sǽləmi
cinema	sínəmə	sənέmə
dilemma	dəlέmə	díləmə

Nonwords were constructed by mutating the consonants of a real word, but keeping the vowels constant. Each consonant was changed by at least one phonological feature, but the consonants were always kept within the same sonority class: stops were changed to other stops, fricatives to other fricatives, nasals to other nasals, and liquids to other liquids. The lexical neighborhood density of each nonword was measured

using the Generalized Neighborhood Model (Bailey and Hahn, 2001)[3]. All nonwords used in the experiment had a GNM value of less than 0.01, corresponding to very sparse neighborhoods.

Table 2.4. Example nonce words

Spelling	real word IPA	nonce words IPA
elegy	ɛ́lədʒi	ɛ́ɹətʃi / əɹɛ́tʃi
salami	səlɑ́mi	fəɹɑ́ni / fǽɹəni
cinema	sínəmə	θímənə/ θəmɛ́nə
dilemma	dəlɛ́mə	kəɹɛ́nə/ kíɹənə

All stimuli, words, mis-stressed words, and nonwords, were transcribed into the international phonetic alphabet (IPA) and pronounced in a random order. The speaker was a native speaker of English, and each word was spoken in the frame sentence "Say X again." The words were then spliced out of the frame sentence and presented in isolation in the experiment.

Acoustic measurements were taken of all the stimuli, to make sure that the acoustic cues to stress placement were comparable in the different types of words. Beckman (1986) discusses duration, intensity, F0 excursion (maximum F0 - minimum F0), and vowel quality as cues to stress in English.

Table 2.5. Acoustic measurements of stressed syllables in all experimental items

Realized Stress	Parameter	Word	Mis-stressed	Nonword	Statistics
Antepenultimate	Duration (ms)	97 (20)	102 (27)	99 (21)	$F_{(2,77)}=0.23$
	Intensity (dB)	74 (3)	74 (2)	75 (3)	$F_{(2,77)}=1.1$, p=0.35
	F0 excursion (Hz)	6.6 (4.3)	7.3 (5.7)	10.4 (13.2)	$F_{(2,77)}=1.3$, p=0.29
Penultimate	Duration (ms)	106 (24)	99 (24)	98 (25)	$F_{(2,77)}=0.79$
	Intensity (dB)	74 (3)	73 (3)	73 (3)	$F_{(2,77)}=0.67$
	F0 excursion (Hz)	17.5 (28.1)	8.2 (4.14)	14.9 (21.1)	$F_{(2,7)}=1.1$, p=0.34

---

[3]Thanks to Adam Albright for sharing his implementation of the GNM model

2.3.3 Procedure

Participants were seated comfortably in front of a computer screen. Each trial began with a fixation cross, during which the stimulus was presented auditorily. Following the stimulus presentation, participants rated how likely each word was to be an actual word of English, responding on a 1-4 scale using a button box. Participants were told that they would hear actual words and made up words mixed together, but that some of the real words might be very unusual and they would not have heard them before. Two examples of very rare actual words were given: 'gálea', and 'tomálley' (both trend-violating). Participants were told that when they didn't recognize the word, they should use the sound of the word to guess how likely it was to be an English word.

Each trial began with a fixation cross varying in duration from 100 to 1000 milliseconds. The fixation cross then persisted throughout the presentation of the auditory stimulus, and lasted until the response screen appeared, 1500 milliseconds after the stimulus onset. The mean duration of the stimuli was 486 ms (sd= 69ms, min=307ms, max=650ms). Participants were instructed to hold still and avoid eye movements or blinking from the time the fixation cross appeared until they saw the response screen. The participant controlled when each trial began by pressing a button on a 'Ready' screen. Participants were allowed to blink and move their eyes during the response screen and before the beginning of each trial.

Stimuli appeared in 5 blocks of either 160 or 320 trials each. Subjects 1 and 2 saw 320 trials in all five blocks, and remaining subjects saw 160 trials in the first two blocks, and 320 trials in the last three blocks for a total of 1280 trials (1600 trials for subjects 1 and 2). Each item appeared once (blocks 1 and 2) or twice (blocks 3-5) per block, and order within each block was randomized. The entire duration of the experiment was about 2.5 hours.

2.3.4 EEG recording

Electroencephalogram (EEG) was recorded continuously by means of a 128-electrode HydroCel Geodesic Sensor Net (Electrical Geodesics Inc., Eugene, OR), at a sampling rate of 250Hz and a bandwidth of 0.01-100 Hz. For recording, the reference electrode was placed at the top of the head (Cz). The impedance for each electrode was maintained below 50kΩ throughout the experiment. A 60 Hz notch filter was applied to the EEG signal offline after recording. The continuous EEG was divided into 1600ms epochs, timelocked to (a) the beginning of the stimulus, (b) the onset of the second syllable, and (c) the onset of the third syllable. Relative to each onset, the epochs began 100ms before that onset, and continued 1500ms after. Because all effects of interest appeared timelocked to the first syllable, I do not discuss data from epochs timelocked to the second and third syllables.

Trials containing eye blinks, eye movements, or other artifacts were excluded automatically using ERPLAB. Eye blinks were detected by subtracting voltages below each eye from voltages above each eye, and eye movements were detected by subtracting voltage to the left of the eyes from voltage to the right of the eyes. Subject-specific voltage thresholds were chosen for eye blink/movement channels (between 300 and 500 μV) as well as for the whole head (between 400 and 500 μV). Any trial where the threshold was exceeded was marked as artifactual. In all conditions, between 22% and 27% of trials contained artifacts.

Table 2.6. Onset times for each syllable of the stimuli, in milliseconds from stimulus onset

	s2 onset	s3 onset	end
mean	177	346	486
min	70	223	306
max	364	500	650

Figure 2.3. Twelve regions of interest, with 9 electrodes in each, used for analysis of Experiments 1 and 2. The 128 electrodes were divided up into 12 regions, each containing 9 electrodes. Average ERPs were calculated, the average of all 9 electrodes in a region. Electrodes which were excluded from analysis, including Cz and the mastoid references, are represented by a grey dot.
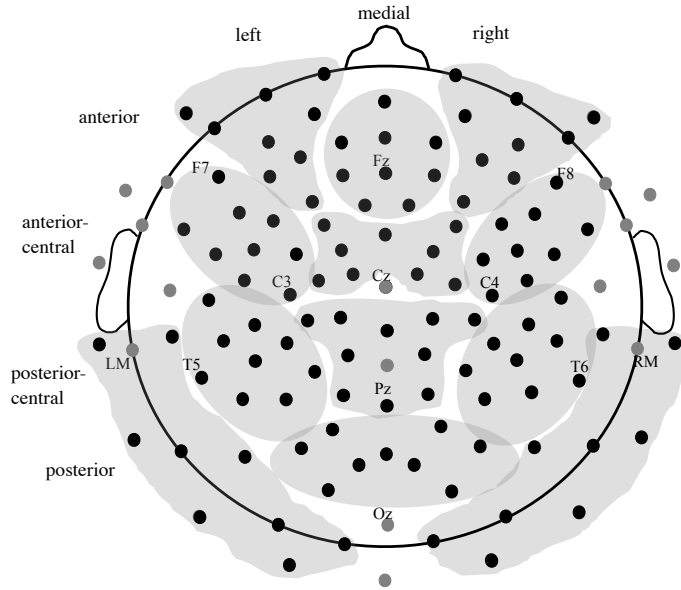
2.3.5 Analysis

ERPs were computed for each participant, condition, and electrode, and were re-referenced to the average of the two mastoid electrodes, and baseline-corrected using the 100ms pre-stimulus-onset as a baseline. The 128 electrodes were divided into 12 regions, each containing 9 electrodes. The regions were chosen so as to divide the head up symmetrically into left, medial, and right, and into four degrees of anterior-posterior. These regions are illustrated in Figure 2.3. Average ERPs were calculated for each region, averaging together all 9 electrodes in each region.

Mean voltage was calculated at four time windows of interest, which were chosen based on visual inspection of the averaged waveforms. Those time windows were 120-180ms, corresponding to the peak of the N1 auditory component; 280-380ms, corresponding to an early difference between violators and observers; 400-1000ms, corresponding to the N400, and 1000-1500ms, corresponding to the Late Positive Component (LPC). The time window for both the N400 and LPC are somewhat later

than is usually found in the literature on these components - this delay is expected since the stimuli in the experiment are somewhat long, and the acoustic information about a word's status as a trend-observer or violator is distributed over the entire duration of the word.

Repeated-measures ANOVAs were calculated using as factors Prosodic Form ([i]-final Antepenultimate, [i]-final Penultimate, [ə]-final Antepenultimate, [ə]-final Penultimate), Word Status (Word, Misstressed, Nonword), Anteriority (Anterior, Anterior Central, Posterior Central, Posterior), and Laterality (Left, Medial, Right).

2.4 Experiment 1 Results

2.4.1 Ratings

Average ratings for each type of word presented in the experiment are shown in Figure 2.4. Ratings range from 1 (very unlikely to be an English word) to 4 (very likely to be an English word). A within-subjects ANOVA was conducted with mean ratings from individual subjects as the dependent variable[4]. Factors were WordStatus (Word, Mis-stressed, Nonword), FinalV (-[ə], -[i]), and Stress (antepenultimate, penultimate).

As expected, the ratings for nonwords were much lower than the ratings for actual words, and the mis-stressed words were somewhere in between (In the ANOVA, there was a main effect of word type $F(2, 42) = 149.97$, $p \ll 0.05$). Ratings for mis-stressed words were much more variable than the other ratings, both within and across subjects. Ratings for actual words were nearly always 4, with very little variability. Two experimental items, the words "jalopy", and "taffeta", were an exception to this, receiving average ratings of 2.9 and 2.7 respectively (average ratings for the other words ranged from 3.5 to 4). Ratings for these words as well as their mis-stressed counterparts were excluded from the analysis.

---

[4]These mean ratings were roughly normally distributed for the mid-stressed and nonwords, but not for real words, whose ratings were clustered at the high end of the scale

Figure 2.4. Mean ratings. Error bars represent standard errors based on the distribution of individual subject means.
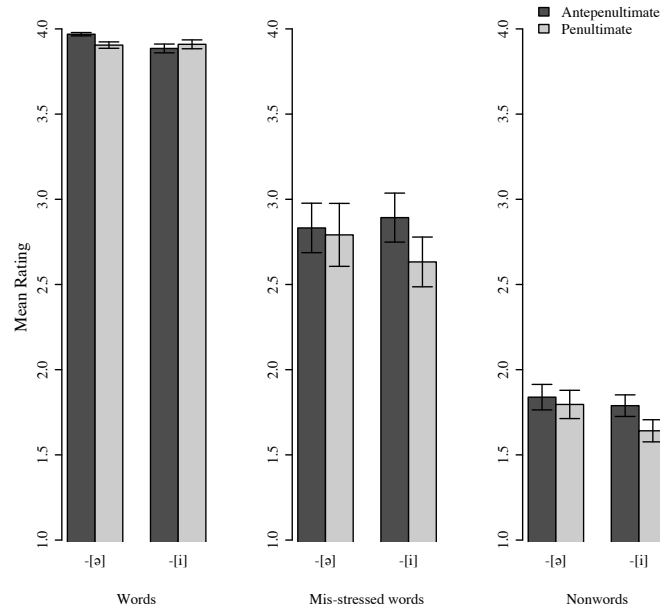
Antepenultimately stressed items received higher ratings than penultimately stressed items in both mis-stressed words and nonwords, but not in real words (main effect of Stress: $F(1,21) = 9.61$, $p < .05$ WordStatus x Stress: $F(2,42) = 4.24$, $p < .05$).

In nonwords, the difference between antepenultimate and penultimate stress was bigger for items ending in -[i] than for items ending in -[ə]. (WordStatus x FinalV x Stress: $F(2,42) = 4.44$, $p < 0.05$). The mean rating for antepenultimately stressed [i]-final words was 0.14 points higher than the mean rating for their penultimately stressed counterparts. However, the difference in means between antepenultimately and penultimately stressed [ə]-final nonwords was only 0.04 points.

2.4.2 ERP data

Figures 2.5-2.10 show the grand average ERP's timelocked to the onset of the word. Each electrode pictured is the average of 9 electrodes, calculated over the regions of interest pictured in Figure 2.3 above. All experimental items elicited the

typical sequence of P1-N1-P2, with the earliest positivity peaking at 80ms, the earliest negativity at 150ms, and the second positivity at 210ms.

2.4.2.1 Early effects

Both real words and mis-stressed words (Figures 2.5 and 2.6) exhibited early differences based on whether the word was trend-observing or trend-violating, regardless of which trend it observed or violated. Trend-violating real words (e.g. áfrica, spaghétti) as well as trend-violating mis-stressed words (e.g. bánana, recípe) exhibited a negativity relative to their trend-observing counterparts, from 280-380ms. An ANOVA conducted on only medial electrodes revealed a significant effect of Prosodic Form (four levels: [i]-final Antepenultimate, [i]-final Penultimate, [ə]-final Antepenultimate, [ə]-final Penultimate) for both real words (F(3,60)=4.9, $p < .05$), and mis-stressed words (F(3,60)=2.8, $p < .05$). Additionally, a contrast of the two trend-violating patterns ([i]-final Penultimate, [ə]-final Antepenultimate) vs. the two trend-observing ones ([i]-final Antepenultimate, [ə]-final Penultimate) was significant for real words (t=4.94, $p < .05$), and for mis-stressed words (t=-2.31, $p < .05$).

In mis-stressed words only, there was an even earlier effect of trend-observing vs trend-violating items, on the N1 peak. The mean amplitude (120-180ms) of this peak was significantly higher for trend-violating than for trend-observing items over left and medial, posterior-central, central and anterior electrode sites (an ANOVA found a significant main effect of prosodic form: F(3,60)=2.9, $p < .05$, and a contrast of trend violating vs trend-observing was also significant (t=-3.73, $p < .05$). Antepenultimately-stressed trend-violating misstressed words (bánana) elicited a more negative N1 than penultimately-stressed trend-violating misstressed words (recípe) (t=-3.79, $p < .05$).

These early effects occur well before any acoustic information about the final vowel is present (starting at about 300ms).

Figure 2.5. Early differences between trend-observing and trend-violating patterns in words (left) and mis-stressed words (right). No such effect was found for non-words. In both words and mis-stressed words, there was a greater negativity to trend-violating than trend-observing items between 280-380 ms post-stimulus onset (before information about the final vowel is available). In mis-stressed words only, there was also a greater negativity for trend-violating items between 120-180 ms, in the time window of the N1 auditory potential. Shown in the graph are average electrodes based on the Anterior, Anterior Central, and Posterior ROIs. Onset times for each syllable are represented in the legend as boxplots showing the distribution of onset times.
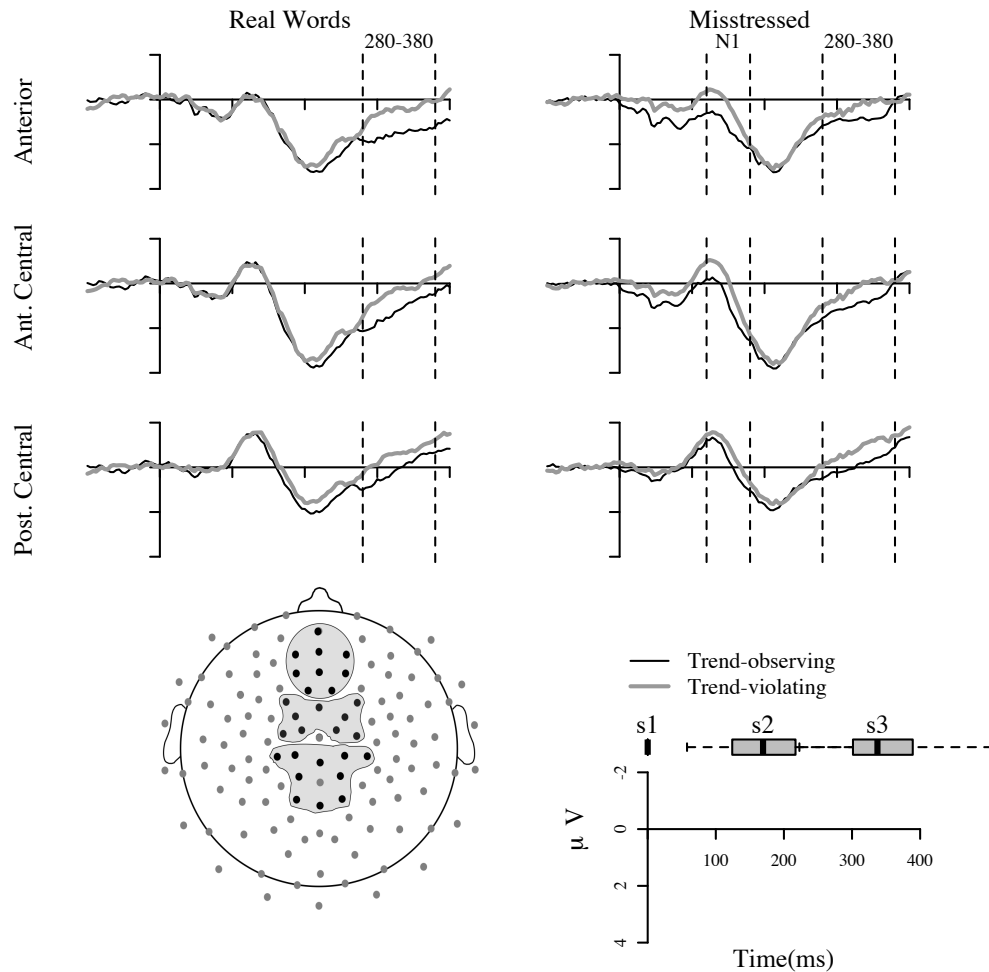
Figure 2.6. Early differences between trend-observing and trend-violating patterns in words (left) and mis-stressed words (right). No such effect was found for non-words. In both words and mis-stressed words, there was a greater negativity to trend-violating than trend-observing items between 280-380 ms post-stimulus onset (before information about the final vowel is available). In mis-stressed words only, there was also a greater negativity for trend-violating items between 120-180 ms, in the time window of the N1 auditory potential. This negativity was greater for [ə]-final mis-stressed words than for [i]-final mis-stressed words. Onset times for each syllable are represented in the legend as boxplots showing the distribution of onset times.
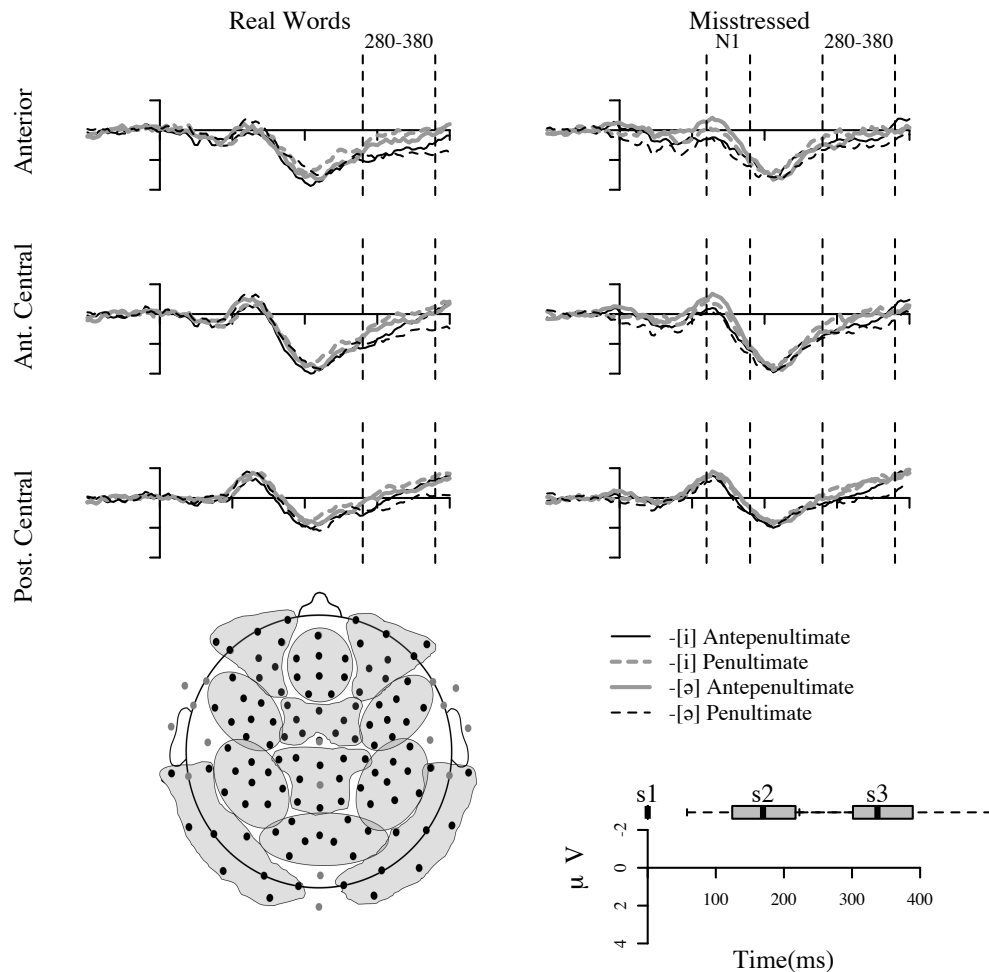
2.4.2.2 N400

Figure 2.7 shows ERPs to all the items in the experiment, split up by word type (real word, mis-stressed word, nonword). The N400 (400-1000ms) to real words is much smaller than that of the other two categories. In an omnibus ANOVA, there was a significant main effect of word type ($F(2,40)=31.2$, $p < .05$). Over just medial electrodes, the N400 was significantly smaller for mis-stressed words than for nonwords: an ANOVA with real words excluded yielded a significant main effect of word type (Mis-stressed, Nonword) ($F(1,20)=5.4$, $p < 0.05$).

Figure 2.8 shows only the schwa-final, penultimately stressed items. In these words, there are three distinct sizes of N400 across the whole head: real words (e.g. banána) have the smallest N400, nonwords (e.g. famáka) have the largest, and mis-stressed words (e.g. afríca) elicit a medium-sized N400. An omnibus ANOVA across the whole head showed a main effect of word type ($F(2,40)=17.0$, $p < 0.05$). ANOVAs with just the real and mis-stressed words ($F(1,20)$, $p < 0.05$), and with just the mis-stressed words and nonwords ($F(1,20)$, $p < 0.05$) revealed significant effects of word type. Thus, the N400 elicited by mis-stressed words is significantly smaller than the N400 elicited by nonwords, but significantly bigger than the N400 elicited by real words.

Other combinations of stress and final vowel do not have this property, however. Figure 2.9 shows the average sizes of the N400 for all items. In all cases except the [ə]-final penultimate, mis-stressed words pattern exactly the same as nonwords (and differently from real words).

Figure 2.9 also illustrates that the N400 is generally smaller for all [ə]-final penultimate items than for other items. This effect is significant for words ($t=3.62$, $p < 0.05$), and mis-stressed words ($t=4.99$, $p < 0.05$), but not for nonwords ($t=0.772$, $p = 0.44$).

Figure 2.7. N400s (400-1000ms) for words, mis-stressed words, and nonwords. Mis-stressed words and nonwords have significantly larger negativities than real words. Additionally, for medial electrodes only, mis-stressed words are less negative than nonwords. Onset times for each syllable are represented in the legend as boxplots showing the distribution of onset times.
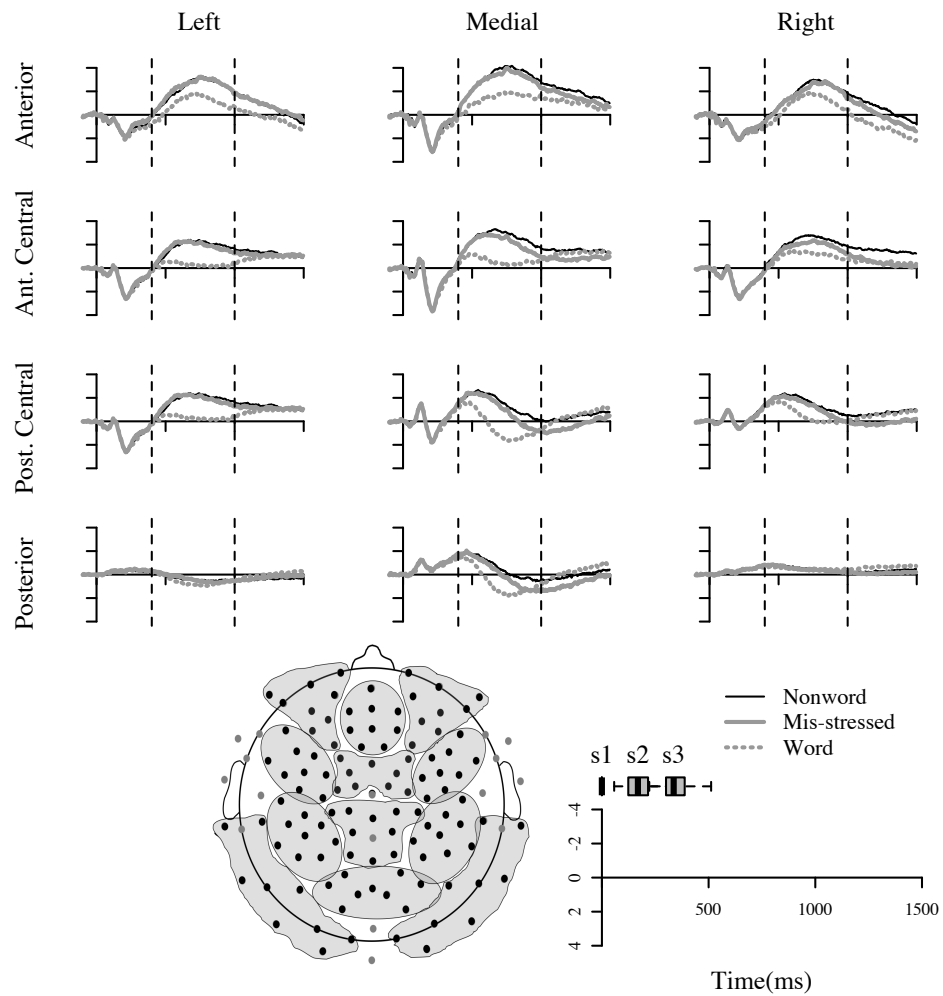
Figure 2.8. N400s (400-1000ms) with just the words, mis-stressed words, and non-words which are [ə]-final, penultimately stressed. Mis-stressed words and nonwords have significantly larger negativities than real words. Additionally, for medial electrodes only, mis-stressed words are less negative than nonwords. Onset times for each syllable are represented in the legend as boxplots showing the distribution of onset times.
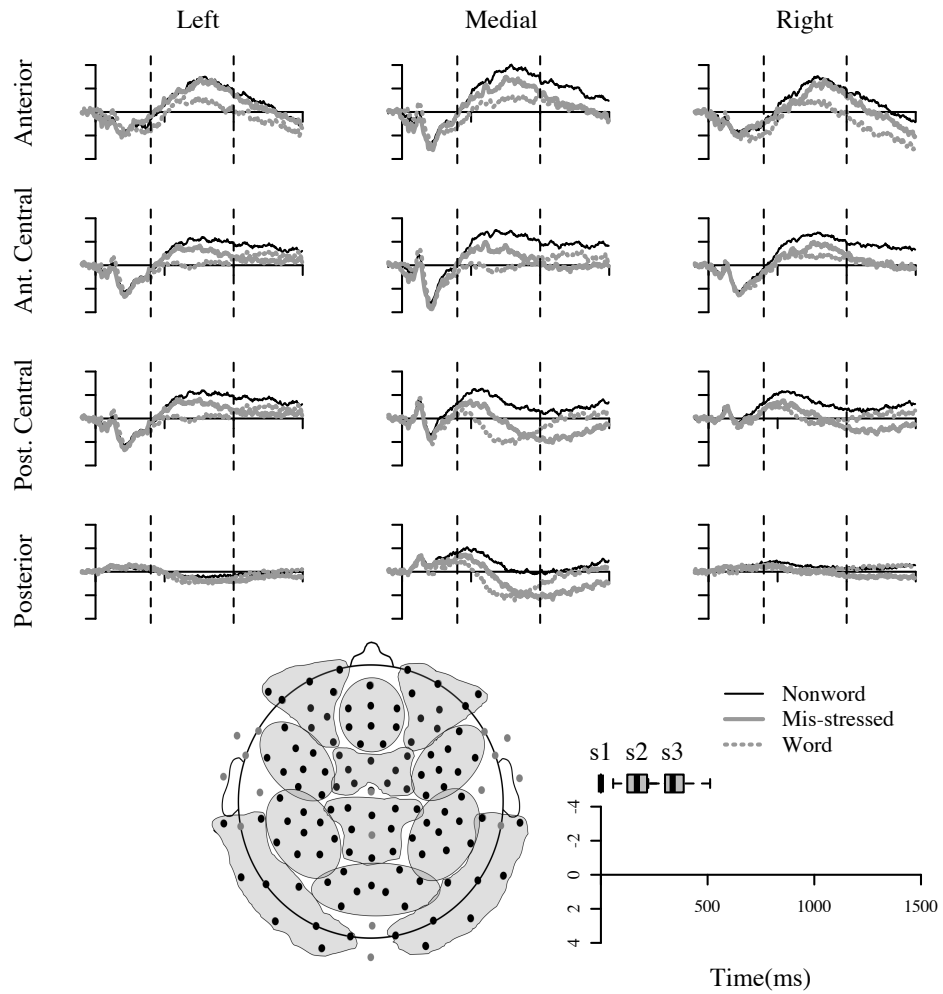
Figure 2.9. Mean N400 amplitude (400-1000ms) for each condition. Error bars represent standard errors based on one value for each average (ROI) electrode for each subject.
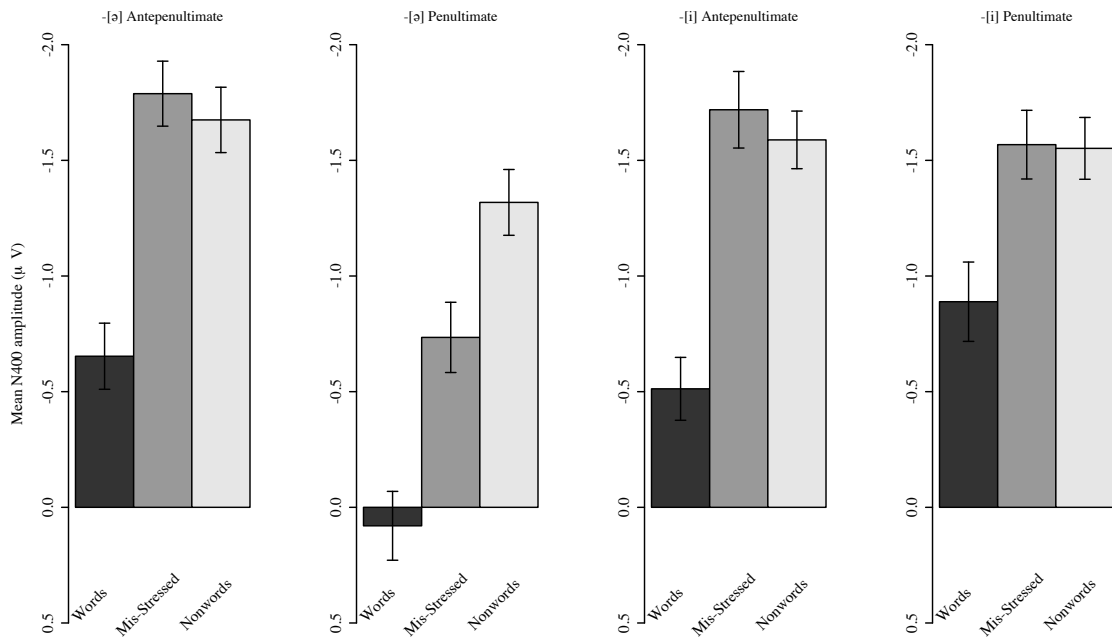
2.4.2.3 Late Positive Component (LPC)

Figure 2.10 shows ERPs to the four different types of nonwords, averaged over just the first block of the experiment. In this block only a single instance of each experimental item was presented. A contrast of i-final penultimate nonwords compared to the other three types was significant (t=3.5, $p < 0.05$).

By the very next block, this effect is gone, however. Figure 2.11 shows the size of the late positivity for each prosodic type by block. An ANOVA on data from blocks 1 and 2 together, with block as a factor, yielded a significant interaction of Block and Prosodic Form (Prosodic Form x Anteriority x Block: $F(9,180)=2.21$, $p < 0.05$).

Because of the small number of trials in the first block of the experiment, a follow-up was run in an attempt to replicate the late positivity in the first block of the experiment. The follow up was identical in design to the experiment reported here except that each block of the experiment contained new nonwords, so that no nonwords were repeated. In this follow-up, the early negativity and the N400 effects discussed above were replicated, but the late positivity was not. No late positivity was found in any block of the follow-up. Because of this failure to replicate, I will treat this late positivity as spurious and will not discuss it further.

2.5   Experiment 2 Methods

Experiment 1 found a late positivity for i-final penultimate nonwords relative to all other nonword types, but this late positivity persisted only throughout the first block of the experiment. i-final penultimately stressed nonwords were also rated as less wordlike than all other nonword types, and this rating difference persisted throughout the entire experiment. Experiment 2 is designed to test two possible explanations for this pattern of results. One possibility is that participants are un-learning the strong i-final trend because they are being exposed to an equal distribution of both word and nonword types within the context of the experiment. In that case, participants

Figure 2.10. ERPs to nonwords in Block 1 of the experiment. In this block only, there is a late positivity (1000-1500ms) to [i]-final Penultimate nonwords relative to all other nonwords. There is no difference in that time window between [ə]-final Antepenultimate and Penultimate nonwords. Onset times for each syllable are represented in the legend as boxplots showing the distribution of onset times.
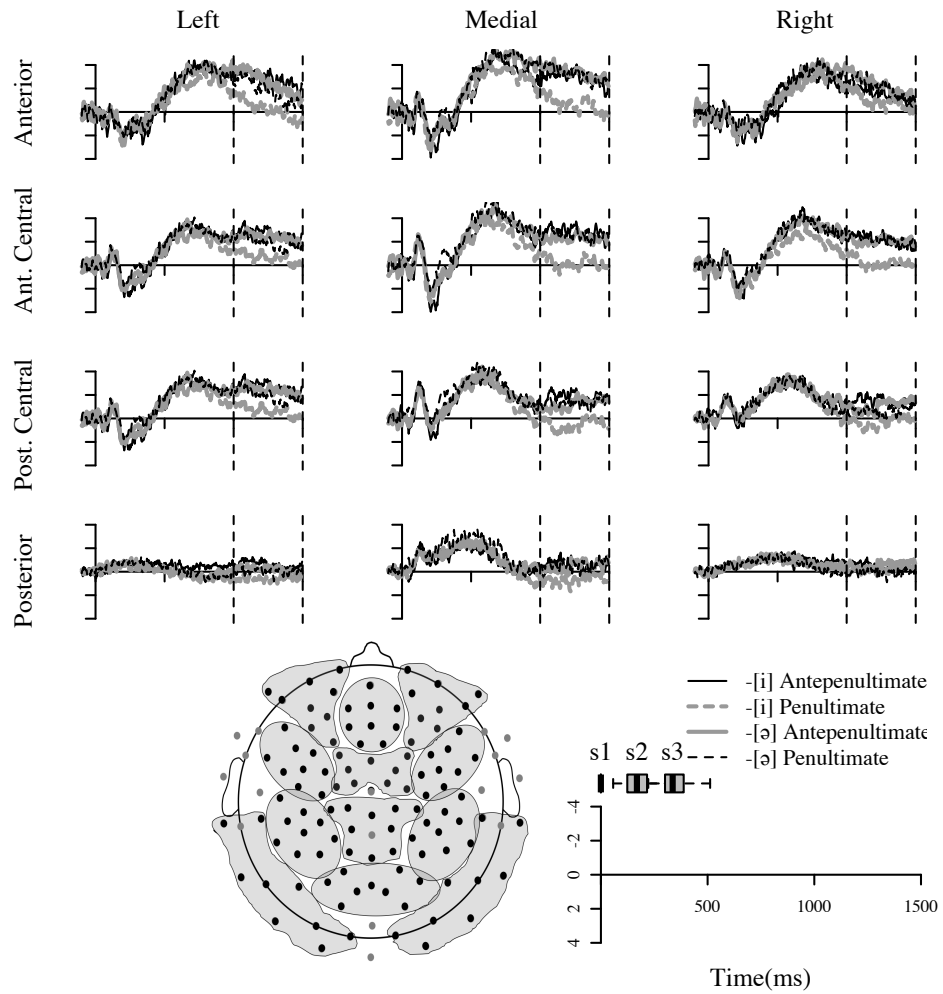
Figure 2.11. Mean late (1000-1500ms) amplitude in the first two and last blocks of the experiment. Error bars represent standard errors based on one value for each average (ROI) electrode for each subject.
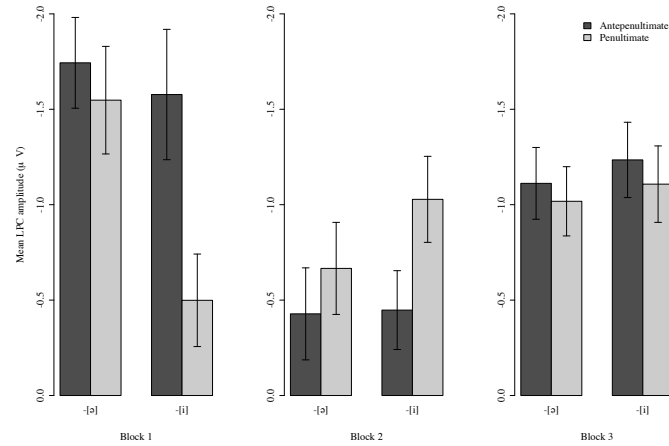
would have to be implicitly or explicitly remembering their ratings for each nonword throughout the experiment, since the rating difference between observers and violators of the trend does not change. A second possibility is that the reduction in size of the LPC is a priming effect for each particular nonword, analogous to the reduction in size of P600 effects found for repeated ungrammatical structures.

Experiment 2 is designed to tease apart these two possibilities by using new nonwords in each block of the experiment rather than repeating the same nonwords throughout. If the reduction in size of the LPC is due to un-learning of the pattern, the same results should obtain in experiment 2 as in experiment 1. However, if the reduction in size is due to each specific nonword being primed, then introducing new nonwords in each block should result in an LPC which persists throughout the experiment.

2.5.1 Participants

Participants were 25 right-handed native speakers of English, between the ages of 18 and 30. All had normal hearing, normal or corrected to normal vision, no

known neurological conditions, and all were undergraduate or graduate students at the University of Massachusetts Amherst.

2.5.2 Materials

As in experiment 1, there were 40 real words, and their mis-stressed counterparts. Instead of repeating the same 80 nonwords 8 times, 640 separate nonwords were used. The 80 nonwords from experiment 1 were included, and an additional 560 nonwords were constructed via a script.

Nonwords were counterbalanced for final vowel, so that each item appeared in all four stress x vowel conditions. Each participant saw each nonword in two stress conditions with the same vowel, but across participants each item appeared with both vowels (e.g. one participant heard [bǽməki] and [bəmǽki], and another heard [bǽməkə] and [bəmǽkə]).

Lexical neighborhood density was calculated for all nonwords using the Generalized Neighborhood Model (Bailey and Hahn, 2001), and mean phoneme bigram probability, phoneme trigram probability, and phoneme positional probability were calculated for each nonword using the Irvine Phonotactic Online Dictionary (Vaden and Hickok). Nonwords from the four conditions (stress x final vowel) were roughly matched on all four measures.

Items were recorded with the same procedure as in Experiment 1. Table 2.8 shows that the stressed syllables of all word types were comparable in duration and in F0 excursion. For both antepenultimate and penultimate stressed items, however, the mean intensity was significantly less by an average of 5-6 dB for nonwords than for real or mis-stressed words.

2.5.3 Procedure and EEG recording

The presentation and EEG recording procedure were almost identical to that of Experiment 1, including location, setup of the equipment and EEG recording pa-

Table 2.7. Mean (standard deviation) neighborhood density for each value for the 640 (80 old, 60 new) nonwords used in Experiment 2. The neighborhood density of all items is very low (a neighborhood density of 1 corresponds to a single neighbor), as are their phoneme transitional probabilities.

	Neighborhood Density			**Bigram transitional prob.**	
	initial	*penultimate*		*initial*	*penultimate*
ə	5.0e-4 (9e-4)	3.1e-4 (6e-4)	ə	3.6e-3 (2e-3)	3.7e-3 (2e-3)
i	5.6e-4 (10e-4)	1.6e-4(3e-4)	i	3.1e-3 (2e-3)	3.1e-3 (2e-3)

	Trigram transitional prob.			**Phoneme positional prob.**	
	initial	*penultimate*		*initial*	*penultimate*
ə	1.6e-4 (1e-4)	1.1e-4 (1e-4)	ə	0.068 (0.008)	0.054 (0.007)
i	1.3e-4 (1e-4)	0.8e-4(1e-4)	i	0.055 (0.008)	0.042 (0.007)

Table 2.8. Acoustic measurements of stressed syllables in all experimental items. ANOVA's were run using type III sum of squares, since there are 640 observations in each nonword cell, but only 40 in each word and mis-stressed cell. For some syllables, it was not possible to get an accurate F0 measurement, so these are excluded from the analysis (92 antepenultimately stressed itesms, 78 penultimately stressed items).

Realized Stress	Parameter	Word	Mis-stressed	Nonword	Statistics
Antepenultimate	Duration (ms)	91 (19)	91 (29)	85 (27)	$F(2,685)=1.02$, $p=0.36$
	Intensity (dB)	75 (3)	74 (2)	69 (5)	**$F(2,685)=24.24$, $p \ll 0.05$**
	F0 excursion (Hz)	6.1 (4.2)	7.0 (5.3)	5.0 (10.3)	$F(2,593)=0.47$, $p=0.63$
Penultimate	Duration (ms)	107 (24)	99 (25)	105 (32)	$F(2,685)=0.35$, $p=0.71$
	Intensity (dB)	74 (3)	73 (3)	67 (6)	**$F(2,685)=35.59$, $p \ll 0.05$**
	F0 excursion (Hz)	17.5 (28.2)	8.2 (4.1)	12.4 (37.9)	$F(2,607)=0.32$, $p=0.73$

rameters, trial structure, and instructions given to participants. As in Experiment 1, stimuli appeared in 5 blocks of either 160 or 320 trials each. The first two blocks consisted of 160 trials while the remaining three blocks had 320 trials. Each real word and mis-stressed word appeared once in blocks 1 and 2, and twice in the remaining blocks. Nonwords were randomly distributed throughout the blocks so that participants did not see the same nonword twice. Due to a coding error, subjects 1-6 did see some of the nonwords exactly twice throughout the course of the experiment, but

remaining subjects did not. The order of presentation of items within each block was randomized. The entire experiment took about 2.5 hours.

As before, trials containing eye blinks, eye movements, or other artifacts were excluded automatically using ERPLAB. Six subjects were excluded completely from analysis due to too high an incidence of artifacts. Three subjects had too many blinks, two subjects exhibited too much low–frequency 'drift' artifacts, and one exhibited too much noise from muscle tension. Subject-specific voltage thresholds were chosen for eye blink/movement channels (between 300 and 700 μV) as well as for the whole head (between 400 and 1000 μV). Any trial where the threshold was exceeded was marked as artifactual. For all included participants, between 4% and 45% of trials contained artifacts (all participants but one were below 30%).

As before, epochs were timelocked to the beginning of the word, the beginning of the second syllable, and the beginning of the third syllable, but as all effects of interest appeared timelocked to the first syllable data timelocked to the first and second syllables will not be discussed further.

Table 2.9. Onset times for each syllable of the stimuli, in milliseconds from stimulus onset

	s2 onset	s3 onset	end
mean	164	330	442
min	34	171	266
max	364	500	650

2.5.4 Analysis

ERPs were computed for each participant, condition, and electrode, and were re-referenced to the average of the two mastoid electrodes, and baseline-corrected using the 100ms pre-stimulus-onset as a baseline. The 128 electrodes were divided into the same 12 regions, each containing 9 electrodes, as in Experiment 1. These are depicted

in Figure 2.3. Average ERPs were calculated for each region, averaging together all 9 electrodes in each region.

Mean voltage was calculated at four time windows of interest, which were chosen based on the results of Experiment 1, and on visual inspection of the averaged waveforms. Those time windows were 280-380ms, corresponding to the early difference between violators and observers found in Experiment 1; 400-1000ms, corresponding to the N400, and 1000-1500ms, corresponding to the Late Positive Component in Experiment 1.

Repeated-measures ANOVAs were calculated using as factors Prosodic Form ([i]-final Antepenultimate, [i]-final Penultimate, [ə]-final Antepenultimate, [ə]-final Penultimate), Word Status (Word, Misstressed, Nonword), Anteriority (Anterior, Anterior Central, Posterior Central, Posterior), and Laterality (Left, Medial, Right).

2.6 Experiment 2 Results

2.6.1 Ratings

Average ratings for each type of word presented in Experiment 2 are shown in Figure 2.12. The pattern of ratings is very similar to that in Experiment 1. As in Experiment 1, a within-subjects ANOVA was conducted with mean ratings from individual subjects as the dependent variable. Factors were WordStatus (Word, Misstressed, Nonword), FinalV (-[ə], -[i]), and Stress (antepenultimate, penultimate). Once again, "jalopy", and "taffeta" were excluded from analysis since they received low average ratings (2.8 and 3.1 respectively; all other words received average ratings of 3.4 or higher, and 27 were over 3.9)

In the ANOVA, there was a main effect of word type $F(2, 48) = 170.7$, $p \ll 0.05$). Ratings for nonwords were much lower than for actual words, and mis-stressed words were in between. Antepenultimately stressed items received higher ratings than penultimately stressed items in both mis-stressed words and nonwords, but not

Figure 2.12. Mean ratings. Error bars represent standard errors based on the distribution of individual subject means.
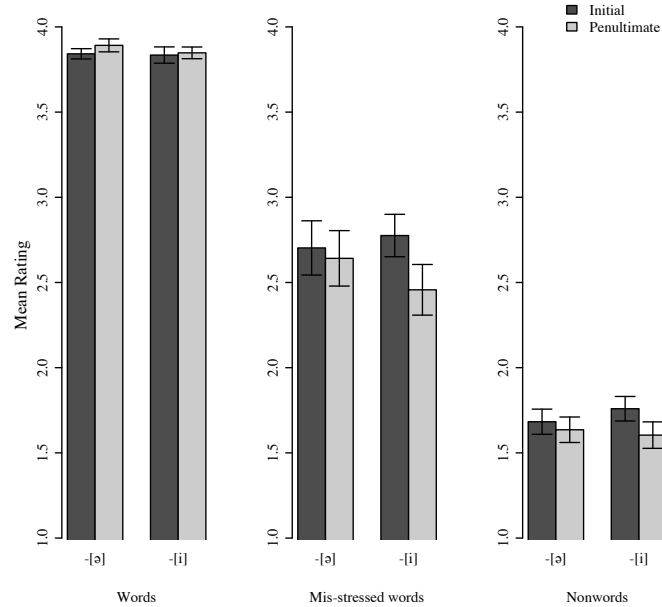
in real words (main effect of Stress: $F(1, 24) = 26.6$, $p \ll .05$ WordStatus x Stress: $F(2, 48) = 13.2$, $p \ll .05$).

The difference between antepenultimate and penultimate stress was bigger for items ending in -[i] than for items ending in -[ə] (FinalV x Stress across all word shapes: $F(2, 24) = 6.3$, $p = 0.02$) but the effect was biggest for nonwords (WordStatus x FinalV x Stress: $F(2, 48) = 10.35$, $p \ll 0.05$). The mean rating for antepenultimately stressed [i]-final words was 0.16 points higher than the mean rating for their penultimately stressed counterparts. However, the difference in means between antepenultimately and penultimately stressed [ə]-final nonwords was only 0.04 points.

2.6.2 ERP data

Figures 2.13-2.15 show the grand average ERP's timelocked to the onset of the word. Each electrode pictured is the average of 9 electrodes, calculated over the regions of interest pictured in Figure 2.3 above. All experimental items elicited the

typical sequence of P1-N1-P2, with the earliest positivity peaking at 60ms, the earliest negativity at 100ms, and the second positivity at 200ms.

2.6.2.1   Early effects

Unlike in Experiment 1, only real words and not mis-stressed words exhibited an early negativity for pattern violators relative to observers. Also unlike Experiment 1, the negativity was only observed for violators of the i-final trend, and was not observed for violators of the ə-final trend. Figure 2.13 compares penultimately stressed words which end in -i (pattern violators) to penultimately stressed words ending in schwa (pattern observers). An ANOVA conducted on just the two penultimately stressed word shapes revealed a significant effect of Prosodic Form over the whole head (two levels: [i]-final Penultimate, [ə]-final Penultimate; $F(1,20)=5.8$, $p = 0.03$). An ANOVA on just the antepenultimately stressed word shapes ([i]-final Antepenultimate, [ə]-final Antepenultimate) revealed no significant effect ($F(1,20)=0.42$). These early effects occur well before any acoustic information about the final vowel is present (starting at about 300ms).

2.6.2.2   N400

Figure 2.14 shows ERPs to all the items in the experiment, split up by word type (real word, mis-stressed word, nonword). The N400 (400-800ms) to real words is much smaller than that of the other two categories. In an omnibus ANOVA, there was a significant main effect of word type ($F(2,40)=25.6$, $p < 7x10^{-8}$). Over just medial electrodes, the N400 was significantly smaller for mis-stressed words than for nonwords: an ANOVA with real words excluded yielded a significant main effect of word type (Mis-stressed, Nonword) ($F(1,20)=5.3$, $p = 0.03$).

In Experiment 1, the N400 on ə-final, penultimately stressed words and mis-stressed words was smaller than in other word shapes. In Experiment 2, only the mis-stressed words showed an amelioration of the N400 for ə-final penultimate items.

Figure 2.13. An early negativity between i-final and ə-final penultimately stressed words. No such effect was found for nonwords or misstressed words.
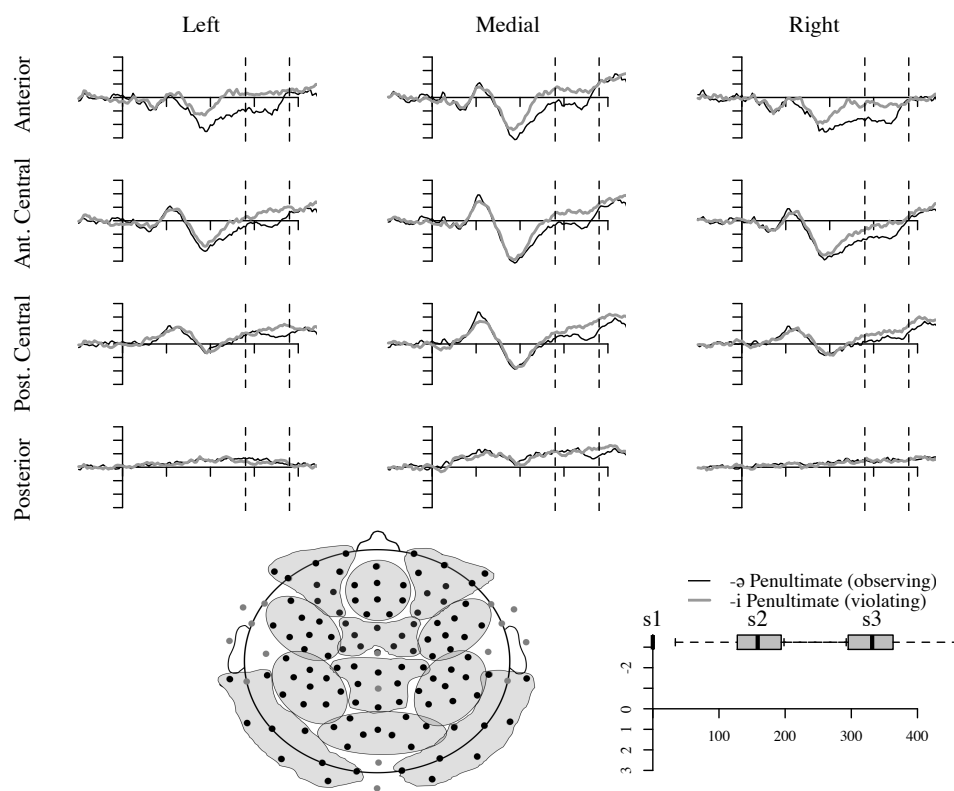
Figure 2.14. N400s (400-1000ms) for words, mis-stressed words, and nonwords. Mis-stressed words and nonwords have significantly larger negativities than real words. Additionally, for medial electrodes only, mis-stressed words are less negative than nonwords. Onset times for each syllable are represented in the legend as boxplots showing the distribution of onset times.
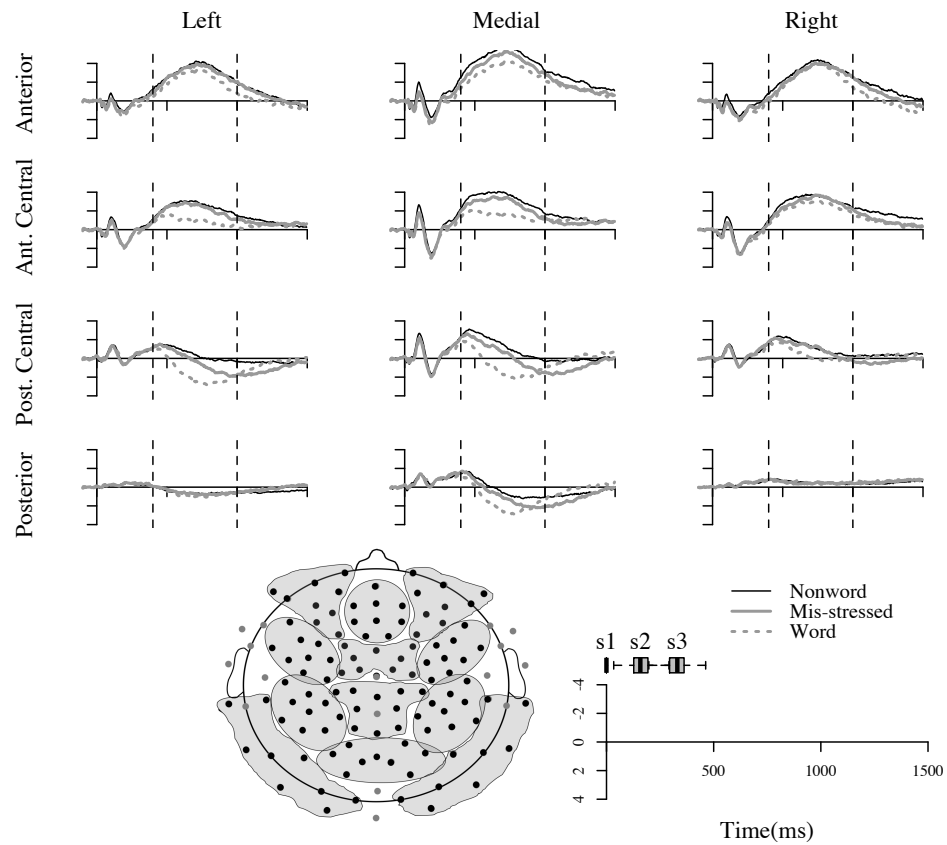
Figure 2.15. Mean N400 amplitude (400-1000ms) for each condition. Error bars represent standard errors based on one value for each average (ROI) electrode for each subject.
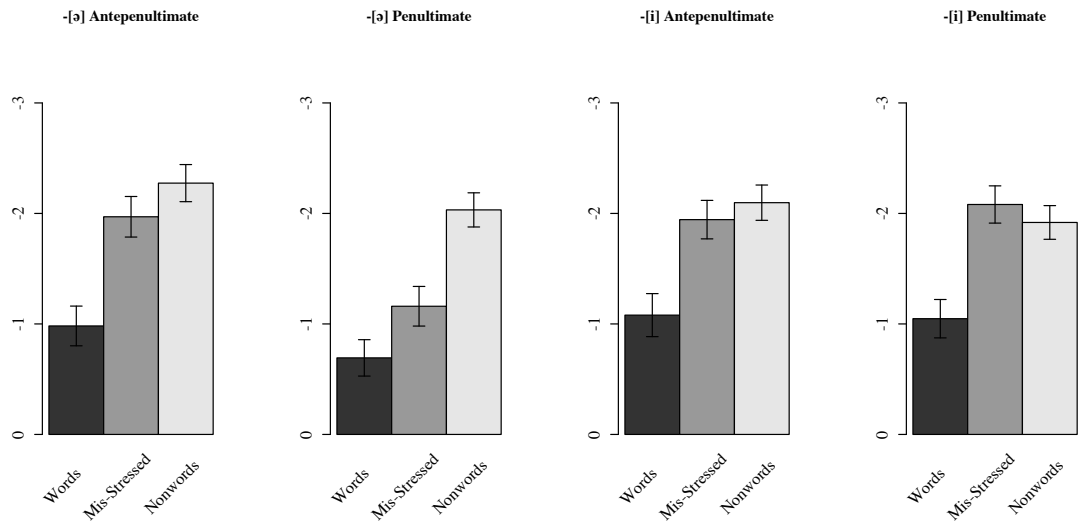
Figure 2.15 shows the mean N400 size for each condition, split up by word shapes. Within mis-stressed words, the difference between ə-final penultimate items and other items is significant (t=4.11, $p < 0.05$), but it is not significant for real words (t=1.67, p=0.1) or for nonwords (t=0.35, $p = 0.72$).

2.7 Discussion

2.7.1 Ratings

This experiment tested the psychological reality and cognitive status of two probabilistic trends in the English lexicon: the tendency for [i]-final words to take antepenultimate stress, and the tendency for [ə]-final words to take penultimate stress. Only the strong i-final trend was 'productive' in the ratings. Nonwords that violated it ([bəmǽki]) were rated as less likely to be an English word than nonword observers ([bǽməki]). The weak ə-final trend was not productive: no rating difference was observed between observers ([fəmákə]) and violators ([fáməkə]). These two trends

52

differed in terms of their degree of observance in the lexicon. The strong trend is observed over a large scope of forms - about 2000 words, including morphologically complex forms and words containing heavy syllables, and of those forms about 90% observe the trend. The weak trend is observed over about 200 words, which are morphologically simple words with no heavy syllables. Of those 200 words, about 70% observe the trend.

All three differences between the two trends (scope, percent observance, morphology and weight as conditioning factors) could contribute to differences in learnability between the two trends. The higher the scope of a trend, the more experience the learner will get with that trend, and the higher the percentage of observers of the trend, the more reliable the trend will be. The extra conditioning factors of syllable weight and morphological structure make the statement of the weak trend more complex than the statement of the strong trend ('i-final words are antepenultimately stressed' vs. 'unsuffixed ə-final words with a light penult are penultimately stressed').

The addition of the 'unsuffixed' condition is particularly problematic. Features like 'unsuffixed' which can only be defined negatively are not typically used in models of phonological grammar. Assuming that such a feature can be used to describe a pattern though, the language learner would not be able to reference the set of unsuffixed things until she had acquired the relevant morphology. In this case, the relevant morphology is derivational affixes and is acquired fairly late. Tyler and Nagy (1989) find that fourth graders have not completely acquired the syntactic properties of various derivational affixes in English, and Jarmulowicz (2002) examined seven and nine year olds' knowledge of the stress-shifting nature of a selection of English derivational suffixes, finding that neither group were as good as adults at correctly judging the stress properties of a particular affix. On the other hand, children can and do begin learning generalizations about the stress pattern of their language as early as 9 months of age (Jusczyk et al., 1993).

How is the strong trend represented? Based on the behavioral data collected in this experiment, there is no way to tell whether the trend is represented abstractly as a rule or constraint in a phonological grammar, or whether speakers' judgments are the results of an analogical process or a similar mechanism which consults the lexicon directly in judging new words. Chapter 3 provides evidence that it is in fact abstractly represented by explicitly testing for effects of specific lexical items on speakers' behavior. However, for the moment the most that can be said is that the trend is probabilistic in nature. In participants' ratings, the numerical difference between strong-trend-observing nonwords and strong-trend-violating nonwords was quite small, only 0.14 points. There was no floor effect - both i-final antepenultimate and i-final penultimate nonwords were rated as 2's more than as 1's (their average ratings were 1.78 and 1.64 respectively). If this trend is abstractly represented, that representation appears not to demand 100% satisfaction on nonwords.

2.7.2 Early negativity

The early negativity to violators of the i-final trend (and both trends in Experiment 1) shows that this trend is active early during the process of lexical access. This effect does not obtain in nonwords, and it is early enough that in many cases subjects have not yet heard the word's final vowel. For both these reasons, this component cannot be directly related to the perception of the acoustic string that violates the i-final trend. Both the stress pattern and the content of the final vowel are necessary information to tell whether a string is trend-observing or trend-violating. By 280-380ms post-stimulus-onset, subjects have processed acoustic information about the word's stress pattern but not about it's final vowel. This distinction between trend-observers and trend-violators may be a consequence of subjects accessing information in the lexical entry of a word, either phonological information about the final vowel (before they have actually heard it), or direct information about whether the

word has a high- or low- probability shape. In nonwords, there is no stored lexical representation, so subjects cannot perceive a violation of a trend before they have heard (and processed) the entire string.

This negativity may be related to the 'N325' observed by Böcker et al. (1999) in response to Dutch words with low-probability stress patterns (their experiment did not contain nonwords). Böcker et al. argue that their N325 is related to the process of extracting the metrical stress pattern from the acoustic signal, and point out that it is in the expected time window for lexical or slightly pre-lexical processes. The early negativity observed in the present experiment cannot have exactly this explanation: because it occurs only in real words, and because it distinguishes between violators and observers at a time when not all acoustic information necessary to make that distinction has unfolded, must relate both to the word's metrical structure and to the process of accessing a particular lexical item.

One possibility is that both the early negativity and Böcker et al.'s N325 are instances of a component which is a manifestation of one stage of lexical access, namely the identification of a word's stress pattern. A greater negativity at this stage obtains if the word's stress pattern is atypical in some way. I will consider two possible mechanisms by which the greater negativity could arise. The first possibility is that the lexical encoding of stress may be different for items which violate grammatical trends than for items which do not. In the phonological literature, many such mechanisms have been proposed - for English stress specifically Hayes (1980); Selkirk (1984); Chomsky and Halle (1968); Burzio (1994); Pater (2005); Becker (2009) have all proposed some kind of extra diacritic or feature on words with exceptional stress patterns. Building on this body of work, Moore-Cantwell and Pater (2015) propose that each word's lexical representation has some weight and competes with grammatical constraints. A stress pattern which violates a lexical trend would have a higher weight (be more strongly represented) so that it can overcome the pressures

of the grammar. A stress pattern which observes a lexical trend would have a lower weight, and be less strongly encoded. If this analysis is correct, then accessing the stress pattern of an exceptional word would proceed differently than accessing the stress pattern of a trend-observing word. However, it is not clear that accessing a more strongly encoded stress pattern should result in a greater negativity - rather one might imagine that accessing or recognizing a more strongly encoded stress pattern would be *less* effortful than accessing or recognizing a less strongly encoded stress pattern.

A second possible mechanism that may give rise to the observed pattern of early negativity is competition between the grammar and a word's lexically encoded stress pattern. For words whose stress pattern agrees with the probabilistic trend in the grammar, there is no competition, as the grammar's predictions and the lexical encoding of stress align. For trend-violating words however, the lexically encoded stress pattern violates the grammar's expectations. Competition between the grammar's expectations and the actual stored stress pattern of a word makes the process of correctly recognizing that stress pattern more effortful, and therefore gives rise to a greater negativity at the stage when the word's stress pattern is recognized.

2.7.3   N400

The N400 was much bigger for nonwords than for words - an effect replicating the large N400 for nonwords found by Rugg and Nagy (1987) and many others. In real words and mis-stressed words only, and in Experiment 2 only in mis-stressed words, the N400 was also modulated by the word's prosodic structure - Schwa-final penultimately stressed words elicit a smaller N400 than all other prosodic shapes. The size of the N400 in nonwords was not modulated by the words' prosodic structure, suggesting that the effect is an effect of the structure of the lexicon only, and does not reflect an abstract grammatical generalization.

Schwa-final penultimately stressed words are observers of the weak trend, suggesting that the weak trend is active in the lexical access process. One possibility is that lexical items which observe the weak trend are easier to access than violators are, potentially because more other lexical items share their stress and final vowel combination. Just as more frequent lexical items elicit a smaller N400 than less frequent lexical items (Rugg, 1990 an others), more frequent *types* of lexical items are eliciting a smaller N400 than less frequent types.

However, the strong trend does not modulate the N400 in the same way - both observers and violators of the strong trend elicit a large negativity, similar in size to that elicited by violators of the weak trend. If [ə]-final Penultimate words are easy to access because many lexical items share their shape, then [i]-final Antepenultimate words should also be easier to access since many lexical items also share their shape.

A different possibility is that the N400 is not being modulated by the weak trend specifically, but by larger-scale statistical generalizations in the English lexicon. Over all word shapes, penultimate stress is the most common stress placement in multisyllabic English words, accounting for 55% of all words at least two syllables long. Penultimately stressed words may be easier to access (and thus elicit a smaller N400) than antepenultimately stressed words because penultimately stressed words are more common overall in English. However, penultimately stressed i-final words are a special case, and are severely underrepresented, making words of that shape more difficult to access. Thus, antepenultimately stressed words with any final vowel, and penultimately stressed i-final words are all more difficult to access than schwa-final penultimately stressed words.

The distribution of stress patterns in English words is shown in 2.16. These counts include all multisyllabic words, in which different syllable counts are not equally represented. In particular, two-syllable words are the most common, and in that category stress can only be penultimate or final; the other stress categories are impossible.

Figure 2.16. Counts of each main stress category, where categories are aligned to the right edge of the word. This graph includes all multisyllabic words found in the corpus, including multimorphemic words.

Main stress counts (all words)

2 syllables: 45%
3 syllables: 31%
4 syllables: 18%
>4 syllables: 7%
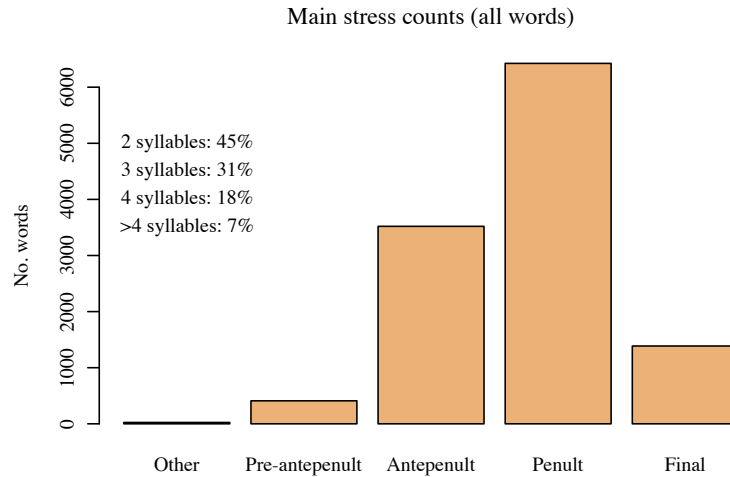
This claim that penultimate stress is the 'expected' stress pattern overall in English words runs contrary to previous literature on lexical processing in English, which observes that the majority of English words begin with a stressed syllable (primary or secondary stressed) rather than an unstressed one (Cutler and Carter, 1987 et. seq). The corpus used here shows a similar effect. Figure 2.17 shows the distribution of main stress locations counting from the left edge of the word (still only multisyllabic words), and indeed initial stress is the most common. This figure does not include the approximately 6,000 monosyllabic words in the lexicon, all of which are also stressed on the first (and only) syllable.

A great deal of literature has shown that English speakers use stressed syllables as a cue to word boundaries, assuming that a stressed syllable is the beginning of a word, both in nonwords and in real words (Cutler and Norris, 1988; Cutler and Butterfield, 1992). English speaking infants (as early as 7 months) use stress as a cue, even relying on it over statistical segmental cues (Thiessen and Saffran, 2003; Johnson

Figure 2.17. Counts of each main stress category, where categories are aligned to the right edge of the word. This graph includes all multisyllabic words found in the corpus, including multimorphemic words.

Main stress from left edge (all words)

2 syllables: 45%
3 syllables: 31%
4 syllables: 18%
>4 syllables: 7%
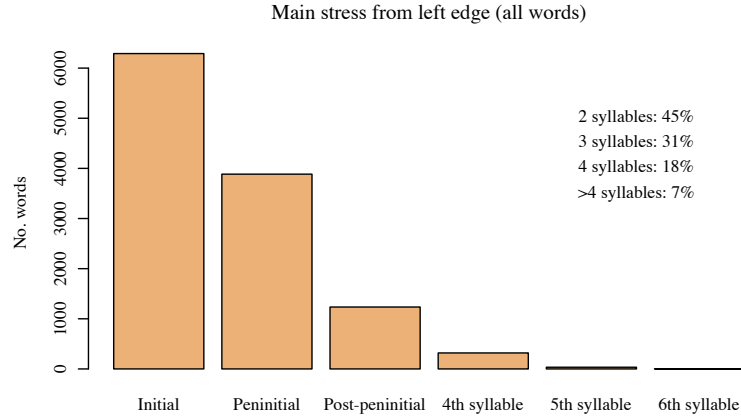
and Jusczyk, 2001). Additionally, adults give more initial than final stress to two-syllable nonwords (Kelly and Bock, 1988; Guion et al., 2003; Domahs et al., 2014), and children prefer initial stress to final stress on two-syllable nonwords (Jusczyk et al., 1993).

The majority of this research has been conducted using two-syllable words, for which initial stress and penultimate stress are equivalent. The behavioral evidence cannot distinguish between a preference for initial stress and a preference for penultimate stress. A few word-segmentation studies have used three-syllable words, showing that children (Houston et al., 2004) and adults (Sanders and Neville, 2000) prefer to posit word boundaries at stressed syllables, resulting in antepenultimately stressed words. However, judgment studies on three-syllable words (Olejarczuk, 2014; Domahs et al., 2014) do not find a clear preference for initial (antepenultimate) over penultimate stress. The generalization that drives speakers' word segmentation choices may be different from the generalization that drives speakers' judgments. Finding word edges may be done (at least in part) without any direct reference to the details of a word's lexical representation, but the process of accessing a segmented word's lexical

entry, and also the process of judging that word's phonological form, necessitates reference to all the details of the stores phonological form.

In sum, although listeners expect word boundaries at stressed syllables, they do not expect words to be initially stressed. Rather, they 'expect' words to have a right-aligned trochee. Words with antepenultimate stress are treated as a low-frequency type of word, and therefore elicit a larger N400 than words with penultimate stress - the more common right-aligned trochee. Penultimately stressed words with a final i are different than other penultimately stressed words. Since speakers have learned a generalization that this type of word is dispreferred, or rare, it elicits a larger negativity than other penultimately stressed words.

2.8   Conclusion

The strong trend for i-final words was observed in participants' ratings of nonwords, and it affected early stages of the lexical access process in nonwords. The ə-final trend also affected early stages of lexical access, but only in the first experiment - the effect did not replicate. I argue that this early negativity is the same component as the 'N325' found by Böcker et al. (1999), and that this component originates from the stage of lexical access at which the word's stress pattern is recognized. This recognition draws upon the acoustic signal as well as stored information about a particular lexical item. The size of the negativity is modulated by the typicality of the word's stress pattern, which could be encoded either in the strength of representation of the pattern (less typical patterns being represented more strongly) or, if words are grouped according to their stress pattern, in the size of the group of words to which that word belongs.

I examine two probabilistic trends within the English stress system, the strong i-final trend, and also a weak ə-final trend for penultimate stress. Only the strong i-final trend affected participants' ratings of nonwords, and affected participants' per-

ception of real words in both experiments (effects of the weak trend were observed in Experiment 1 but were not replicated in Experiment 2). The weak ə-final trend is observed over only a small number of words in the lexicon, specifically monomorphemic words all of whose syllables are light. I argue that this trend does not affect speakers' behavior because it is either too complex (requires reference to a large number of features in its definition) or there is not enough evidence for it in the lexicon. On the other hand, I find strong evidence that the strong i-final trend, although it is still quite complex, is not only known by speakers and accessible for making wordlikeness judgments, but is active in the perception process.

CHAPTER 3

STATISTICAL PATTERNS IN PRODUCTION

3.1   Introduction

Word stress in English is lexically conditioned - words with nearly identical segmental content can have idiosyncratically different stress patterns: *cánnery* vs. *canáry*. However, a number of generalizations have been described by e.g. Chomsky and Halle (1968); Liberman and Prince (1977); Halle and Vergnaud (1987); Hayes (1980, 1982); Kager (1989); Burzio (1994); Alcántara (1998). Perhaps the most notable among these is the 'Latin Stress Rule' which governs stress in longer words of English.

(1)   **Latin Stress Rule for English:** If a word's penultimate syllable is heavy, then that word receives penultimate main stress. If the penultimate syllable is light, then the word receives antepenultimate main stress.

This rule, like most generalizations in the English stress system, has exceptions. Words like *gálaxy* violate the first clause, and words like *vanílla* violate the second clause. However, speakers of English do demonstrate probabilistic knowledge of this and other generalizations (Olejarczuk, 2014; Domahs et al., 2014). This paper investigates speakers' knowledge of an exceptionful generalization in the English stress system. Participants in a wug-test extend this generalization probabilistically to novel words. One possible explanation for this behavior is that participants analogize to existing words in order to choose a stress pattern for a nonword. In this case, the statistics of the lexicon would 'automatically' be transferred to participants' production of nonwords. The production experiment directly tests for this possibility, but

does not find evidence that participants use particular actual words to make their choice of stress on nonwords. I argue that the trend is abstractly represented as part of the phonological grammar.

While the first clause of the Latin Stress Rule has very few exceptions in the lexicon, the second clause has many exceptions. Pater (1994) argues that antepenultimate and penultimate stress compete when the penultimate syllable is light, and that neither is clearly the rule or clearly the exception. In this paper, a search of the CMU pronouncing dictionary reveals that the degree of preference for antepenultimate stress in words with light penults varies based on the word's final vowel. Light penult words ending in [i], [l̩], or [ɹ̩] are more likely take antepenultimate stress than light penult words ending in [n̩/m̩/ŋ̩] or [ə].

A web-based production experiment compares participants' choices of stress on novel words ending in [i] vs. novel words ending in [ə]. Additionally, each participant provides an 'analogical base' for each nonword by listening to that nonword and filling in a blank with a similar real word. The stress patterns of the given bases can then be compared to participants' choice of stress for each nonword. Participants treat i-final and ə-final words differently, preferring antepenultimate stress on i-final words, but exhibiting no preference on ə-final words. Although in aggregate the participant-provided analogical bases also follow the lexical trends, the stress of the bases does not predict the stress participants produced on particular nonwords. I argue that the preference for i-final words to take antepenultimate stress is productive because it is grammatically encoded. A model using Maximum Entropy Grammar (Goldwater and Johnson, 2003) is provided.

3.2 Models of productivity

If a phonological trend in the lexicon is generalized to new words, what is the underlying mechanism? One possibility is that speakers have learned an abstract

generalization which is a cognitive object independent from the lexicon. Such a generalization would be learned from the lexicon, but once learned would influence production and perception without making direct reference to the lexicon. Another possibility is that participants can extend trends in their lexicon to nonwords without referencing an explicit abstract generalization about the trend. For example, if during the production or perception of a nonword, the contents of the lexicon are accessed in a way that directly affects a person's behavior on that nonword, this influence of the lexicon could lead to an extension of trends in the lexicon to nonwords. This paper examines a probabilistic trend in the English stress system which speakers productively extend to new words, directly testing for the effects of a 'non-abstract' mechanism on speakers' productions.

Speakers' ability to generalize a phonological pattern to new words proceeds from a system of linguistic knowledge containing minimally a lexicon and a production and perception mechanism. An important question is whether or not this system also contains a grammatical component - a system of abstract rules or constraints. Models of phonological knowledge such as Chomsky and Halle (1968) and Prince and Smolensky (1993/2004) which rely on abstract representations of forms and generalizations across them have been successful in modeling both individual speakers' knowledge of their language, and the typology of phonological patterns across languages.

However, because the ability to generate a new form is used in day-to-day speech relatively rarely (compared to the ability to generate a new sentence, which is used almost constantly), it is tempting to posit a phonological system that is simplified compared to the syntactic system - a system which instead of a grammatical component uses independently necessary cognitive mechanisms to account for phonological productivity. For example, McClelland and Elman (1986) develop a model of the lexical access process which also predicts, essentially as a side effect, that generalizations in the lexicon should affect speakers' perception of novel words. Other strategies

include 'statistical learning' mechanisms for phonology (Seidenberg et al., 2002 and references therein), which simplify the linguistic system by assuming that general cognitive mechanisms for pattern learning in a broad array of contexts apply also to a language's lexicon. Spreading-activation models (Rumelhart and McClelland, 1986) and analogical models (Skousen, 1989; Daelemans et al., 1994; Nakisa et al., 2001) do posit some extra mechanism beyond the lexicon, but one without explicit representation of any generalizations.

It is not just categorical trends in the lexicon of a language but also probabilistic trends which can be generalized to new words. A growing body of work (Zuraw, 2000, 2010; Ernestus and Baayen, 2003; Hayes et al., 2009; Becker et al., 2011) demonstrates speakers' ability to 'probability match', or apply certain statistical generalizations in the lexicon to nonwords. Ernestus and Baayen find that some obstruents are more likely to be voiced word-internally than others (p/b is voiced in about 9% of words while s/z is voiced in about 33% of words). When given the chance to choose the voicing of a word-internal obstruent, participants mimic these probabilities in their productions, voicing p/b in 4% of responses, but s/z in 23% of responses.

Probability-matching behavior presents an interesting challenge to models of phonological knowledge which rely on abstract generalizations. Systems of rewrite rules (Chomsky and Halle, 1968) and interacting constraints (Prince and Smolensky, 1993/2004) are typically designed to consistently yield a single output for any given input, a quality which effectively models speakers' categorical phonological knowledge. Possibilities for modeling variable outputs include encoding the most common pattern as grammatical, and items which deviate from that as exceptions, and encoding no preference for either output in the grammar. The former predicts that only the most common pattern should extend to nonwords, and that it should extend categorically. The latter predicts that speakers should vary in their treatment of novel forms, but not in a way that matches the specific statistics of the lexicon.

Models which have adapted the constraint-based system of Optimality Theory (OT) to predict probabilistic behavior include systems of partially ordered constraints (Anttila, 1997), Stochastic OT (Boersma and Hayes, 2001), Noisy Harmonic Grammar (Pater, 2008), and Maximum Entropy Grammar (Goldwater and Johnson, 2003). Each of these models provides a unified explanation of categorical and probabilistic phonological behavior, using the same technology for both. In the most widely-used of these, Maximum Entropy (MaxEnt) grammar, constraints are each assigned weights, and when constraints compete their relative weight determines the probability that the outcome will observe one constraint or the other. In this way, abstract constraints can predict particular probabilities for particular outcomes. This type of model could predict, for example, that p/b will be voiced 9% of the time in a particular context.

Unlike categorical phonology, variable phonological patterns are typically conditioned by a wide array of factors (Bayley, 2002). Models which use analogy or spreading activation can easily capture the effects of many different factors on the outcome of a production or judgment. Because such models refer only to specific instances of lexical items and not to any features or classes of lexical items, it is actually *more* difficult for these models to capture a pattern the less different factors affect it. In models with abstract generalizations, the addition of more generalizations, or the addition of more features to the generalizations makes the system more complex.

While grammatical models of variable phonology and agrammatical models both have strengths and weaknesses, it is surprisingly difficult to distinguish the predictions of the two types of model for probability matching behavior. Ernestus and Baayen (2003) consider six different models for their data, two of which are OT-type models, and four of which do not use a grammatical mechanism. All models achieve a close fit to the experimental data, and Ernestus and Baayen ultimately decide in favor of the agrammatical models on the grounds that they are simpler. However, they also point out that their simplicity metric is flawed in that it counts parameters of each

model, but does not does not take into account other types of complexity such as the complexity of the learning process or the computational time required to produce a given output (which can be quite long for some analogical models).

In Experiment 3 presented below, I explore a different strategy for distinguishing between a grammatical model of probability matching, and a model which refers directly to the contents of the lexicon. If participants refer directly to specific actual words of their lexicon to choose a response for a novel word, then participants should behave differently on individual novel words based on which real word(s) those novel words are closest to. Novel words which are phonetically very close to a specific actual word should behave like that actual word. Novel words which are phonetically dissimilar from any actual word should vary in their behavior at close to chance, not following the probabilistic trends in the lexicon. Baker and Smith (1976) examined speakers' knowledge of English stress generalizations. They tested nonwords which were very close to particular actual words (e.g. *cinempa*), finding that participants' chosen stress patterns for the novel words matched the stress pattern of the neighbor, rather than the stress pattern predicted by the phonological generalizations proposed for English stress in e.g. (Chomsky and Halle, 1968). More recently, Guion et al. (2003) conducted a wug test of English stress generalizations in which nonwords were neither particularly close to any actual words nor particularly distant. They asked participants to provide possible analogical bases for each nonword, and found that the stress of that potential analogical base did influence participants' choice of stress for a nonword, but structural properties of the word (syllable weight and part of speech) exerted a greater influence. They argue that an analogical process is operating alongside a grammatical mechanism.

In the experiment presented here, participants are given nonwords which are phonetically distant from any actual words. Following the methodology of Guion et al., participants are asked to provide potential analogical bases for each nonword. A

67

system which does not represent the probabilistic trend grammatically predicts that such distant nonwords should either (a) be able to analogize to some word or group of words despite the phonetic distance, (b) vary between all possible stress patterns at chance (not following the generalization in the lexicon), or (c) be unstressable. Rather, I find that speakers can generalize the trend to nonwords, but that their chosen analogical bases do not relate to their choices of stress for the nonwords.

This result does not rule out the possibility that some analogical mechanism is part of the phonological system, but like Guion et al., I argue that a grammatical mechanism, containing abstract representations of probabilistic generalizations, is at work.

3.3 The stress generalization

The CMU pronouncing dictionary (Weide, 1994) together with the SUBTLEX$_{\text{US}}$ corpus (Brysbaert and New, 2009) was used to examine the distribution of stress in long words of English. The placement of stress in longer words of English has been described by the 'Latin Stress Rule' (Chomsky and Halle, 1968; Hayes, 1982, et seq.), which states that words with a heavy penultimate syllable take main stress on the penult (*bonánza*), while words with a light penultimate syllable take main stress on the antepenultimate syllable (*cínema*). While the first clause is nearly exceptionless in the lexicon [1], the second clause has many exceptions (Pater, 1994). An examination of the lexicon reveals that in words with a light penult, the probability of antepenultimate stress is conditioned by the word's final vowel. Specifically, words that end in [i] tend to take antepenultimate stress, while words that end in [ə] have no preference.

For purposes of this search, a light syllable is defined as one (a) with no coda consonant, and (b) whose vowel is a monophthong (one of [ɑ, æ, ɛ, ɔ, ɪ, ʊ, i, u, ə, ɚ]).

---

[1]exceptions include *gálaxy, cháracter, ádjective*

The maximal onset principle was used to determine syllabification, so e.g. 'palmistry' is treated as [pɔ.mɪ.stɹi], and included in the set of all-light words. However, a word like 'bonanza' ([bənæn.zə]) would count as heavy, since its first syllable must have a coda consonant - [nz] is not a legal onset in English.

Many long words are morphologically complex. Because particular affixes can influence the stress pattern of a word, either by attracting stress or by behaving as extrametrical (Chomsky and Halle, 1968; Burzio, 1994; Halle and Vergnaud, 1987), words were automatically coded for whether or not they were morphologically complex, using the spelling of the word as a proxy. For example, words ending in 'tion' were considered to end in the '-tion' affix. The list of suffixes and prefixes in Tescner and Whitley (2004), chapter 2, was taken to be exhaustive and words in the corpus with any of these strings at the appropriate edge of the word were marked as morphologically complex. Some affix strings were excluded because more simple words fit them than complex words. Examples are 'ab-', 'ad-', 're-', '-y' and '-o'. Ultimately, this method of marking words was variably successful depending on the length of the word - for two syllable words it was relatively conservative, tending to err in the direction of marking simplex words as morphologically complex, but since longer words are more often morphologically complex, the direction of error shifts.

The success of this morphological discrimination was assessed by randomly sampling 100 words from each category (morphologically simple, morphologically complex) for lengths of 2 syllables, 3 syllables, and 4 syllables. A native English speaker (the author) then checked these randomly sampled words (600 total) and noted the number of incorrect categorizations in each sample. The results of this categorization are shown in Table 4.6.

A total of 5101 words met the criteria of having a light penult, and being long enough to potentially take antepenultimate stress (at least trisyllabic). In these words, the most common final vowels were [ə] (1198 words), [i] (1531 words), [n̩/m̩/ŋ] (949

Table 3.1. Number of incorrect categorizations in each random sample of 100 words

| | Categorized as: | | |
	Simple	Complex	F1 score
2 syllables	8	72	0.43
3 syllables	12	13	0.87
4 syllables	26	2	0.84

words), and [l̩] (674 words). Main stress was almost always penultimate or antepenultimate. A small number of words took preantepenultimate main stress (365) or final main stress (155). How likely each stress pattern was depended on the final vowel. Words ending in [-i] or [-l̩] were more likely to take antepenultimate stress than penultimate stress, but words ending in [-ə] or syllabic nasals were roughly equally likely to take penultimate or antepenultimate main stress. These distributions are recorded in Figure 3.1.

Word-final [i], [l̩], and also [ɹ̩] have been treated as special by Chomsky and Halle (1968); Liberman and Prince (1977); Hayes (1982), all of whom noticed that words ending in these segments were more likely to have exceptional preantepenultimate stress, e.g. *álligator, állegory, párticiple*. In these words, the final [i/l̩/ɹ̩] seem to be ignored for purposes of stress placement - possibly because (Chomsky and Halle; Liberman and Prince) the [i/l̩/ɹ̩] are underlyingly consonantal [j/l/ɹ]. In this case, álligator would be stressed as the three syllable form with a final heavy syllable [æ.lɪ.ge͡ɪtɹ], and receive antepenultimate stress. In this case, the surface form obtains through a rule of sonorant syllabification. Another possibility (Hayes) is that final [i/l̩/ɹ̩] are marked as extrametrical and are simply ignored by stress assignment rules.

The corpus data in Figure 3.1 demonstrates that words with final [i/l̩/ɹ̩] also tend to have much more antepenultimate stress than penultimate stress. This would be predicted if these segments tend to be extrametrical - a three syllable word like ćannery would be treated as a two syllable word, [kæ.nə], and take initial stress.

Figure 3.1. Words with light penultimate syllables and the five most common final syllable nucelii, from the CMU pronouncing dictionary (North American English).
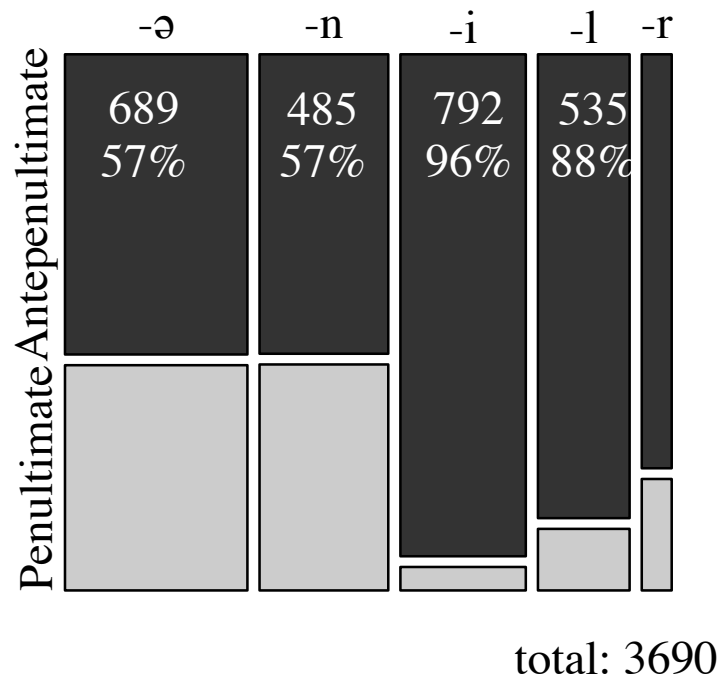
Many [i/l̩/ɹ̩]-final words in English are morphologically complex, and specifically contain a suffix which is stressless or which demands antepenultimate stress, e.g. -y (yellow ∼ yellowy), -er (yellow ∼ yellower), -ity (absúrd ∼ absúrdity), -able (avért ∼ avértable). However, the statistical differences between these three final vowels and final schwa persists in monomorphemic words. Figure 3.2 illustrates this with just -i and -ə.

Experiment 3 presented below will directly compare participants' behavior on items with final -i vs. final -ə. Items are exactly three syllables long, and have only light syllables.

Figure 3.2. Words with light penultimate syllables, classified as morphologically simple or complex, from the CMU pronouncing dictionary (North American English). Only words which end in either [i] or [ə] are included.
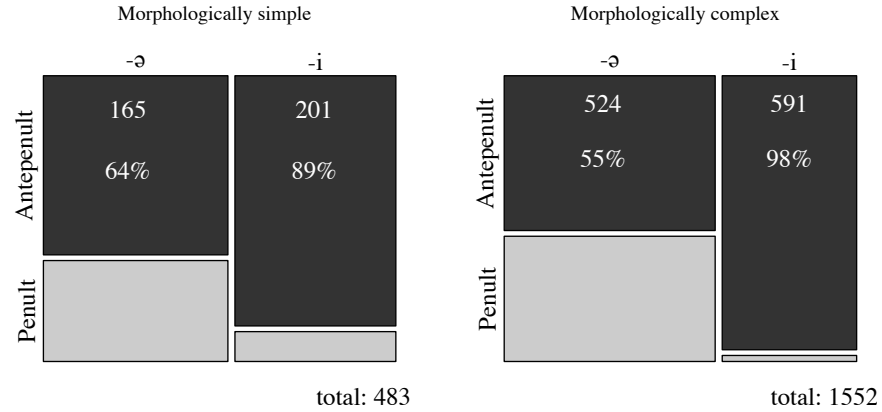
3.4   Analogical models of Productivity

While there are many models on the market which explain phonological productivity through more or less direct reference to the lexicon, I will focus here on those in which analogy proceeds in real-time during the production process, and in which no special learning process is required. Some examples of models which do not meet these criteria are connectionist models (Rumelhart and McClelland, 1986) which require a learning phase, the Generalized Context Model (Nosofsky, 1990; Nakisa et al., 2001), which relies on speakers' gestalt knowledge of their entire lexicon; and Skousen's Analogical Modeling of Language (Skousen, 1989; Eddington, 2000), and Daelemans' Instance Based Learner (Daelemans et al., 1994), which require a learning process for the model to assign different degrees of importance to different features of a lexical item. Because of this learning process, it is harder to differentiate the predictions of these models from the predictions of models with an explicit grammatical component. Instead, I will focus on the TRACE model (McClelland and Elman, 1986). This is a model of lexical access that at least in some cases can generalize trends in the lexicon to new words 'for free'.

McClelland and Elman (1986) develop the TRACE model of speech perception, which is a cascading activation model designed to solve the problem of how listeners use acoustic information to discover the intended lexical item, given that that acoustic information unfolds gradually over time, and that often information for different phonemes overlaps in time. Their model also predicts, essentially as a side effect, that patterns in the lexicon of a language could influence the perception of novel words. The example they give is of the difference between [sl] and [sɹ] onsets in English - [sl] onsets are abundant, while [sɹ] onsets are disallowed. In perception, listeners will categorize an ambiguous sound as [l] more often than as [ɹ] after an s than after an f, where both are allowed (Massaro and Cohen, 1983). According to the TRACE model, an ambiguous [sl/ɹ] sequence would activate actual words with [sl] onsets, like 'sleet' and 'sleep', but would not activate any [sɹ] sequences since those are absent. Because activation feeds backwards from those activated lexical items, the result would be a percept of an [sl] cluster rather than an [sɹ] cluster.

To extend this explanation to the stress case, and to a production paradigm, imagine that a listener heard a novel word with an ambiguous stress pattern (all syllables would have full vowels, and have matching pitch, loudness, etc.). The segmental material in the nonword would cause particular lexical items to become activated, and those actual words would each have a specific stress pattern. If the listener then had to choose a stress pattern with which to produce the nonword, the stress pattern(s) of the activated real words would be primed. If a stress pattern, A, is more pervasive in the lexicon than a competing stress pattern, B, then it will be more likely overall that the words activated by the perception of the nonword will themselves have stress pattern A. Because it is more likely that words activated in the perception of the nonword have stress pattern A, it will also be more likely that a participant would produce stress pattern A on that nonword.

73

To work through an example, suppose the nonword [dækæθi], with no syllable clearly hosting main stress. According to the TRACE model, the acoustic information in this nonword would cause actual words to be activated that are similar in some ways, such as having three syllables, æ-nucleii, the consonants d, k, or θ, or a final i. For example, the word 'apathy' has three syllables, a final i, and an æ-nucleus in the first syllable. The word 'decathlon' also has three syllables and an æ-nucleus, but no final i. However, it shares all three onset consonants with the nonword. Which components of the nonword are most influential in the spreading-activation process could depend on a variety of factors, including the acoustic details of the segments, the order in which the segments are encountered, statistical information about each segment in the lexicon, and even aspects of the discourse context. Suppose that the d in [dækæθi] had a relatively long VOT, making it acoustically somewhat t-like. In that case, words with both d and t would be activated. The order of the segments matters too. In TRACE, activation spreads upward immediately as soon as acoustic information is available. That means that the listener would activate the phonemes d and æ before activating phonemes for the following segments, and activation would immediately spread from these phonemes to word-level units containing them. Information later in the word could override this initial activation on words containing the initial segments, but there will still be some preference.

If a listener perceiving the ambiguously-stressed nonword [dækæθi] arrived at a state where 'ápathy' received the most activation (let's call this the case where 'ápathy' is *accessed*), that listener would then be more likely to produce the word as [dǽkəθi], with antepenultimate stress. On the other hand, if she arrived at a state where 'decáthlon' received the most activation ('decáthlon' is accessed), she would be more likely to produce [dəkǽθi], with penultimate stress. If the structure of the lexicon is such that 'ápathy' or other antepenultimately-stressed words are accessed more often than 'decáthlon' or other penultimately stressed words, then in aggregate

more speakers will choose antepenultimate stress for [dækæθi]. For some nonwords, there will be a single actual word which most speakers will access, while for others there may be multiple actual words which are equally similar in segmental content to the nonword, so that different speakers will access different ones. In either case, across many nonwords, the accessed word will be more likely to have the stress pattern which prevails in the lexicon than to have the minority stress pattern, and therefore in aggregate speakers will produce more of the prevailing stress pattern.

In Experiment 3, participants first listen to novel words with ambiguous stress patterns and produce them (choosing a stress pattern). In the second half of the experiment, they are given the nonwords again and asked for each 'What real word does this remind you of?' They fill in a blank with their answer. This basic methodology has been used before by Guion et al. (2003) to assess whether characteristics of particular lexical items were influencing participants' choices of stress pattern in production. A crucial assumption here is that each participant will, at least most of the time, access the same actual word every time they hear a particular nonword. If this is true, then the word participants fill in the blank with can be assumed to be the same word that was accessed during the production task earlier in the experiment. If participants are deciding on a stress pattern for each nonword based solely (or mainly) on the stress pattern of the 'nearest' lexical item, then the stress pattern of each provided real word should strongly correlate with that participant's choice of stress pattern for that nonword in production.

3.5 Experiment 3

3.5.1 Introduction

This experiment was modeled after Guion et al. (2003), in which the productivity of certain trends in the English stress system was tested by asking participants to pronounce novel English words. The challenge for a production task for the English stress

75

system is that English orthography is non-transparent, and different participants may interpret one orthographic string in many different ways. Vowels are especially difficult to represent unambiguously in English orthography, which is problematic because the quality of a vowel is one factor which can affect the stress of a word. On the other hand, it is difficult to present a novel word auditorily without giving it some stress pattern. Guion et al. solved this problem by presenting auditorily strings of individual syllables, each pronounced as a separate prosodic word, and asking participants to string the syllables together into a word. I copy that methodology here.

3.5.2 Methods

3.5.2.1 Participants

The experiment was presented via the world wide web, and participants were recruited through word of mouth, and through Amazon Mechanical Turk. All participants were over 18 years of age, and had an age range of 19 to 61 (mean age: 33). Only IP addresses originating in the United States were accepted. Participants were asked where they were from, and "when you speak English, where do people think you are from?" If their answer to the second question was a location within the united states, they were assumed to be a native speaker of American English. Participants were paid at a rate of $0.91 for the experiment, which took about 20 minutes. Data was collected from a total of 104 participants, and data from 65 participants was used. The remaining participants were excluded because of problems with the sound recording and native speaker status. The process of excluding participants is described in detail in the results section.

3.5.2.2 Items

Items were a subset of the items used in the ERP experiments, so that all were three syllables long, consisting only of light syllables (codaless syllables whose vowel is a monophthong). Both novel words and real words were used. Because participants

76

were asked to pronounce a novel word after listening to three isolated syllables, real words were included in order to encourage participants to make their productions as like real English words as possible - in particular to encourage them to reduce unstressed vowels in their productions. Real words used in the experiment were evenly split among the four word shapes used in the ERP experiments (-i final antepenultimate stress, -i final penultimate stress, -əfinal antepenultimate stress, -əfinal penultimate stress).

Each item consisted of three auditorily presented individual syllables, and two auditorily presented versions of the full word, with different stress patterns (Antepenultimate and Penultimate). Participants first heard the syllables, then pronounced the word, then heard the two stress option, and chose between them. An example item is shown in Figure 3.3.

Figure 3.3. Example item from the production section of Experiment 3: all presentation was auditory.

◀)) [bæ] [mæ] [ki]
RECORD...
What did you say?
◀)) [bǽməki]   ◀)) [bəmǽki]
☐   ☐
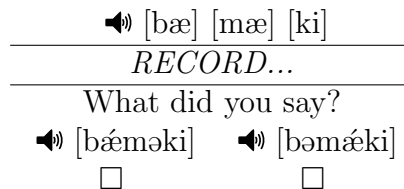
Items were 80% novel words and 20% actual words (32 novel words and 8 actual words). When they were actual words, the two stress choices were (1) the actual word, and (2) a mis-stressed version of the actual word, e.g. [kǽnədə] and [kənǽdə]. All items (words and nonwords) had the same stressed vowel in each stress version, and in all cases, the first two of the three individual syllables in an item had the same stressed vowel.

For the two stress versions of each item, the stimuli were transcribed into the international phonetic alphabet (IPA) and pronounced in a random order by a male native speaker of American English, in the frame sentence "Say X again." The words

Figure 3.4. Pitch contour for individual syllables presented to participants. All syllables were resynthesized to have this contour
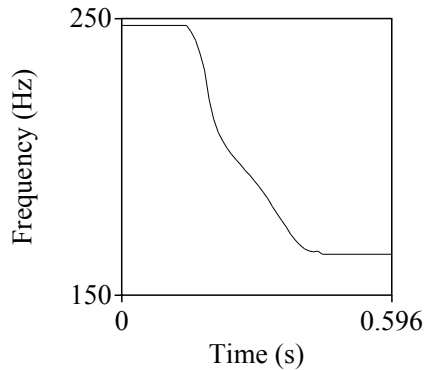
were then spliced out of the frame sentence. Eight real words were also included - two words for each stress x vowel quality condition. In each case, the stressed vowels of the correctly stressed and misstressed versions of the word matched.

Nonwords were counterbalanced for their final vowel. Two lists were made: in the first list, each nonword was randomly assigned a final vowel so that half were [i] and half were [ə]. In the second list, each item appeared with the final vowel opposite that used in the first list. Participants were assigned one of the two lists at random. The lexical neighborhood density of each nonword was measured using the Generalized Neighborhood Model (Bailey and Hahn, 2001). All nonwords used in the experiment had a GNM value of less than 0.01, corresponding to very sparse neighborhoods.

The isolated syllables were constructed in the following way: A female native speaker of American English (the author) read a list of individual syllables written in IPA. These recordings were then resynthesized in Praat (Boersma and Weenink, 2011) so that each vowel was approximately 400ms long, and faded into silence over the final 100ms. The pitch contour of the syllables was also resynthesized to be identical (a H* pitch accent followed by a H-H% boundary tone, shown in Figure 4.3). The intensity of the syllables was also normalized.

The same 32 nonwords that were used in the production task were used in the analogical base task. Participants were divided into two groups: Group 1 heard full word pronunciations, where one stress pattern or the other was randomly selected for each item. Group 2 heard the individual syllable prompts.

3.5.2.3 Procedure

The experiment was presented via the world wide web, using software built on Experigen Becker and Levine. Each participant first completed the production study, and then completed the analogical base study. When participants arrived at the site, they were first asked to electronically sign a consent form, and then they completed a sound check to test that their microphone and speakers were working. Next, for part 1, they were instructed that they would hear a sequence of three syllables, and that they should speak the whole word fluently as if it were a real word. They were given an example nonword sequence of syllables and two examples of those syllables strung together into a pseudoword - once with antepenultimate stress and once with penultimate stress. Next, they were given a sample trial which was a real word (they were told in advance that it would be a real word). In each trial of the experiment, they first heard the three syllables, then were asked to speak the word fluently, then they listened to the two stress options for that item and clicked a radio button to choose one. There were 32 nonword items and 8 real word items.

In the second part of the experiment, participants were instructed to listen to a stimulus, and fill in the blank with a real word that it reminded them of. The same 32 nonwords used in the production task were used in the analogical base task. Participants were divided into two groups. Group 1 heard full-word pronunciations of each nonword. For each nonword, antepenultimate or penultimate stress was randomly selected, and participants heard only that version of the word. In Group 2, participants heard the same three-syllable prompts which were used in the production task.

3.5.3 Results

Participants' success at the production task was assessed in two parts. First, did participants produce the syllables fluently together as a single word, with a single main stress? Second, did their produced stress agree with the stress they reported producing? For each participant, 10 (out of 32) nonword recordings were randomly selected. The author listened to these and annotated whether the production had a single stressed syllable or not, and transcribed the location of the main stress if it had one. Stress was assessed based on vowel reduction and pitch. If a production had a full vowel in every syllable, or both of the first two syllables, it was counted as 'incorrect'. Also, if the production contained a pitch fall on any syllable but the last, or pauses between the syllables, it was counted as 'incorrect'. A participant was excluded from analysis if more than three of the examined 10 nonwords counted as 'incorrect'. Participants who did not successfully record any sound were also excluded. In total, 22 (out of 104) participants were excluded for these reasons. Two additional participants were excluded because they were not native speakers.

For the 'correct' productions, which followed the criteria of being a single prosodic unit in which at most one syllable bears main stress, participants' accuracy at reporting their own stress pattern was assessed. For 65 participants, their choice of stress pattern in the forced choice task agreed with the author's transcription of their produced stress at least 9 times out of 10. 15 Participants had less than 90% accuracy on the forced choice task, and were therefore excluded from analysis. Both the quality of the productions, and the agreement between participants' productions and the forced choice task were held to relatively stringent standards because the assumption about this data is that it is production data, not perception data. If a participant did not produce the items in a word-like fashion, then their choice of stress patterns in the forced choice task cannot be based on the action of the production system. Likewise, if a participant produced one stress on a form, but chose another after hearing both

80

Figure 3.5. Counts of stress choices for each final vowel in nonwords

Forced choice responses
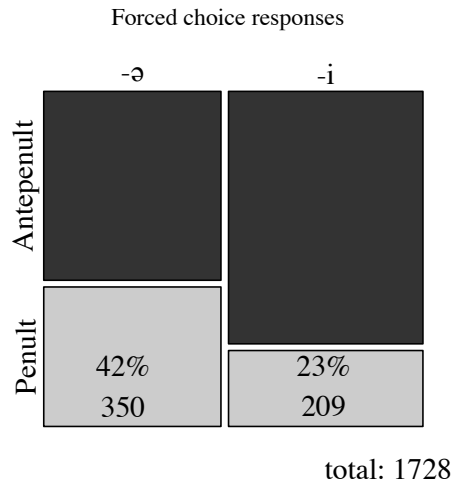
options, then their choice is not based on the action of their production system, but rather on a combination of the perception of each stress and a high-level judgment.

3.5.3.1 Production results

Both groups of participants did the same production task, so their results were analyzed together. Overall, participants' productions of nonwords probabilistically obeyed the strong i-final generalization, and did not obey the weak ə-final generalization.

Figure 3.5 shows the counts of each type of stress response for each type of final vowel. These counts are responses in the forced choice task, but recall that participants are only included if their choice of stress agreed with their produced stress at least 90% of the time. This means the counts can be thought of as production counts. Additionally, trials in which participants failed to listen to both stress options before responding were excluded.

Overall, participants preferred antepenultimate stress for both ə-final nonwords and i-final nonwords, but this preference was slight in the ə-final case, and relatively

strong in the i-final case. This pattern matches the distribution in the lexicon among all three-syllable and longer words. Note also that participants responded with a probabilistic rather than a categorical preference for antepenultimate stress on i-final nonwords.

A mixed effects logistic regression was fitted to this data, with produced stress as the dependent variable, final vowel as the predictor, and including random slopes and intercepts for both subjects and items. Penultimate stress was the baseline value, so negative coefficients indicate a preference for antepenultimate stress. There was a slight preference for antepenultimate stress over penultimate stress over all (Intercept=-0.35, p=0.049), and a stronger preference when the final vowel was [i] (β= -1.27, p¡0.001).

3.5.3.2 Analogy results

The results of the analogical task were analyzed separately for the two groups - recall that group 1 heard full word productions with a particular (randomly selected) stress pattern, while group 2 heard three-syllable utterances with ambiguous stress just as in the production task. There were 33 participants in group 1 and 32 in group 2. For both groups, participants gave a mixture of real-word responses, short phrases (e.g. 'the panda' for [tǽməpə]), and transcriptions of the nonwords. Responses which were misspelled single real words were included in the analysis with spelling corrected. The percentage of single, real-word answers in each group was similar - 56% for group 1 and 57% for group 2. This number varied greatly among participants, ranging from 0 to 84% with a median of 63%.

In group 1, participants mostly responded with two and three syllable words. Their three-syllable responses tended to match the stress pattern of the prompt they were given. If the prompt was an antepenultimately stressed sequence like [bǽməki] they were more likely to write an antepenultimately stressed word in the blank, while

if the prompt was penultimately stressed, like [bəmǽki], they were more likely to write down a penultimately stressed word in the blank. Figure 3.6 shows the relationship between the stress of the prompt and the stress of the response, for all three syllable responses.

Figure 3.6. Agreement between the stress of the prompt and the stress of the word participants give in response: only three-syllable responses are included here.
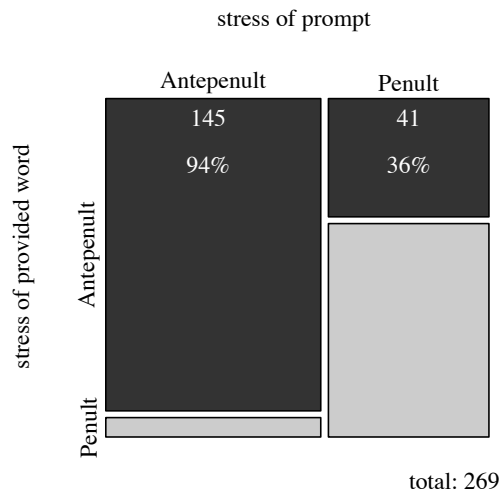
When the prompt is antepenultimately stressed, the vast majority of responses were also antepenultimately stressed. When the prompt was penultimately stressed, the majority of words were penultimately stressed, but the trend for agreement was much less strong.

The stress of the analogical bases provided by group 1 did not in general predict their choices of stress in the production task. Figure 3.7 shows the relationship between the stress of the word participants wrote down in the analogical base task and the stress they produced for that item in the production task. The majority of produced stresses were antepenultimate, regardless of the stress of the word given in the analogical base task, and there does not appear to be any preference for a participant's produced stress to match the stress of their chosen analogical base. This

83

is also true if only bases that disagree with the stress of the prompt are examined, although in this latter case the numbers are quite small.

Figure 3.7. The relationship between the stress of the participant-provided analogical base, and that participant's produced stress for that item: The left figure shows all analogical bases, while the right shows only those which disagree in stress with the prompt
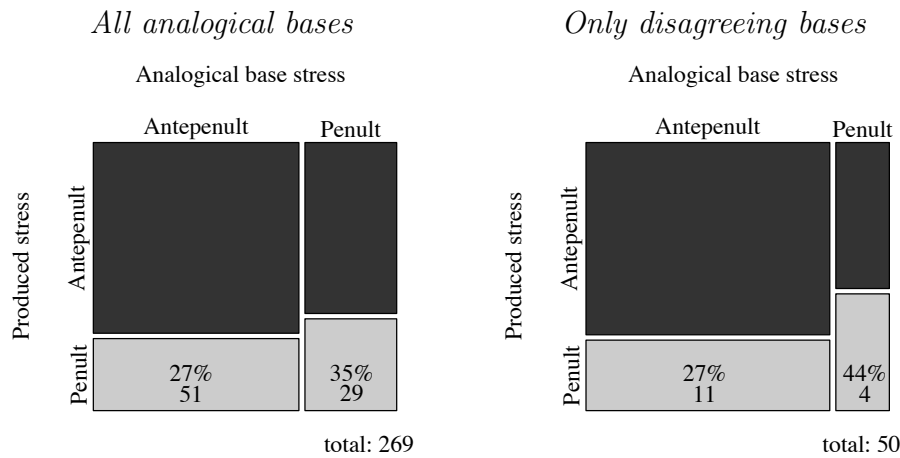
Although the task given to group 1 was relatively natural - listen to a felicitous nonword and write down a near actual word neighbor - the bases provided in this task may not be the same bases which would be accessed during the production task. From participants' responses, it is clear that the stress pattern of the prompt in the analogical task partially determined their choice of real word, but in the production task the prompts did not have a stress pattern. The words that a participant accessed while listening to three separate monosyllables with no stress relationship between them might be different than the words accessed while listening to a nonword with a particular stress pattern.

Participants in group 2 were given the exact same prompts in the analogical task as in the production task. Words given in response to these prompts varied in how long they were, with three-syllabled words being the most common. Table 3.2 shows the number of words produced with in each length category.

Table 3.2. Lengths of words given in group 2's analogical task - about half are long enough to host antepenultimate stress

Number of Syllables				
1	2	3	4	5
118	175	298	40	4
19%	28%	54%		

Analogical bases provided by the participants typically had the same final vowel as the nonword prompt. For i-final prompts, the given analogical base ended in [i] 63% of the time overall, and 92% of the time in three-syllable bases. For ə-final prompts, the given analogical base ended in [ə] 36% of the time overall, and 50% of the time in three-syllable bases. In all cases, agreement was the most common outcome. The analogical bases provided by participants tended to be antepenultimately stressed overall, but especially when they ended in [i]. Figure 3.8 illustrates this.

Figure 3.8. Stress of provided analogical bases which ended in [i] or [ə]: final vowel indicates the final vowel of the actual provided base, not the final vowel of the nonword prompt.

Analogical base final vowel

total: 303
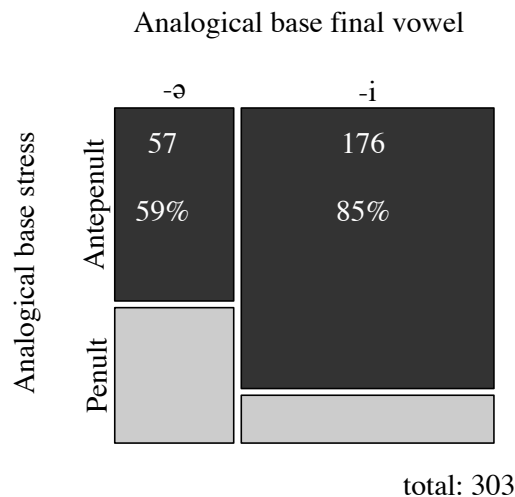
The set of provided analogical bases are a sample of the lexicon of English - because of this, the default assumption should be that they would follow the trends in the lexicon of English. It is therefore not surprising that i-final bases strongly tend to be antepenultimately stressed, while ə-final bases tend only weakly to be antepenultimately stressed. These tendencies mirror the tendencies exhibited by participants in the production task. The question is: is participants' behavior in the production task the result of analogizing to some existing word on each trial? Or, does their behavior reflect the presence of an abstract generalization that prefers antepenultimate stress on i-final words?

Figure 3.9 shows the relationship between the stress of the analogical base provided for an item by a participant, and that participant's produced stress pattern on that item. Only analogical bases long enough to host antepenultimate stress (3 syllables long or longer) are included. In both the produced stress patterns and in the choice of analogical bases, participants give more antepenultimate stress on i-final words than on ə-final words, but there does not appear to be a direct relationship between the stress of the given base and the produced stress on that item. The items are divided by final vowel, and in neither i-final nor ə-final items is there a clear relationship.

A logistic regression on just data from group 2 was fitted. The produced stress pattern was the dependent variable, and fixed effects were the final vowel of the item and the stress pattern of the analogical base provided for that item by that participant.

In the model with both factors, reported in Table 3.3, the final vowel of the stimulus has a large coefficient and is highly significant. The stress of the analogical base has only a small coefficient and is not significant. In order to assess the relative contributions of the two factors to the overall model fit, each one was dropped from the model and the fit of that simpler model was compared to the fit of the model with both factors.

86

Figure 3.9. The relationship between the stress of the participant-provided analogical base, and that participant's produced stress for that item: The left figure shows i-final items, and the right figure shows ə-final items. Only analogical responses 3 syllables long and longer are included.
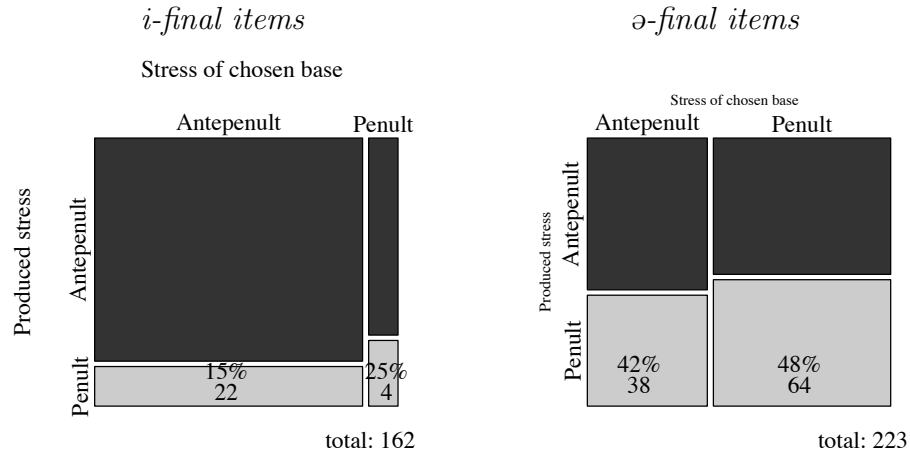
Table 3.3. Logistic regression with two factors. Negative coefficients mean greater chance of antepenultimate stress, while positive coefficients mean greater chance of penultimate stress

Model: *Produced Stress ∼ Final Vowel + Analogical Base Stress*		
	Estimate	p
Intercept	-0.54	0.02
Final Vowel = i	-1.22	0.0001
Analogical Base Stress = Penult.	0.42	0.20
AIC:	290	

Table 3.4. Model comparison between regression models. Each factor was dropped from the model, and that resulting simpler model was compared to the full model.

remove:	change in AIC	Likelihood ratio	p
Final Vowel	+13	15.66	0.0001
Analogical Base Stress	0	1.7	0.20

The results shown in Table 3.4 demonstrate that if the stress of the analogical base is removed from the model, the fit is not worsened. On the other hand, when

final vowel is removed as a predictor, the fit of the model is significantly worse. Based on these results, it is possible to conclude that the stress of the analogical base is not related to the produced stress of that item.

If the assumption holds that each participant accesses the same real word each time they hear a particular nonword stimulus, then these results indicate that participants are not using that accessed nonword to make their choice about stress on the nonword. However, words may be accessed non-deterministically. For each item, there may be a set of 'nearest neighbors', one of which will be accessed when a participant hears a certain stimulus, but perhaps not the same one each time. In this case, there may not be a direct correlation between the stress of an analogical base provided by a participant, and that participant's chosen stress for that item. Instead, items with more antepenultimately-stressed 'nearest neighbors' will be more likely to take antepenultimate stress, and words with more penultimately-stressed nearest neighbors will be more likely to take penultimate stress.

In order to test this possibility, the set of words given as analogical bases for each item was examined. For each item, the percentage of analogical base responses with each stress pattern was calculated. If participants do not access exactly the same lexical item every time they hear a particular nonword stimulus, but instead access one of a relatively circumscribed set of lexical entries, then for each item in the experiment, the rate of occurrence of a particular stress pattern in the analogical bases provided for that item should predict the rate of occurrence of that stress pattern in production.

Table 3.5 lists the different bases that participants wrote down for the item [rɛ mɛ nə]. Two-thirds of the bases had antepenultimate stress and one-third had penultimate stress. Only bases long enough to take antepenultimate stress (3 syllables or longer) are analyzed here, although results including two-syllable bases as well are very similar. This was about half of the responses.

Table 3.5. Example list of analogical base responses for a stimulus. For this stimulus, a base with antepenultimate stress was given 2/3 of the time.

Stimulus	[rɛ mɛ nə]			
Analogical Base	banána	rémedy	hármony	pánama
no. Responses	2	2	1	1

67% Antepenult, 33% Penult

For some items, participants strongly agree with each other in their choices of analogical bases. For example, for the item [sɛ lɛ kə], 9 participants wrote down the word 'silica', and only one wrote down the word 'saliva'. For other items, there is less agreement. For the item [tɛ pɛ di], participants wrote down 'tahiti', 'parody', 'tapestry', and 'tragedy'.

For most items, the most common stress pattern among the given analogical bases was antepenultimate. This was also modulated by the item's final vowel. Table 3.6 shows that more ə-final stimuli than i-final stimuli had majority penultimately stressed bases. This is expected, since there are more ə-final penultimately stressed words in the English lexicon than i-final penultimately stressed words. This table shows the same effect as Figure 3.8 - namely that participants' choices of analogical base generally tend towards being antepenultimately stressed but otherwise mimic the trends found in the lexicon.

Table 3.6. Most common stress among the given analogical bases, for each experimental item. Items are not included if no given base for them was 3 syllables long or longer.

Final Vowel	Most common stress		
	Antepenultimate	Penultimate	Pre-antepenultimate
-ə	21	10	-
-i	26	4	1

Figure 3.10 shows that there is no correlation between how many bases given for an item were antepenultimately stressed and how likely that item was to be

89

given antepenultimate stress in the production task. Each point in the plot is one nonword item. Overall, i-final items take higher percentages of antepenultimate stress in production than ə-final items do, but items in both categories vary widely in the percentage of antepenultimate stress among their analogical bases. Whenever stress was not antepenultimate it was almost always penultimate, so the relationship between the two rates of penultimate stress is similarly noncorrelative. There were just a few given analogical bases which were neither antepenultimate nor penultimate, but pre-antepenultimate. Ten nonword items got at least one pre-antepenultimate base - 8 i-final and 2 ə-final, but only one got majority pre-antepenultimate bases, namely [dɛ lɛ si], for which 8 participants wrote 'délicacy'.

Figure 3.10. For each nonword item: the percentage of provided bases which had antepenultimate stress vs. the percentage of productions of that item with antepenultimate stress. There is no correlation between these two percentages.

% Antepenultimate stress by item
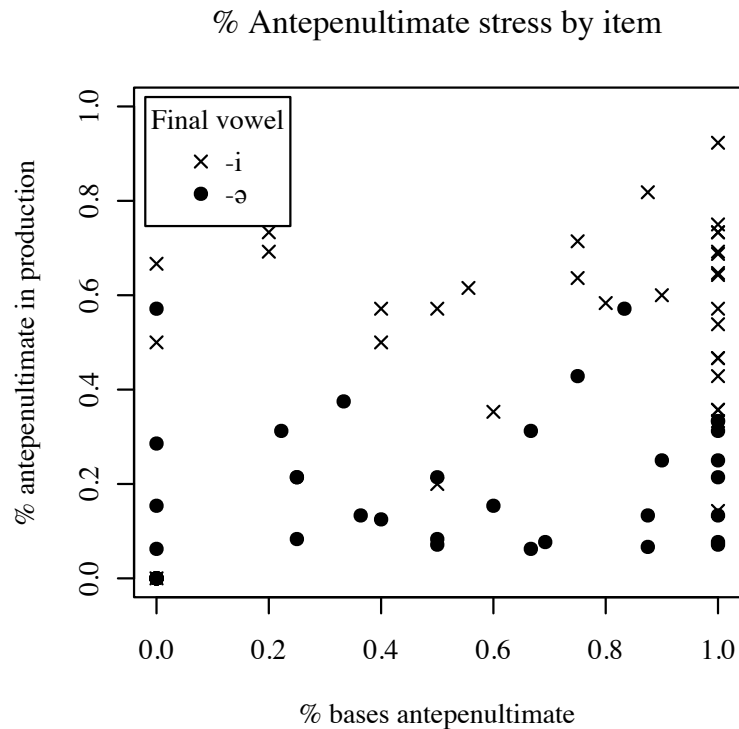
To summarize the findings of the analogical base task: The analogical bases provided by participants in aggregate followed the lexical trends: i-final bases were antepenultimately stressed more often than ə-final bases were. This same trend was obeyed in production, but there is no evidence that participants were using their lexicon directly to perform the production task. Individual participants' choice of analogical base did not correlate with their choice of stress on that item in production, and the composition of the set of analogical bases given for an item did not predict the distribution of stress patterns on that item in production.

3.5.4 Discussion

The results of the production task showed that participants differentiate between i-final and ə-final words, preferring antepenultimate stress much more on i-final words. The set of words which participants provided in the analogical base task also roughly matched these lexical statistics, but the stress of the bases did not directly predict participants' choice of stress for each nonword.

In both the production task and the analogical base task, participants exhibited a slight overall preference for antepenultimate stress. In the production task, this preference could be the result of participants preferentially preserving the vowel quality of the first syllable in the stimulus. Preferences for preserving material in initial position over material in other positions have been noted in many languages, and are typically formalized in Optimality Theory as positional faithfulness (Beckman, 1997; Kawahara and Shinohara, 2011). Ernestus and Baayen (2003) observed a similar faithfulness to the nonword stimulus in their probability-matching experiment.

In the analogical base task, participants are explicitly asked to retrieve a real word based on the given nonword input. In this case, the preference for finding real words with antepenultimate stress may be due to the importance of a word's initial syllable in lexical retrieval (Nooteboom, 1981; Horowitz et al., 1968). Participants may be

more likely to find actual words whose initial syllable matches the stimulus's first syllable. Since the stimulus's first syllable was always stressed, and contained a full vowel, the retrieved word would then typically be stressed on the first syllable.

Another possible explanation for the overall preference for antepenultimate stress is that participants are choosing more antepenultimate stress not because it is antepenultimate but because it is initial. Initial stress is arguably the most common type of stress in the English lexicon, and preference for initial stress on nonwords has been shown both in word-segmentation tasks (Thiessen and Saffran, 2003) and in judgment tasks (Guion et al., 2003; Cutler and Carter, 1987; Jusczyk et al., 1993). If this overall preference for antepenultimate stress can be characterized as a preference for initial stress, this would mean that participants are exhibiting two distinct 'layers' of knowledge about the lexicon in their productions. They know the relatively detailed generalization that a word prefers to be antepenultimately stressed just in case it ends in [i], and they also know the more general preference for antepenultimate stress across all types of words.

The preference for antepenultimate stress on i-final words is probabilistic - participants did not observe it with every production, but rather they observed it in aggregate 77% of the time, a rate which is comparable to the generalization's rate of observance in the lexicon of English. Recall (Figure 3.2) that in the lexicon i-final words took antepenultimate stress between 89% (monomorphemic words) and 98% (morphologically complex words) of the time.

Is this probabilistic behavior the result of an abstract generalization which is part of participants' phonological grammar? Or is it the result of directly consulting the lexicon on each production? The analogical base task was designed to provide evidence to differentiate between these two possibilities.

When a participant hears a speech stimulus, even if it is a nonword, an automatic lexical access process must ensue. This process will ultimately be unsuccessful, but

may activate one or more actual lexical items which are similar to the stimulus in some way. If this lexical access process is deterministic, then given a particular lexicon the same word or words will become activated each time the system is exposed to the same stimulus. Individual participants in the experiment may have sightly different lexicons, with different lexical items and different frequencies over them, but each participant should use their lexicon in a similar way each time they encounter the same stimulus. If this is so, then the word a participant writes down during the analogical base task will be the same actual word which was most activated when that participant performed the production task. If participants used the stress pattern of this 'nearest neighbor' to perform the production task, then each participant's response on the fill-in-the-blank task would predict the stress pattern that they chose for that item.

However, there does not appear to be a clear relationship between the real word items that participants report on the analogical base task, and their choice of stress patterns on the production task. Group 1 and group 2 of participants performed slightly different analogical tasks, but no relationship between analogical base stress and produced stress was found in either case.

On the other hand, the lexical access process on a nonword may be non-deterministic. In this case, there would not be a single lexical item which becomes activated each time a participant heard a particular stimulus, but instead all words in the lexicon would have some probability of being accessed upon each exposure to the stimulus. Most of the probability would fall on some relatively small set of lexical items, though - words which are similar to the nonword in some way. If a nonword's set of most-similar words contained mostly antepenultimately stressed words, participants would be most likely to access an antepenultimately stressed word, and therefore give that nonword antepenultimate stress. If that set contained mostly penultimate stress, they would be more likely to give the nonword penultimate stress. In the analysis section,

I examined the set of real words provided by participants for each item in Group 2[2]. The percentage of provided analogical bases with a particular stress pattern did not predict the rate of occurrence of that stress pattern in production.

In aggregate, participants' choices of analogical base followed the same trend as participants' productions: in both cases, i-final words were more likely to take ante-penultimate stress than ə-final words. However, this experiment found no evidence that participants were directly using the stress pattern of particular lexical items to make their choice about the stress pattern of a nonword in the production task.

The nonword items in this experiment were specifically designed to not be particularly similar to any actual words. All had sparse neighborhoods according to the generalized neighborhood model (Bailey and Hahn, 2001), and participants in the experiment had trouble coming up with words in the analogical base task. The average number of valid responses per item in this task was 10, even though 32-33 participants saw each item. These results therefore do not rule out the possibility that participants' *could* analogize to a particular lexical item in producing a nonword if the nonwords were constructed so as to be closer to some particular lexical item, or to have dense neighborhoods. Baker and Smith (1976) specifically test nonwords that are very close to actual words (the example they give is 'cinempa'), and find in fact that the behavior of nearby words affects the behavior of nonwords more than things like syllable weight and part of speech (specifically, noun vs. verb). In an experiment very similar to the one presented here, Guion et al. (2003) found with two-syllable nonwords that the stress pattern of participants' choice of analogical base predicted their choice of stress on a nonword, but not as reliably as syllable weight and part of speech. Guion et al neither specifically manipulated their nonwords' similarity to

---

[2]Group 1's responses were not examined because these participants were responding to an item which already had a stress pattern, and in most cases their response also had that stress pattern. Because of this effect, the set of stress patterns given in response to a particular nonword would be artificially balanced.

particular words nor specifically designed their words to be dissimilar to any actual words. In general, two-syllable words will have denser neighborhoods than three-syllable words, just because there are many more two-syllable than three-syllable words in English. Thus, Guion et al's nonwords likely had denser neighborhoods than the nonwords used in the experiment presented here.

What the analogical base task presented here has shown is that when participants cannot or do not use the behavior of a particular lexical item to choose a stress pattern in production, they still observe the i-final generalization. That is, they still prefer antepenultimate stress on i-final words which are long enough to take antepenultimate stress. No such preference is observed on ə-final words. This i-final preference is probabilistic in nature and does not apply 100% of the time, but rather about 80% of the time. A number which is very close to the observed rate of antepenultimate stress on i-final words in the lexicon.

3.6   Grammatical representation of the i-final trend

This section presents a Maximum Entropy (MaxEnt) analysis of the trend for words ending in [i] to take antepenultimate stress. The analysis builds on the analyses of primary stress placement given in Pater (2000), and Alcántara (1998), both of which use the constraints ALIGN-HEAD-R, NONFINALITY, FOOTBINARITY, and TROCHEE. In this paper, TROCHEE, the constraint which demands that feet be left-headed, is assumed to have a very high weight, and only candidates which satisfy it are considered. In order to grammatically model the i-final trend, I add to this constraint set a version of NONFINALITY which applies just to words that end in [i]. This constraint has the same effect as an extrametricality rule marking word-final [i] as extrametrical (Hayes, 1982).

Main stress in English typically occurs within a three-syllable window at the right edge of the word. This is predicted by the joint action of an alignment constraint

demanding that main stress be as close as possible to the right edge of the word (ALIGN-R), and a nonfinality constraint demanding that the final syllable of a word be unfooted (FOOT-NONFINALITY) (Pater, 2000, p. 240, Alcántara, 1998, p. 120-121).

(2)   ALIGN-HEAD-R (ALIGN-R): Assign a violation for every syllable intervening between the right edge of the word and the main stressed syllable

(3)   FOOT-NONFINALITY (NONFIN): Assign a violation if the final syllable of the word is parsed into a foot.

Figure 3.11. Violations assigned by ALIGN-R and NONFIN to several candidate stress patterns.

/σσσσ/	NONFIN	ALIGN-R
→ a. σσσ(σ́)	1	
(→) b. σσ(σ́σ)	1	1
→ c. σσ(σ́)σ		1
(→) d. σ(σ́σ)σ		2
e. σ(σ́)σσ		2
f. (σ́σ)σσ		3
g. (σ́)σσσ		3

In Table 4.8, candidates d-g violate ALIGN-R too many times and are harmonically bounded by candidate c, while candidate b violates both NONFIN and ALIGN-R and is harmonically bounded by both a and c. However, candidates a and c, each with a single-syllable foot, will only satisfy FOOTBINARITY when the foot contains a single heavy syllable. In words with light penultimate and final syllables, candidates b and d will not be harmonically bounded, since they satisfy FOOTBINARITY.

In Hayes (1995), English is analyzed with moraic trochees, which can consist of a single heavy syllable (H́) or two light syllables (ĹL) but not of a single light (*Ĺ) or a heavy plus a light (*ĹH, *H́L). FOOT-BINARITY penalizes these disallowed foot

shapes. Kager (1999) discusses several examples of other stress patterns for which such a constraint is necessary.

(4) FOOT-BINARITY (FTBIN): Assign a violation to any foot which does not contain exactly two moras (which is not binary).

If this constraint is high enough ranked, parses such as a, c, and e in Table 4.8 would only be allowed if the syllable bearing the foot is heavy, whereas parses like b and d would only be allowed if the two syllables in the foot were both light. Words of shape LHL could only be parsed with a foot on the penultimate syllable: L(H́)L, *(Ĺ)HL, *(ĹH)L. Words with only light syllables, such as those in the experiment, could only have penultimate stress with a right-aligned (ĹL) trochee, as in b, or antepenultimate stress, as in d. Penultimate stress would be preferred when ALIGN-R ≫ NONFIN, and antepenultimate stress would be preferred when NONFIN ≫ ALIGN-R.

When the final vowel is [ə], participants follow the lexicon in producing equal percentages of antepenultimate stress and penultimate stress. When the final vowel is -i, antepenultimate stress is preferred. This situation can be modeled with an extra constraint, NONFIN-i, which assigns a violation just in case a final syllable whose nucleus is [i] is parsed into a foot.

(5) FOOT-NONFINALITY-i (NONFIN-i): Assign a violation if a word-final [i] is parsed into a foot.

This constraint has the same effect as a rule marking certain types of final syllables (in this case those with an [i] nucleus) as extrametrical (Hayes, 1982). Hayes and also Chomsky and Halle (1968); Liberman and Prince (1977) noticed that words ending in [i], [l̩], and [ɹ̩] all tend to behave as if the final syllable is extrametrical. An analysis including [l̩] and [ɹ̩] in the group of segments that prefer to be extrametrical could either use three separate constraints, e.g. NONFIN-l̩, or a single constraint referring to the three segments as a class: NONFIN-[i/l̩/ɹ̩].

Figure 3.12 shows an OT grammar which predicts that ə-final words should vary between antepenultimate and penultimate stress, and that i-final words should take antepenultimate stress only. In this grammar, FtBin is high-ranked, ruling out candidates with a single light syllable as a foot. Align-R and NonFin are unranked with respect to each other, but are both outranked by NonFin-i. For i-final inputs, only antepenultimate stress is allowed, but for inputs with other final vowels, antepenultimate and penultimate stress are equally grammatical.

Figure 3.12. OT grammar predicting that ə-final words should vary between antepenultimate and penultimate stress, while i-final words should only take antepenultimate stress. Only inputs with light penultimate syllables are considered here. 'X' indicates either a light or a heavy initial syllable, while 'L' indicates a light penult.

	FtBin	NonFin-i	Align-R	NonFin
/XLə/				
→ (XL)ə			**	
X(L)ə	*!		*	
→ X(Lə)			*	*
XL(ə)	*!			*
/XLi/				
→ (XL)i			**	
X(L)i	*!		*	
X(Li)		*!	*	*
XL(i)	*!	*		*

Participants' behavior in the experiment was stochastic - they produced both penultimate and antepenultimate stress on nonwords. There was a slight preference for antepenultimate stress in ə-final items, and a strong but still probabilistic preference for antepenultimate stress in i-final items. Based on the results of the analogical base task, I argue against a model in which the distinction between final [i] and final [ə] obtains because participants analogize to particular real words when choosing nonword stresses. I argue that instead, participants' phonological grammar must be able to represent the i-final trend probabilistically.

In Maximum Entropy grammar (Goldwater and Johnson, 2003), interacting constraints predict a probability distribution over candidate outputs. Constraints are not strictly ranked but are assigned weights. The negative dot product of those weights with the violation profiles of each candidate is the harmony score ($\mathcal{H}$) of the candidate. A probability distribution can then be calculated over those candidates using the exponential of $\mathcal{H}$, regularized over all candidates for a given input. Figure 3.13 shows a set of weights over the four constraints in Figure 3.12 which predicts a 50-50 distribution over penultimate and antepenultimate stress for ə-final inputs with a light penultimate syllable, and a 70-30 distribution in favor of antepenultimate stress for i-final inputs with a light penultimate syllable.

Figure 3.13. MaxEnt grammar predicting a 50-50 distribution over antepenultimate and penultimate stress for ə-final words, and an approximately 70-30 distribution for i-final word, favoring antepenultimate stress. $\mathcal{H}$ indicates harmony scores for each candidate, while P indicates the model's predicted probability for that candidate. Only inputs with light penultimate syllables are considered. 'X' indicates either a light or a heavy initial syllable, while 'L' indicates a light penult. Weights are listed immediately beneath constraint names, and are fitted by hand. Note that probabilities listed as zero are actually nonzero but very very small (on the order of 10^{-5}).

	P	$\mathcal{H}$	FTBIN	NONFIN-i	ALIGN-R	NONFIN
			10	1	1	1
/XLə/						
→ (XL)ə	0.5	-2			2	
X(L)ə	0	-11	1		1	
→ X(Lə)	0.5	-2			1	1
XL(ə)	0	-11	1			1
/XLi/						
→ (XL)i	0.73	-2			2	
X(L)i	0	-11	1		1	
→ X(Li)	0.27	-3		1	1	1
XL(i)	0	-12	1	1		1

In this tableau, FTBIN has a high weight, so that candidates with a foot consisting of a single light syllable are assigned a vanishingly small probability. The other

constraints all have equal weight, but candidates with a final [i] parsed into a foot incur a violation of both NonFin and NonFin-i, leading them to have a lower harmony score than candidates which have a final ə parsed into a foot. If the weight on NonFin-i were lower, the distribution over i-final candidates would be more similar to the distribution over ə-final candidates (if its weight were zero, the two distributions would be equivalent).

If NonFin were given a weight higher than the weight of Align-R, then the model would predict a higher probability of antepenultimate stress than penultimate stress for both types of input. The greater the difference between the two weights, the greater the preference for antepenultimate stress. Likewise, if Align-R had a higher weight than NonFin, the model would predict a higher probability on penultimately stressed candidates.

Each actual word of English has a single correct stress pattern, and does not vary: *cánnery* is always pronounced with antepenultimate stress, and *canáry* is always pronounced with penultimate stress, despite the fact that they both have a phonological shape for which the grammar predicts 70% antepenultimate stress. In a MaxEnt framework, a high weighted faithfulness constraint can ensure that all real words are pronounced correctly. One must assume that all words in the English lexicon are underlyingly specified for their stress pattern. Only nonwords with no prespecified stress pattern are subject to within-word variation. Figure 3.14 includes a high-weighted faithfulness constraint, and illustrates the behavior of a real word input, specified for an exceptional stress pattern compared to the behavior of a nonword input in which stress is not specified.

In typical language use, inputs with no specified stress pattern are the exception - the vast majority of words uttered are from the lexicon and are specified for their stress pattern. The most important fact for a learner of English to acquire is that Faith-Stress is high weighted, and that individual stress patterns must be stored.

Figure 3.14. MaxEnt grammar with high weighted faithfulness, illustrating the different behavior of an input from the lexicon, whose stress pattern is pre-specified, and a nonword input without a specified stress pattern. Only inputs with no pre-specified stress pattern have variable outputs. FAITH-STRESS ensures that when there is stress in the input, it is preserved. Weights are listed immediately beneath constraint names, and are fitted by hand. Note that probabilities listed as zero are actually nonzero but very very small (on the order of 10^{-5}).

	P	$\mathcal{H}$	FAITH-STRESS 50	FTBIN 10	NONFIN-i 1	ALIGN-R 1	NONFIN 1
/canáry/							
(cana)ry	0	-52	1			2	
ca(na)ry	0	-11		1		1	
→ ca(nary)	1	-4				1	1
cana(ry)	0	-63	1	1			1
[bæmæki]							
→ (bæmə)ki	0.73	-2				2	
bə(mæ)ki	0	-11		1		1	
→ bə(mæki)	0.27	-3			1	1	1
bəmə(ki)	0	-12		1	1		1

Given this, one might wonder what is the mechanism by which the markedness constraints and their relative weightings are acquired. Zuraw (2000, 2010) examines a similar case, in which each actual lexical item has only one correct pronunciation, but 'subterranean' constraints mirror the lexical statistics, and predict participants' choices on nonwords. Zuraw uses the Gradual Learning Algorithm (GLA) (Boersma and Hayes, 2001), a learning mechanism which updates the weights of all constraints with each learning datum. She shows that because of this indiscriminate updating, relative weightings of markedness constraints are automatically learned along with the high weight of faithfulness.

3.7   Conclusion

Models of phonological grammatical knowledge such as Maximum Entropy grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008; Hayes et al., 2009)

represent the phonological grammar as a series of probability distributions over possible outcomes. A MaxEnt representation of a preference for antepenultimate stress would predict antepenultimate stress on a given output a majority of the time, but not 100% of the time. Whether that majority is 60%, 80%, or 99.9% is determined by the weights on the constraints involved. Such a system is capable of representing the probability matching behavior found in many wug-tests of probabilistic phonological patterns (e.g. Becker et al., 2011; Hayes et al., 2009; Ernestus and Baayen, 2003; Zuraw, 2000, 2010).

In Experiment 3, I examined an alternative explanation for such probability matching behavior, namely that participants match lexical probabilities because they are directly referring to the contents of the lexicon when they produce novel words. I assume that lexical access is a process that is automatically applied to nonwords as well as real words. Spreading activation models of lexical access such as TRACE (McClelland and Elman, 1986) predict that when the lexical access process is applied to nonwords, some real words should become activated more than others, specifically the real words which are most like the nonword. In section 4, I sketched out how the task of producing a novel word would proceed in this case. The participant would perceive a nonword stimulus, in this case one with ambiguous stress, and certain lexical item or items would be activated. The stress patterns of those lexical items would then be primed when the participant selected a stress pattern for the nonword, causing the nonword to be most likely to have the stress of the most activated lexical item. A pattern that is found frequently in the lexicon (say, antepenultimate stress on i-final words) would show up in participants' productions frequently because the lexical access process would find words which obey the pattern more frequently than words which do not.

In Experiment 3, participants produced stress patterns on novel words, but they also gave similar real words for each novel word. If the process sketched in section 2

were responsible for participants' matching of the lexical trends, then the similar real words ('analogical bases') given by participants for each nonword item should match the stress pattern chosen for that item. No relationship was found between the stress pattern of the analogical bases and the stress patterns chosen in production, although both followed the lexical trends.

CHAPTER 4

THE LATIN STRESS RULE: A CASE OF UNDERMATCHING LEXICAL STATISTICS

4.1 Introduction

In productivity experiments, speakers often 'frequency match' a trend in the lexicon on novel words (Hayes et al., 2009; Ernestus and Baayen, 2003; Zuraw, 2000), producing a distribution of responses which matches the distribution in the language's lexicon. An exact match of the lexical frequencies can be achieved through a wide variety of possible mechanisms, as Ernestus and Baayen demonstrate. Rather, it is often mis-matches with the lexical statistics which can best shed light on the phonological acquisition and representation system (Becker et al., 2011; Becker et al.; Griner, 2001; Zhang et al., 2006). In this chapter I examine a case of such mismatch, in which a pattern that is nearly categorical in the lexicon is reproduced on nonce words as a (undermatched) probabilistic trend.

The pattern to be examined here is the 'Latin Stress Rule' for English stress (Chomsky and Halle, 1968, et seq.), which states that words (longer than two syllables) with a heavy penultimate syllable receive penultimate main stress. This generalization is nearly unviolated in English, with just a few unambiguous exceptions ('character', 'adjective', 'galaxy'). In an experiment, this trend will be examined together with the trend for i-final words to take antepenultimate stress which was the subject of the previous two chapters. There is very little evidence in the lexicon for which trend should take precedence: examples of the crucial case, i-final words with heavy penults, are rare. These two trends are incorporated here into a single

productivity experiment, in which participants produce stress patterns for all four types of words: i-final, heavy; i-final, light; ə-final, heavy; and ə-final, light.

4.2   The Latin Stress Rule

Stress in long words of English (longer than two syllables) has been described by the 'Latin Stress Rule' (Chomsky and Halle, 1968; Hayes, 1980; Halle and Vergnaud, 1987; Burzio, 1994):

Latin Stress Rule for English: If a word's penultimate syllable is heavy, then that word receives penultimate main stress. If the penultimate syllable is light, then the word receives antepenultimate main stress.

The second clause of this rule is not actually robustly obeyed in the English lexicon - exceptions abound. Pater (1994) examines the number of observers (ex: récipe) and violators (ex: banána) in the English lexicon, and evidence about native speakers' judgments of both types of words, arguing that neither one can properly be called the exceptional pattern. As discussed in the previous two chapters, words in the English lexicon (here, the CMU pronouncing dictionary (Weide, 1994)) with a light penult take antepenultimate stress about 75% of the time, but this percentage is modulated by the word's final vowel. Words with a final [i] take antepenultimate stress almost 90% of the time, but words with a final [ə] are divided 50-50. In addition, speakers prefer antepenultimate over penultimate stress in [i]-final nonwords, but do not exhibit a preference in [ə]-final nonwords.

The first clause of the Latin Stress Rule is more statistically robust in the lexicon of English, but it is also more complex (Olejarczuk, 2014): it requires speakers to form a generalization that heavy syllables attract stress, but only when they are in penultimate position. Learners of this generalization must first learn what counts as a heavy syllable in English (what counts as a heavy syllable is not universal - see Garrett, 1999 for an overview), and then must learn that weight only matters

in certain privileged positions - in this case the penultimate syllable of the word. Heavy final syllables also attract stress in English, but what counts as heavy in the final syllable may be different from what counts as heavy in the penultimate syllable (Burzio, 1994, pg. 64). This difference between final weight and penultimate weight will be returned to in the discussion section.

Chomsky and Halle (1968) and subsequent work on English stress defines 'heavy' penultimate syllables as syllables with either a long vowel, or at least one coda consonant. In the search of the CMU pronouncing dictionary presented below, the monophthongs [ɑ, æ, ə, ɔ, ɛ, ɪ, ʊ, i, u] as well as the syllabic consonants [l̩, ɹ̩, n̩, m̩, ŋ̩] are counted as short vowels, while the high tense vowels and all diphthongs [e͡i, a͡i, o͡ʊ, a͡ʊ, o͡i] were counted as long.

According to the Maximal Onset Principle (Kahn, 1976), consonants between two vowels should be syllabified in such a way that as many consonants as possible are assigned to an onset rather than a coda. For example, in a word like 'tapestry', with three consonants between the final two vowels ([tæpəs̲t̲ɹ̲i]), all three consonants would be assigned to the onset of the final syllable, and the penultimate syllable would be coda-less, and therefore light: [tæ.pə.s̲t̲ɹ̲i]. Any string of consonants which can legally begin a word in English can begin a syllable (for example, [stɹ] begins the word 'string'), but a sequence of consonants which cannot begin a word must be broken up across the coda and the onset. For example, in a word like 'galaxy' ([gælək̲s̲i]), the [k] must belong to the coda of the penultimate syllable, since [ks] cannot begin a word in English ([gæ.lək.si]).

If syllabification follows the Maximal Onset Principle and is independent of stress assignment, then words like 'tapestry' would have a light penultimate syllable which would not attract stress under the Latin Stress Rule. However, Kahn proposes that the Maximal Onset Principle can be violated in English in the case of stressed syllables - stressed syllables prefer to be heavy, and therefore attract consonants to their coda

position. For example, a word like 'digestive', would be syllabified as [dài.ʤés.təv], rather than [dài.ʤé.stəv] - the [s] is parsed as a coda of the stressed syllable which otherwise would be light. Given this possibility, it is unclear how to characterize words like 'tapestry' - because the penultimate syllable is unstressed, it is light, but if it were to be stressed then it would be heavy.

Table 4.1. Example words exhibiting different kinds of penultimate syllable weight

Weight	Penult syllable	Antepenult stress		Penult stress	
H	CVV	exponent	[ék.spòʊ.nənt]	aroma	[ə.ɹóʊ.mə]
	CVC	galaxy	[gǽ.lək.si]	elixir	[ə.lík.sɹ̩]
H/L	CV.CC	tapestry	[tǽ.pə.stɹi]	digestive	[dài.ʤé.stəv]
L	CV	radio	[ɹéi.ri.òʊ]	bikini	[bɨ.kí.ni]
	CV	cinema	[sí.nə.mə]	banana	[bə.nǽ.nə]

A search of the CMU pronouncing dictionary revealed that the trend in English for heavy penultimate syllables to take main stress is almost never violated, regardless of what type of heavy syllable is considered. Words with a diphthong in the penultimate syllable take penultimate main stress 97% of the time, while words with a monophthong in the penultimate syllable take penultimate stress only 32% of the time (Figure 4.1). Likewise, words with a closed penultimate syllable (regardless of vowel quality) take penultimate main stress 95% of the time, while words with an open penultimate syllable take penultimate main stress 35% of the time (Figure 4.2).

Because words like 'tapestry' and 'digestive' have a flexible syllabification, it is unclear whether they should be counted as exceptions to the Latin Stress rule when they are antepenultimately stressed. In order to address this question, Figure 4.3 compares words whose penult is otherwise light (has no coda, and whose vowel is a monophthong) but whose final syllables have a complex onset. In fact, words whose penult is 'ambiguously heavy' take penultimate stress significantly more often than words whose penult cannot be heavy even if it were stressed (χ^2=118.47, p< 1x10^{-15}). Words in this category do not behave like words with a closed penult or with a long

Figure 4.1. The effect of vowel length of the penultimate syllable on main stress placement. All words three syllables long and longer, both morphologically complex and morphologically simple words. $\chi^2=1255.02$

total: 5926
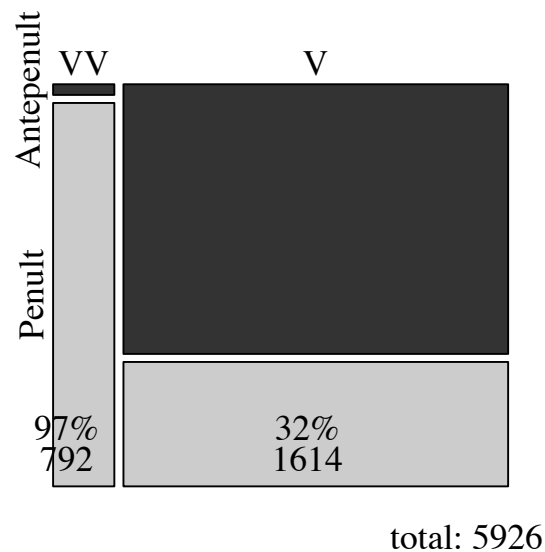
vowel in the penult - they only take penultimate stress 48% of the time. However, from the perspective of the learner trying to match these lexical statistics, 'tapestry' words could still be counted as potential heavy penult words.

If all syllables with either a long vowel or a coda consonant are counted as heavy, then 97% of all words with heavy penults take penultimate stress, and 32% of words with light penults take penultimate stress. If 'tapestry' words are included in the heavy-penult category, then 87% of words with a heavy penult take penultimate stress, and 22% of words with a light penult take penultimate stress. If words with a complex onset in the final syllable are counted as heavy penult words when the penult is stressed, and as light-penult words when the penult is unstressed, the most extreme distribution obtains: 97% of heavy penult words have penultimate stress and 21% of light penultimate words have penultimate stress.

Figure 4.2. The effect of a closed penultimate syllable on main stress placement. Only consonants which cannot legally be syllabified as onset to the final vowel are counted as coda consonants. All words three syllables long and longer, both morphologically simple and complex are included. χ^2=745.99

total: 5910
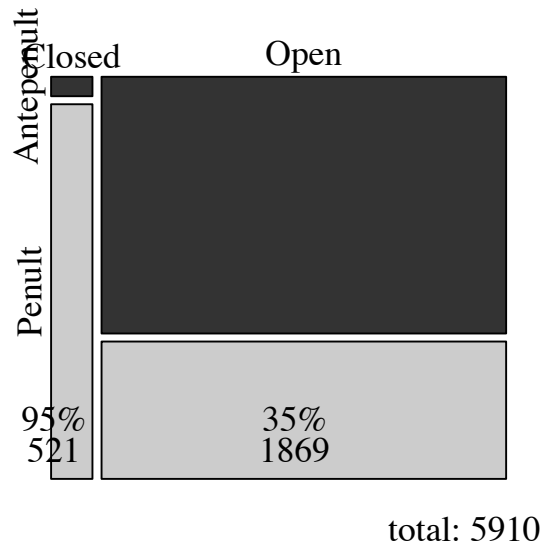
In chapters 2 and 3, words with only light penultimate syllables were examined. Within those words, the final vowel affected words' likelihood of taking antepenultimate vs. penultimate stress. Words ending in [i] took antepenultimate stress the majority of the time, while words ending the [ə] took antepenultimate stress about half the time. In words with heavy penultimate syllables, there is still a numerical trend for i-final words to take antepenultimate stress more often than ə-final words. Because there are so few i-final words with a heavy penult a chi-square value cannot be reliably calculated. Figure 4.4 illustrates the numerical trends.

Experiment 4 tests English speakers' ability to productively apply the Latin Stress rule to nonwords. The generalization that heavy penults are stressed is strongly observed in the lexicon of English - there are a mere 49 exceptions (out of 1368 words with heavy penults) in the entire corpus observed here. The generalization tested in chapters 2 and 3 that i-final words take antepenultimate stress is also strongly

Figure 4.3. The effect of final onset complexity on stress placement. Only words with an otherwise light penultimate syllable (codaless, monophthongal vowel) are included. Both morphologically simple and complex words three syllables long and longer are included $\chi^2=118.47$
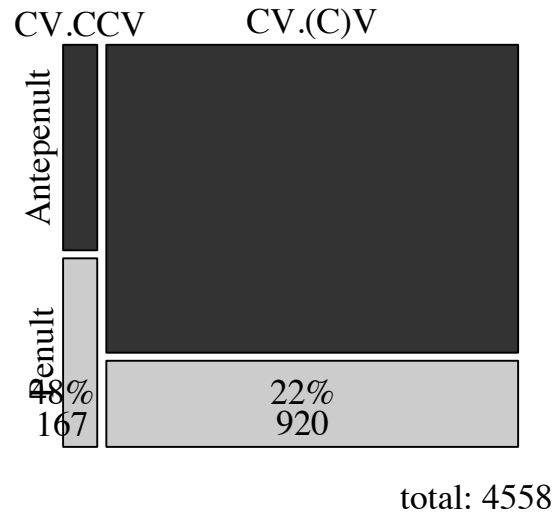
Figure 4.4. Effects of the final vowel, broken down by weight of the penult: i-final words are more likely to take antepenultimate stress in words with heavy penults as well as in words with light penults. All words three syllables long and longer are included
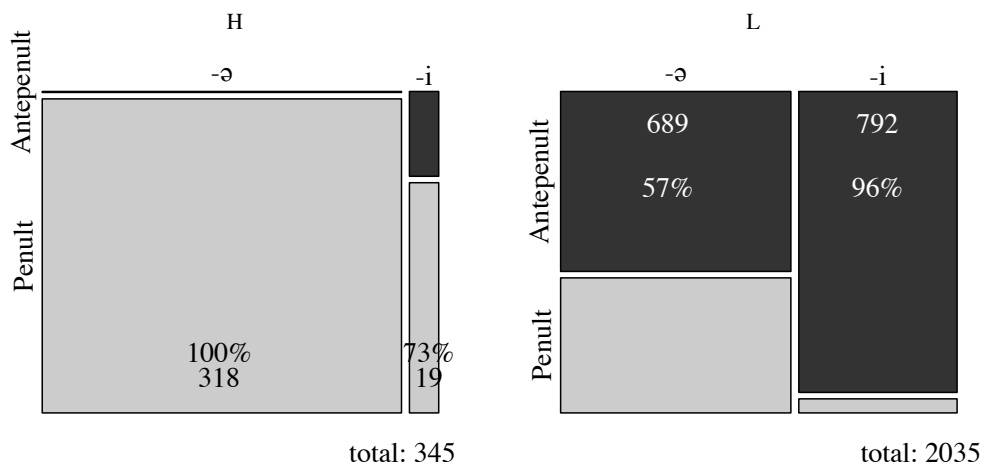
observed in the lexicon, (56 exceptions out of 855 i-final words), but fewer forms in the lexicon contribute to the generalization - about 2/3 as many. Speakers' knowledge of the two trends is compared in Experiment 4 - both the weight of the penultimate syllable and the quality of the final vowel are manipulated.

If speakers match more precisely probability distributions for which they have more evidence (more forms), then participants should match the trend for heavy penult syllables to take penultimate stress better than they match the trend for i-final words to take antepenultimate stress. Since both trends are instantiated over a large number of forms, it may be that speakers match both well, but show a preference for one over the other when they are in conflict, specifically in i-final words with heavy penultimate syllables.

4.3 Methods: Experiment 4

Experiment 4 was similar in methodology to the first part of Experiment 3 in chapter 3. In this case, items' final vowel was manipulated (half were i, half ə) just as in Experiment 3, but additionally the weight of items' penultimate syllable was manipulated. This experiment also tests for a difference in preferred stress pattern based on a word's part of speech. Studies such as Guion et al. (2003); Domahs et al. (2014); Kelly and Bock (1988) found that speakers prefer initial stress more strongly on two-syllable nouns than on two-syllable verbs. Sonderegger and Niyogi (2013) find evidence for the same pressure in historical records of changes in stress patterns on noun-verb pairs.

In the English lexicon, nouns and verbs longer than two syllables also differ in their stress preferences.

Verbs do take slightly more final stress than nouns, even in longer words, but the overall numbers of final stress are low. The main difference between nouns and verbs is that the preference for antepenultimate stress over penultimate stress is stronger

Figure 4.5. Counts of nouns and verbs longer than two syllables. This graphic includes both multimorphemic and monomorphemic words, but leaves out a small number of words with pre-antepenultimate stress (around 300, mostly nouns)

Part of Speech
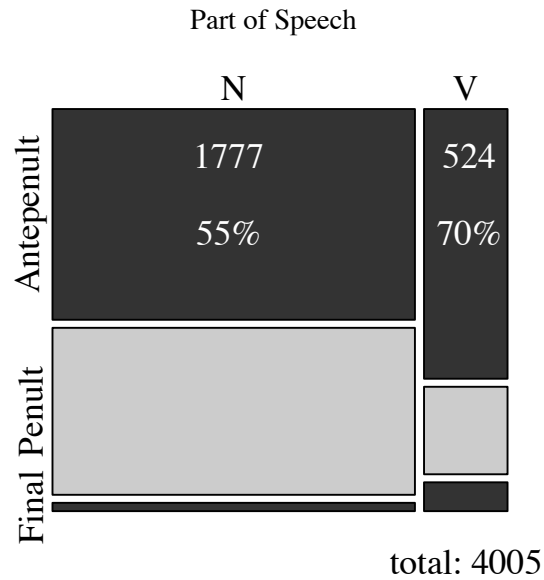
for the verbs. As will be discussed below, the noun-verb manipulation did not affect participants' choices of stress, likely because the presentation of the noun/verb context was problematic.

4.3.1 Participants

Like Experiment 3, Experiment 4 was presented via the world wide web, and participants were recruited through Amazon Mechanical Turk. All participants were over 18 years of age, and had an age range of 18 to 75 (mean age: 34). Only IP addresses originating in the United States were accepted. Participants were asked where they were from, and "when you speak English, where do people think you are from?" If their answer to the second question was a location within the united states, they were assumed to be a native speaker of American English. Participants were paid at a rate of $0.91 for the experiment, which took about 15 minutes. Data was

collected from a total of 100 participants, and data from 42 participants was used.The remaining participants were excluded because of problems with the sound recording and native speaker status. The process of excluding participants is described in detail in the results section.

4.3.2   Items

Items were derived from the items used in Experiment 3 in chapter 3. The nonword items from Experiment 3, all three syllables long and consisting of only light syllables, constituted the light-penult condition of the experiment. From each of these, a heavy-penult item was constructed by adding a coda to the penultimate syllable. Codas were chosen so that they did not form a legal onset cluster with the onset of the following syllable.

Table 4.2. Example item in four conditions. Conditions were counterbalanced so that participants saw each item in only one condition, but saw an equal number of instances of each condition.

	Penult Weight	
Final Vowel	Light	Heavy
-ə	[pǽ] [lǽ] [kə]	[pǽ] [lǽz] [kə]
-i	[pǽ] [lǽ] [ki]	[pǽ] [lǽz] [ki]

Each item consisted of a written frame sentence, followed by three auditorily presented individual syllables, and two auditorily presented versions of the full word, with different stress patterns (Antepenultimate and Penultimate). Participants first heard the syllables, then pronounced the word, then heard the two stress option, and chose between them. An example item is shown below. Participants were asked in the instructions to pronounce the entire sentence fluently, but a minority of participants did so: only 14 out of 100.

I got rid of the _____ yesterday

🔊 [bæ] [mæ] [ki]

_____

RECORD...

_____

What did you say?

🔊 [bǽməki]   🔊 [bəmǽki]

☐   ☐

As in Experiment 3, real words were also included in order to encourage participants to make their productions as fluent and like real words as possible. The real words used were the same as in Experiment 3. All had light penults, and were evenly split among the four word shapes used in the ERP experiments (-i final antepenultimate stress, -i final penultimate stress, -əfinal antepenultimate stress, -əfinal penultimate stress).

Items were 80% novel words and 20% actual words. When they were actual words, the two stress choices were (1) the actual word, and (2) a mis-stressed version of the actual word, e.g. [kǽnədə] and [kənǽdə]. All items (words and nonwords) had the same stressed vowel in each stress version, and in all cases, the first two of the three individual syllables in an item had the same stressed vowel.

Each stress version of each item was transcribed in the international phonetic alphabet (IPA) and pronounced in a random order by a female native speaker of American English, in the frame sentence "Say X again." Isolated syllables were constructed (as in Experiment 3) by first recording a different female native speaker of American English speaking each individual syllable as it was written in IPA. These recordings were resynthesized in Praat (Boersma and Weenink, 2011) so that each vowel was approximately 300ms long. Syllables with codas were manipulated so that the coda consonant was clearly identifiable. In the majority of cases, the consonant's amplitude was boosted, and in some cases a version of that final consonant from a different recording was spliced in. Each syllable was also resynthesized to have an

Table 4.3. Pitch contour for individual syllables presented to participants. All syllables were resynthesized to have this contour
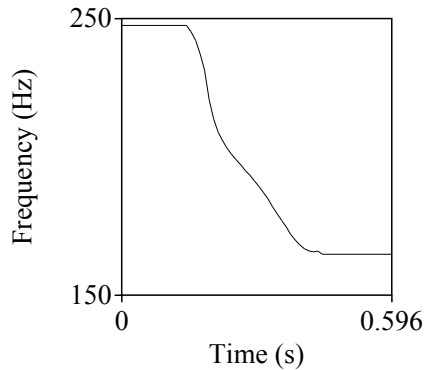
identical pitch contour (a H* pitch accent followed by a H-H% boundary tone), and so that each syllable would be perceived as an independent prosodic word. The pitch contour used was identical to that used in Experiment 3, but it is illustrated here for convenience. Intensity was also normalized across syllables using Praat.

Items were counterbalanced, so that each participant saw equal numbers of each condition, but did not see the same item in multiple conditions. Eight lists were constructed and each participant was randomly assigned a list. Each list contained 8 items in each of the four Penult Weight x Final Vowel condition (H, i-final; H, ə-final; L, i-final; L, ə-final). Noun and Verb frames were randomly assigned to each list so that each final vowel had an equal number of noun and verb frames. Noun and Verb frames were not balanced across Penult Weight conditions. Each item occurred in each of the eight Penult Weight x Final Vowel x Part of Speech condition on exactly one list.

4.3.3 Procedure

The experiment was presented via the world wide web. When participants arrived at the site, they were first asked to electronically sign a consent form, and then they completed a sound check to test that their microphone and speakers were working.

Next, they were instructed that they would see a written frame sentence and hear a sequence of three syllables, and that they should speak the whole sentence fluently, as if the three syllables constituted a word in the sentence. They were given an example nonword sequence of syllables and two examples of those syllables strung together into a pseudoword - once with initial stress and once with penultimate stress. Next, they were given a sample trial which was a real word (they were told in advance that it would be a real word). In each trial of the experiment, they saw a written frame sentence, and first heard the three syllables, then were asked to speak the word fluently, then they listened to the two stress options for that item and clicked a radio button to choose one. There were 32 nonword items and 8 real word items.

4.4   Results

As in Experiment 3, participants' success at the task was assessed based on whether participants (a) produced the syllables fluently together as a single word, with a single main stress, and (b) whether their produced stress agreed with the stress they reported producing. The author listened to each recording and annotated the participant's produced stress for that item as well as whether the participant produced a coda consonant in the penultimate syllable or not. Stress pattern was assessed based on vowel reduction and pitch. If a production had a full vowel in every syllable, or both of the first two syllables, it was counted as 'incorrect'. Also, if the production contained a pitch fall on any syllable but the last, or pauses between the syllables, it was counted as 'incorrect'. A participant was excluded from analysis if more than three of the first 10 nonwords examined counted as 'incorrect'. Participants who did not successfully record any sound were also excluded. Because of a technical problem, a large number of participants - 34 in total, did not record any sound. An additional 6 participants were excluded because of the quality of the recordings, and

116

13 were excluded because they produced too many items as individual syllables rather than fluent words. A total of 38 participants remained after these exclusions.

These 38 participants were all greater than 75% accurate at reporting their produced stress. All agreed with the author's transcription of their produced stress at least 75% of the time. Participants did not always accurately produce a heavy penultimate syllable when one was present in the three-syllable prompt. On 21% of trials with codas on the penultimate syllable of the prompt, the participant left out the coda, producing a light penult instead. On 6% of trials with a light penult in the prompt, the participant produced a coda on the penult. Because of the relatively high level of mismatch between the prompts and participants' productions, the weight (transcribed by the author) of the actual production is used for analysis, rather than the weight of the penultimate syllable of the prompt. Of the 38 participants whose data was analyzed, only 10 recorded full sentences, speaking the item in context using the written prompt. The other 28 recorded only isolated words.

4.4.1 Part of Speech

The part of speech of each item was determined by the written context sentence displayed before each item (Example noun context: "I've been waiting for a ____for months." Example verb context: "I spent all weekend trying to ____."). The part of speech demanded by the context sentence did not affect participants' choice of stress pattern, as can be seen in Figure 4.6.

One possible explanation for this lack of effect is that the presentation of the context sentences was problematic. Participants saw a written context sentence, but the rest of the trial was auditory only. Although participants were instructed to speak the whole sentence, with the word in the blank, the majority of participants did not do this, and could possibly have completely ignored the written context in choosing the item's stress pattern.

117

Figure 4.6. Participants' choices of stress in items with different part-of-speech contexts. No difference is observed between noun and verb contexts.

Forced choice responses

	Noun	Verb
Antepenult		
Penult	54% 268	51% 252

total: 998
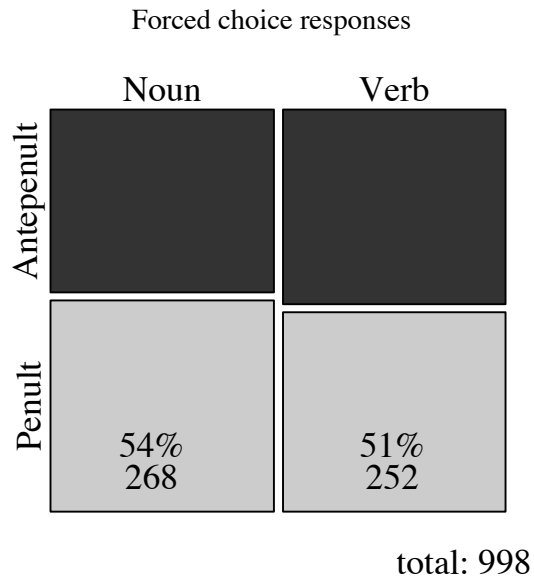
Figure 4.7 illustrates the numbers for just the subset of participants who did speak the context sentences aloud (10 participants). In just this subset of participants, a slight numerical difference between proportion of Antepenultimate stress obtains between Noun and Verb contexts, but that difference is not statistically significant. A mixed-effects logistic regression was run on just this subset of data, with part of speech as a fixed effects, and with random intercepts and slopes for both participants and items. There was no significant change in the likelihood of antepenultimate stress given a verb context compared to a noun context (β=-0.2, p=0.55).

It is possible that even in this subset of participants, the fact that the context sentence was visually presented while the nonword stimulus was auditorily presented could have led participants to make their decision about the nonword's stress without taking its part of speech into account. A better design would incorporate auditorily presented context sentences. However, a more likely explanation is that the difference

Figure 4.7. Participants' choices of stress in items with different part-of-speech contexts: Only the subset of participants who produced the context sentences aloud for each item. There is a small numerical difference in proportion of each stress pattern between Noun contexts and Verb contexts, but the difference is not statistically significant.

Forced choice responses

total: 254
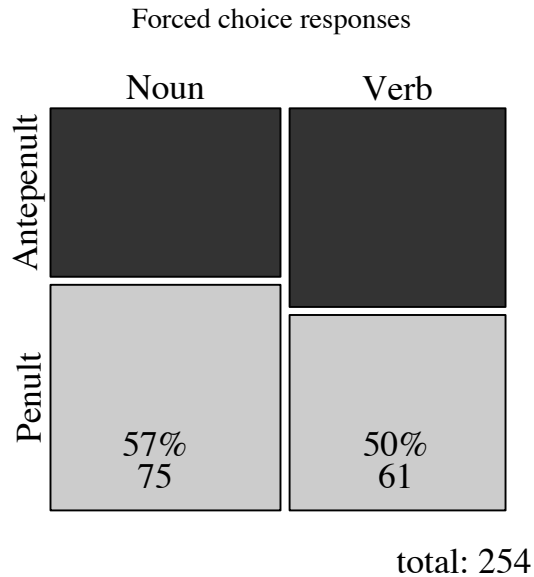
between longer nouns and verbs in the lexicon is not known by speakers. English speakers can reliably reproduce the difference between two syllable nouns and verbs in judgments tasks and in production (Guion et al., 2003), but the statistical trends in the lexicon are much less robust for longer words, since there are relatively few long verbs.

4.4.2 Weight and final vowel

Participants modulated their choice of stress pattern based on the weight of the penultimate syllable. They preferred penultimate stress when the penult was heavy, and antepenultimate stress when the penult was light. In the lexicon somewhere between 87% and 97% of words with a heavy penultimate syllable receive penultimate

119

main stress. Participants under-matched this distribution, producing penultimate stress on words with heavy penults 75% of the time.

Figure 4.8. Participants' choice of stress pattern for different penult weight conditions. Weight is based on the author's transcription of participants' actual productions rather than the weight of the penultimate syllable in the prompts. Because participants more often erred towards producing light penults than heavy ones, there are more responses overall in the light penult category.

Forced choice responses

total: 998
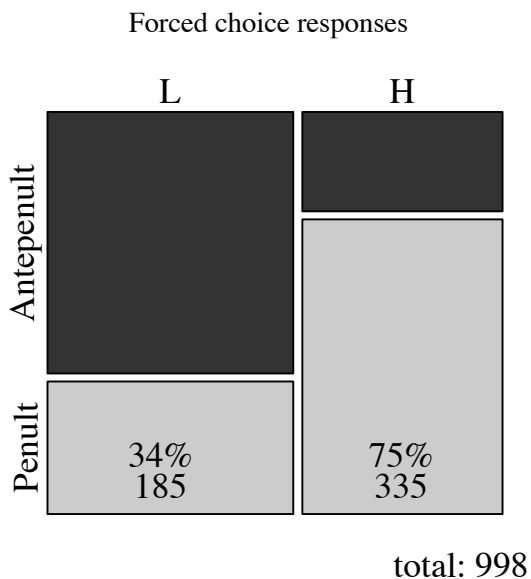
As in Experiment 3, participants produced antepenultimate stress more when the final vowel was [i] than when the final vowel was [ə]. This effect was independent of the weight of the penultimate syllable - both within heavy items and within light items, i-final items are produced with more antepenultimate stress than ə-final items.

A mixed-effects logistic regression was fitted with Final Vowel and Penult Weight and their interaction as fixed effects, and with random intercepts for participants and for items[1]. When the final vowel was [i], items were significantly less likely to receive penultimate stress than when the final vowel was [ə] (β=-0.96, p<6x10^{-6}). When the

---

[1]random slopes were initially included in the model as well, but the model did not converge

Figure 4.9. Participants' choice of stress pattern for different final vowels, broken down by weight category. Regardless of the weight of the penult, participants prefer antepenultimate stress more for i-final items than for ə-final items.
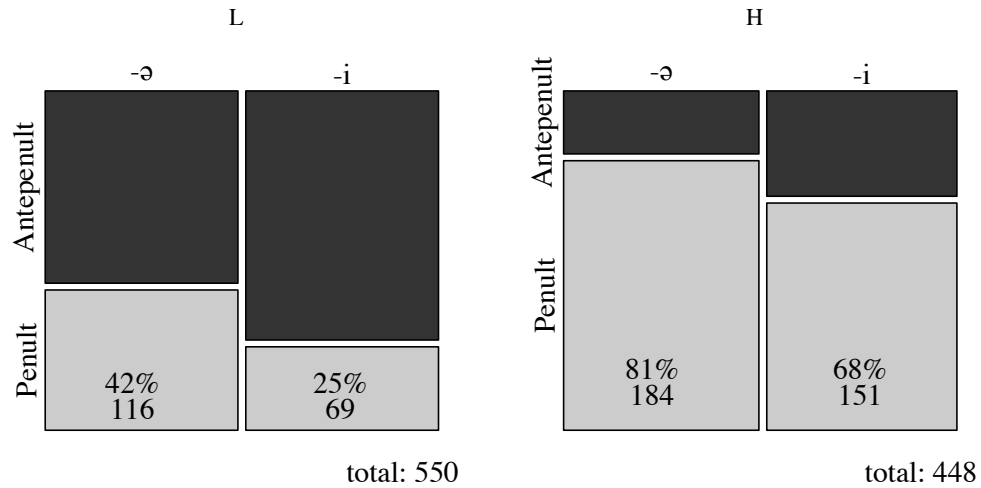
penultimate syllable was closed by a consonant, items were significantly more likely to receive penultimate stress than when the penultimate syllable was open (β=2.11, p<2x10^{-16}). The interaction between the two was not significant (β=0.09, p=0.77).

Table 4.4. Logistic regression with two factors. Produced stress as a function of penult weight and final vowel

Model: *Produced Stress ~ Final Vowel + Penult Weight*

	Estimate	p
Intercept	-0.34	0.119
Final Vowel = i	-0.96	5.26×10^{-6}
Penult = H	2.11	$< 2 \times 10^{-16}$
Penult = H X Final Vowel = i	0.09	0.78
AIC:	1119	

Participants are observing both the trend for i-final words to take antepenultimate stress and the trend for words with heavy penultimate syllables to take penultimate stress, and they are observing both probabilistically. The weight of the penultimate syllable has a stronger effect on stress placement than the final vowel (the magnitude of the coefficient on penult weight is about twice the magnitude of the coefficient on

final vowel), but a heavy penult does not completely trump a final [i]. In fact, the two effects are statistically independent. The quality of the final vowel matters equally in the heavy penult case as in the light penult case.

4.4.3 Individual participants' behavior

The probabilistic effects of final vowel and penult weight observed here could potentially arise from either within-speaker variation, or across-speaker variation. Analyzing the data in aggregate could lead to a misunderstanding about the behavior of individuals - in the most extreme case, speakers could behave probabilistically in aggregate but categorically as individuals. In the present experiment, it could be that some participants observe the Latin Stress Rule or the final vowel dependency categorically, and others do not observe one or the other of the generalizations at all in their responses. Because claims about how probabilistic trends are represented in the phonological grammar are claims about the cognitive architecture of individual speakers, misunderstanding the behavior of individual speakers could potentially lead to wrong conclusions about the cognitive architecture of each individual speaker.

Because relatively little data was collected from each participant (32 items, 8 in each weight x final vowel condition), no strong conclusions can be drawn about particular participants' behavior in this experiment. Instead, consider the histograms in Figure 4.10. Each individual's percentage of initial stress was calculated for each penult weight condition and for each final vowel. The histograms show the number of participants who fell into each percentage bin.

If the overall percentage of initial stress in each condition of the experiment arose from across-participant variation, in which participants were either obeying a trend categorically or not obeying it at all, then these distributions should be bimodal, with one mode close to 1 or 0 and another close to 50%. Instead, the means of each distribution are at or close to the mode of the distribution. Many participants

Figure 4.10. Individual participants' proportion of initial stress for a) each type of penultimate syllable, and b) each type of final vowel. Vertical lines correspond to overall average percentages in each condition: For heavy penult items, 32%; for light penult items 63%, for all i-final items 56%, and for all ə-final items 39%. 38 participants are included.

a)   b)
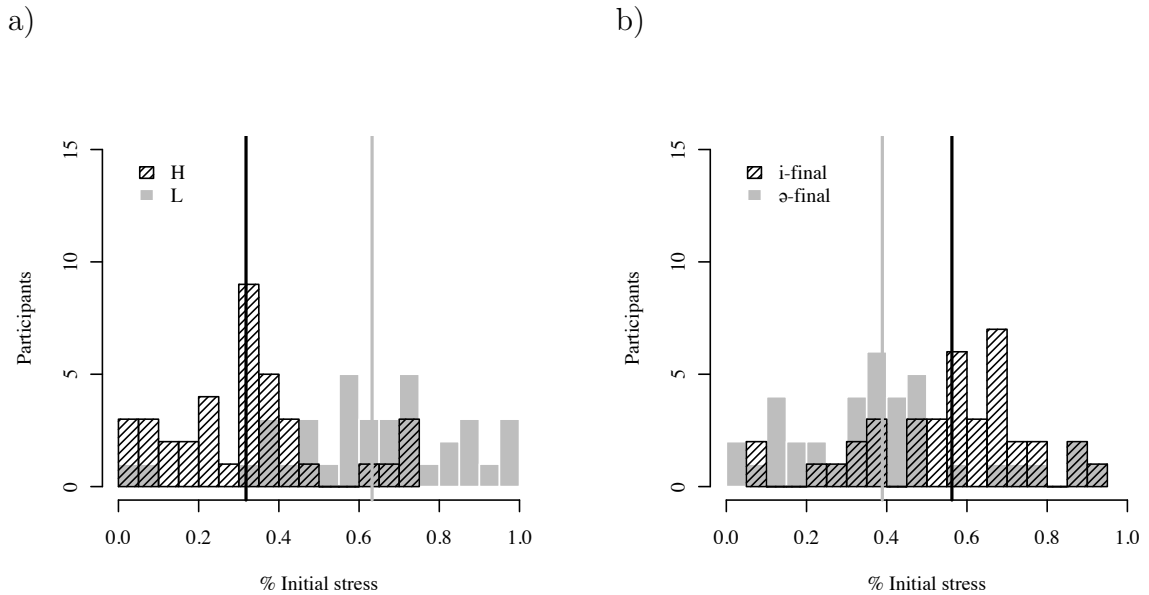
produced rates of initial stress far from 0, 1, or 0.5. Based on this it is unlikely that the observed variation is the result of participants behaving differently from each other, but always categorically. Rather, individuals are each behaving probabilistically.

4.4.4 Discussion

The Latin Stress Rule (if the penultimate syllable is heavy, the word takes penultimate main stress) is a generalization in the English stress system which figures centrally in many analyses (Chomsky and Halle, 1968; Halle and Vergnaud, 1987; Hayes, 1980; Burzio, 1994). The productivity of this generalization was tested here, and it was found that although participants do observe the generalization, they undermatch the lexicon, producing many exceptions. Olejarczuk (2014) found something similar in his wug-test of the Latin Stress Rule. In that experiment, participants produced

more penultimate stress on heavy-penult words than on light-penult words, but the mean proportion of penultimate stress on heavy-penult words was still very low - less than 50%. This low rate of penultimate stress found by Olejarczuk can be explained by the structure of the items: heavy-penult items contained sonorant-obstruent clusters between the final two vowels, which could only be syllabified as a coda plus onset (e.g. ma.dal.paz). However, in stressless syllables sonorant codas can be reduced to a syllabic sonorant so that the syllable no longer has a coda.

Domahs et al. (2014) also tested the Latin Stress Rule, finding that English speakers produced penultimate stress nearly 100% of the time for heavy-penult items (40% of the time for light-penult items). The methodology of Domahs et al. differs somewhat from the methodology of Experiment 4: stimuli were presented orthographically rather than auditorily, no i-final items were included, and several different structures of word (e.g. with antepenultimate heavy syllables) were included. Additionally, similarity to existing words was controlled by excluding items that rhymed with existing items, rather than via a holistic measure of neighborhood density. Further investigation into what effects these methodological differences might have on participants' behavior is in order.

In the next section, I present a Maximum Entropy model of the Latin Stress rule. This model undermatches the lexical statistics similarly to participants when a general weight-to-stress constraint is used. If a more specific constraint demanding main stress on specifically the penultimate syllable is used, closer match to the lexicon is obtained. I argue that participants in the experiment know only the general constraint and not the more specific constraint.

4.5   Modeling English stress with MaxEnt

The stress system of English has a long and rich history as the object of phonologists' interest. Notable comprehensive analyses have been given in Chomsky and

Halle (1968); Liberman and Prince (1977); Hayes (1980); Selkirk (1984); Halle and Vergnaud (1987); Kager (1989); Burzio (1994); Alcántara (1998). Throughout these works, a few major themes arise: Main stress nearly always falls on one of the final three syllables of a word, and within that three-syllable window both heavy penultimate syllables and heavy final syllables can attract main stress. The morphological structure of a word is important. Some affixes shift stress to a particular position within the word (e.g. '-ity' demands antepenultimate main stress, so that elástic ∼ elastícity; see Tescner and Whitley, 2004 for a comprehensive discussion of this), but the stress of a base is sometimes preserved in the derived form as a secondary stress (Pater, 2000; Collie, 2008; Zamma, 2012). A word's part of speech can also affect stress placement, so that nouns prefer penultimate stress or antepenultimate stress, while verbs prefer final stress (Kelly and Bock, 1988; Sonderegger and Niyogi, 2013; Sereno and Jongman, 1995). Stress clashes (two stressed syllables in a row) are rare in English. They are rare in words, and avoided in the production of nonwords (Kelly and Bock, 1988), and also can trigger stress retraction at the word and phrase level (Prince, 1983; Tilsen, 2012; Henrich et al., 2014 and many others).

Another property of English stress that is agreed upon throughout the above literature is that lexical exceptions abound. Some generalizations are more exceptionful than others: exceptions to the preference for nouns to take penultimate stress and verbs to take final stress are plentiful, while exceptions to the preference for words with heavy penultimate syllables to take penultimate main stress are vanishingly rare. However, all accounts of the system have used some mechanism for marking exceptions. Chomsky and Halle (1968) and later Burzio (1994) attribute different stress patterns to different underlying forms, for example the difference between the word 'cínema', and 'banána', both nouns with all light syllables, is that 'banána' is underlyingly /bananna/, with a geminate which does not get realized, but which does make the second syllable heavy. Other approaches use diacritics which determine

the underlying weight of a syllable (Hayes, 1980), or which specify some syllable as extrametrical on a word-by-word basis (Selkirk, 1984).

This section will focus on modeling the quantity-sensitivity in longer words of English: words with a light penultimate syllable weakly tend to take antepenultimate stress, and words with a heavy penultimate syllable strongly tend to take penultimate stress. Speakers' knowledge of this generalization was tested in Experiment 4, where participants chose a stress pattern for three syllable words with a light or heavy penult. In their productions, participants preferred penultimate stress on words with a heavy penultimate syllable, but they produced more antepenultimate stress than expected given the lexical statistics. This is illustrated in Figure 4.11. In the lexicon, 96% of words with heavy penultimate syllables had penultimate main stress, while in the experiment participants produced items with heavy penults with penultimate stress only 75% of the time.

Individual participants consistently undermatched the lexical distribution (though by varying amounts). Undermatching of this type has been argued to indicate that participants have a phonological grammar which does not viridically represent the lexical distribution, either because it cannot, or because something about the acquisition system has prevented speakers from acquiring the necessary grammar (Hayes et al., 2009; Becker et al., 2011; Moore-Cantwell, 2012). In this case, the preference for heavy penultimate syllables to take penultimate stress may be underlearned because speakers' phonological grammar can represent the preference for heavy syllables to be stressed, but it does not represent the specific preference for *penultimate* syllables to be stressed. The specific preference is more complex, referring to more features than the general preference. It also refers to two different kinds of features: weight, a property of the syllable structure and segmental content; and position in the word. Goldrick and Larson (2010); Moreton and Pater (2012) argue that more complex generalizations (in terms of number of features) are more difficult for learners to acquire.

Figure 4.11. Corpus counts compared to participants' choices in a wug-test. The corpus counts include only three-syllable words and longer words. Heavy penultimate syllables either contain a diphthong or are closed by at least one consonant. In the experiment, participants heard a three-syllable stimulus with ambiguous main stress and produced the nonword with either penultimate or antepenultimate main stress. Heavy penults were always closed by a consonant - no diphthongs were present in the experiment.
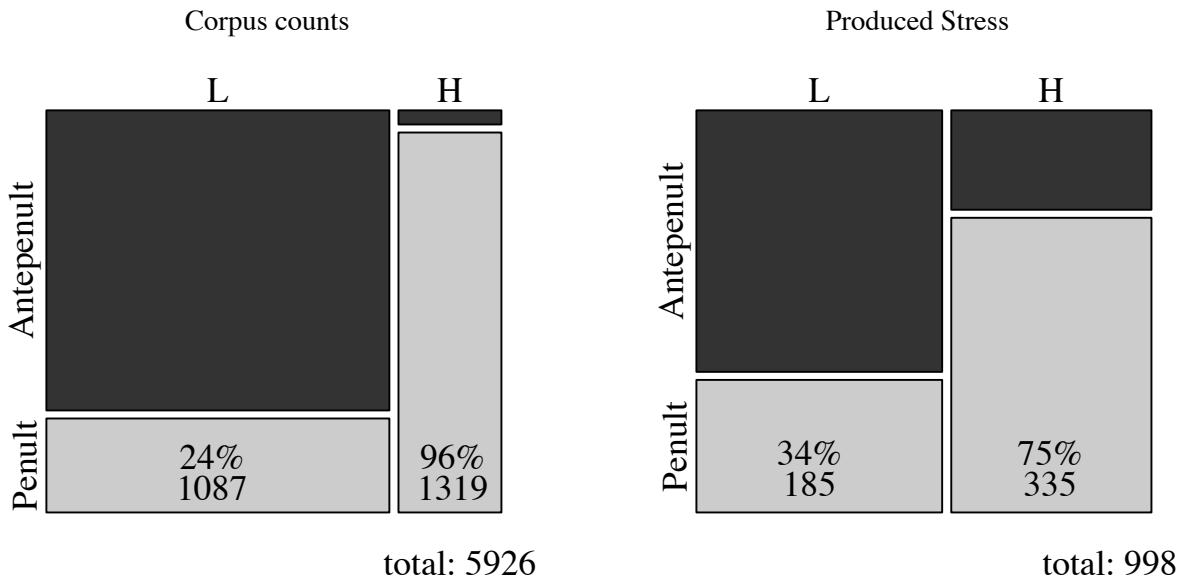
In the case of penult weight in English, the better match to the lexicon provided by the more complex generalization may not be worth cost for the learner of using a more complex constraint.

In what follows, I will first sketch a grammatical analysis of the basic system (heavy penult words strongly prefer penultimate stress, light penult words weakly prefer antepenultimate stress), using a probabilistic grammatical model (MaxEnt). The weights of the appropriate constraints will be fit to the lexical data. A system with a position-specific WEIGHT-TO-STRESS constraint, targeting only penultimate syllables, fits the lexical distribution very closely, and better than the same system with a position-general WEIGHT-TO-STRESS constraint. However, the system with the position-general constraint better predicts participants' behavior in the produc-

127

tion task. I argue that participants undermatch the lexical distribution in the lexicon because they have a grammatical system which incorporates a position-general WEIGHT-TO-STRESS constraint, rather than the position-specific version which would provide a more perfect fit to the lexicon.

4.5.1 Training data: the CMU pronouncing dictionary

The CMU pronouncing dictionary (Weide, 1994) has been used throughout this dissertation to find the lexical frequencies of the trends discussed. In this chapter, it will be used as the training data for fitting a MaxEnt model of the English stress system. The CMU pronouncing dictionary is a dictionary of American English lexical items, and contains about 134,000 entries, phonetically transcribed, and with each vowel annotated for primary, secondary, or no stress. The average adult vocabulary size is much smaller than 134,000 - about 9,000 to 17,000 words (Zechmeister et al., 1995). This means the pronouncing dictionary contains a great many entries that are low frequency enough as not to be present in most adult native speaker vocabularies. The CMU pronouncing dictionary also contains forms with inflectional morphology (both 'banana' and 'bananas' for example). English inflectional morphology does not affect a word's stress, so including these entries inflates the counts of particular word shapes (especially words with final clusters). In order to avoid hyper-low-frequency entries and entries with inflectional morphology, a 'cleaned-up' version of CMU was used, namely the input corpus for Hayes's phonotactic probability calculator (Hayes, 2012). This input file contains 18,034 entries, all of which are frequent enough to be in English CELEX (Baayen et al., 1993). This dataset also avoids entries with inflectional morphology, and certain transcription 'errors' are corrected, such as having multiple primary stresses on a single word.

A series of scripts was then used to annotate this lexicon further. Each word was first syllabified and annotated for its syllable structure (e.g. CVC, CVCC). The

maximal onset principle was used, so that clusters which are legal onsets of English were assumed to be onsets in every case[2]. The CMU transcription system does distinguish between syllabic [ɹ̩] and [ɹ] as a coda, but it does not represent syllabic l's and nasals, instead transcribing them as ə followed by a coda l or nasal. In order to prevent inflated counts of syllables with codas in stressless position, all schwa-l or schwa-nasal sequences in stressless position were assumed to be syllabic sonorants instead of syllables with codas, and were re-transcribed as such. Diphthongs [e͡ɪ, a͡ɪ, o͡ʊ, a͡ʊ, ɔ͡ɪ] were counted as long, while the non-diphthongs [ɑ, a, æ, ɛ, ɪ, i, ʊ, u, ɹ̩, m̩, n̩, ŋ̩, l̩] were counted as short.

Additionally, each entry was cross-referenced with English CELEX for part of speech information, and with SUBTLEX (Brysbaert and New, 2009) for frequency information. Spelling was used as a proxy for annotating derivational morphology through the process described in 4.5.1.1. This calculated information about each lexical item was then used to further annotate for a variety of factors, such as a word's main stress, weight of the penultimate syllable, weight of the final syllable, final vowel, contents of the word-final coda, etc. Only words with at least two syllables were included in the calculations. The total number of words included in each search of the lexicon was 11,765.

4.5.1.1 Morphology Matters

English stress assignment is conditioned by morphology in two ways. First, 'neutral' prefixes and suffixes can be added to a word without affecting its stress pattern, e.g prímitive ∼ prímitive**ness**; cálibrate ∼ **re**cálibrate. Second, many affixes enforce a specific main stress pattern. English has suffixes which enforce final stress, penultimate stress, antepenultimate stress, and even pre-antepenultimate stress (Tescner and Whitley, 2004). Also, as discussed by Chomsky and Halle (1968); Burzio (1994);

---

[2]Thanks are due to Robert Staubs for generously lending his syllabification algorithm

Table 4.5. Examples of stress-shifting affixes. See the appendix for a full list of affixes, which list was compiled based on the list in pronouncing English Chapter 2

Stress shifted to:		
ultima	exámine	exàmin**ée**
penultimate	abólish	àbolí**tion**
antepenultimate	sólid	solíd**if̀y**
pre-antepenultimate	antíque	ántiqu**àry**
no shift	compánion	compánion**able**

Pater (2000); Collie (2008) words derived via a stress-shifting affix can preserve the main stress of their base as a secondary stress, resulting in a stress pattern which is atypical for monomorphemic words.

In the corpus searches presented here, spelling was used as a proxy to detect derivational morphology. For example, words ending in 'tion' were considered to end in the '-tion' affix. The list of suffixes and prefixes in Pronouncing English, chapter 2, was taken to be exhaustive and words in the corpus with any of these strings at the appropriate edge of the word were marked as morphologically complex. Some affix strings were excluded because more simple words fit them than complex words. Examples are 'ab-', 'ad-', 're-', '-y' and '-o'. Ultimately, this method of marking words was variably successful depending on the length of the word, and it was relatively conservative, tending to err in the direction of marking simplex words as morphologically complex rather than the reverse.

The success of this morphological discrimination was assessed by randomly sampling 100 words from each category (morphologically simple, morphologically complex) for lengths of 2 syllables, 3 syllables, and 4 syllables. A native English speaker (the author) then checked these randomly sampled words (600 total) and noted the number of incorrect categorizations in each sample.

Because 2-syllable words of English are morphologically complex relatively rarely, a high percentage of words which end in strings that normally constitute an affix are false alarms (e.g. 'vary' ends in '-ary'). On the other hand, the majority of longer

Table 4.6. Number of incorrect morphology categorizations in each random sample of 100 words

	Categorized as:		F1 score
	Simple	Complex	
2 syllables	8	72	0.43
3 syllables	12	13	0.87
4 syllables	26	2	0.84

Table 4.7. Rate of obedience in the lexicon of stress-shift demands

Claimed stress shift:	% obedience	no. suffixes
ultima	71%	13
penultimate	97%	17
antepenultimate	77%	29
pre-antepenultimate	67%	1

words of English are morphologically complex, so strings which are not usually a separate morpheme in shorter words often are in longer words (For example, a final '-y').

The morphological marking was also used to check the 'accuracy' of stress-shifting affixes. Words containing affixes marked as stress-shifting in (Tescner and Whitley, 2004) did have the prescribed stress pattern the majority of the time. In the following table, all affixes of each class are grouped together.

4.5.2 Modeling the Latin Stress Rule

Main stress typically occurs within a three-syllable window at the right edge of the word. This is predicted by the joint action of an alignment constraint demanding that feet be as close s possible to the right edge of the word, and a nonfinality constraint demanding that the final syllable of a word be unfooted. Both constraint types have a long history in the modeling of stress systems, beginning in McCarthy (1986), and both have been used specifically for English stress by Pater (2000); Alcántara (1998). Table 4.8 illustrates the effects of these two constraints. In this tableau and in subsequent tableaux, only trochees are considered as possible feet in English (Hayes,

1995). A trochee is here defined as a one- or two-syllable foot in which stress is on the leftmost syllable.

(1)   ALIGN-R: Assign a violation for every syllable that intervenes between the main stressed foot and the right edge of the word.

(2)   FOOT-NONFIN: Assign a violation if the final syllable of the word is parsed into a foot.

Table 4.8. ALIGN-R and FOOT-NONFIN jointly predict the three-syllable window in which main stress can occur. No ranking is here assumed between the two constraints. Candidates with final, penultimate, or antepenultimate stress incur only a single violation of the two constraints, while candidates with stress leftward of the antepenult incur two or more violations of the alignment constraint

/σσσσ/	ALIGN-R	FOOT-NONFIN
→ σσσ(σ́)		*
→ σσ(σ́σ)		*
→ σσ(σ́)σ	*	
→ σ(σ́σ)σ	*	
σ(σ́)σσ	**!	
(σ́σ)σσ	**!	
(σ́)σσσ	**!	

Here, no ranking is assumed between ALIGN-R and FOOT-NONFIN. Because the two constraints compete, all candidates violate one of them at least once. Candidates with stress further left than the antepenult violate ALIGN-R twice, making them strictly less harmonic than candidates which incur a single violation of one of the constraints. In order to get the effects of weight, FOOTBINARITY (FTBIN) and WEIGHT-TO-STRESS (WTS) are needed. FTBIN requires feet to have at least two moras. A mora can come from either a vowel, or a coda consonant, such that heavy syllables have two moras. Effectively, the constraint penalizes feet which do not

132

consist of a single light syllable. Feet consisting of a single heavy syllable or any two syllables do not violate the constraint.

(3) FTBIN: Assign a violation for every foot which consists of a single light syllable.

(4) WEIGHT-TO-STRESS(WTS): Assign a violation for every heavy syllable not bearing stress.

Table 4.9. OT grammar with ALIGN-R, NONFIN, WTS, and FTBIN. ALIGN-R and NONFIN do not choose between antepenultimate stress and penultimate stress. FTBIN and WTS together make penultimate stress preferable to antepenultimate stress in words with a heavy penult.

	ALIGN-R	FOOT-NONFIN	FTBIN	WTS
/LLL/				
→ (LL)L	*			
(L)LL	**!		*!	
→ L(LL)		*		
L(L)L	*!		*!	
/LHL/				
(LH)L	*!			*!
(L)HL	**!		*!	*!
→ L(HL)		*		
→ L(H)L	*			

In Table 4.9, FTBIN and WTS both force words with a heavy penultimate syllable to take penultimate stress. Words with a light penultimate syllable can take either antepenultimate or penultimate stress, but FtBin rules out the L(L)L option, so that words with a light penult and penultimate stress must violate nonfinality L(LL). This becomes important in modeling the weak preference for antepenultimate stress in light-penult words. Table 4.10 shows the same candidates and constraints. Each constraint is given a weight, and the harmony and MaxEnt probability of each candidate are shown. Probabilities written as zero are in fact nonzero, but very small. All

the constraints have roughly equal weight, but FOOT-NONFIN has a slightly higher weight than ALIGN-R. This slight difference in weight leads to a roughly 70-30 distribution favoring antepenultimate stress on words with a light penult, and the same distribution favoring the L(H)L form over the L(HL) form when the penult is heavy. WTS rules out antepenultimate stress when the penult is heavy, and FTBIN rules out L(L)L.

Table 4.10. Maximum Entropy model of the 'Latin Stress Rule', according to which words with a heavy penult take penultimate stress while words with a light penult prefer antepenultimate stress. The weights given here were fitted by hand to produce a probability distribution over candidates that matches the one seen in the lexicon. Probabilities are the exponent of the harmony score, normalized by the harmony scores for each candidate for that input. Probabilities written as zero are actually nonzero, but very small.

	P	$\mathcal{H}$	ALIGN-R 10	FOOT-NONFIN 11	FTBIN 10	WTS 10
/LLL/						
→ (LL)L	0.73	-10	-1			
(L)LL	0	-30	-2		-1	
→ L(LL)	0.26	-11		-1		
L(L)L	0	-20	-1		-1	
/LHL/						
(LH)L	0	-10	-1			-1
(L)HL	0	-30	-2		-1	-1
→ L(HL)	0.26	-11		-1		
→ L(H)L	0.73	-10	-1			

These four constraints successfully capture the strong preference for penultimate stress on words with heavy penults and the weak preference for antepenultimate stress on words with light penults. (The higher the weight on FOOT-NONFIN, the stronger the preference for antepenultimate stress on light penult words will be). However, the WTS constraint demands that all heavy syllables, regardless of location, have stress. The language learner does not look just at LLL and LHL words in order to learn the stress generalizations - all word shapes must be taken into account.

Table 4.11 shows the results of a search of the CMU pronouncing dictionary. Words in CMU are broken down by the weight pattern exhibited on their last three syllables, and the proportion of words within each category which take each possible weight pattern is listed. Heavy syllables are syllables with either a long vowel or a coda consonant, except for in final position where the final consonant is counted as extrametrical. [ɑ, æ, ə, ɔ, ɛ, ɪ, ʊ, i, u] as well as the syllabic consonants [l̩, ɹ̩, n̩, m̩, ŋ̩] are counted as short vowels, while diphthongs [ei͡, ai͡, ou͡, au͡, oi͡] were counted as long. Morphology and part of speech are ignored in this table - each cell contains both monomorphemic and multimorphemic forms, and all parts of speech. Two-syllable words, and words in which the main stress does not appear on one of the last three syllables (e.g. 'álligator' or 'rélativism') are excluded from the chart. For this reason, some columns do not sum to one.

In Table 4.11, penultimate stress is a good option whenever the final syllable is not heavy. For all light penult words, antepenultimate stress occurs with a relatively high probability, while for heavy penult words, antepenultimate stress is only rarely attested. In words with heavy final or antepenultimate syllables, it is common to stress those syllables. For example, in LLH words, the final syllable gets a secondary stress the majority of the time (102 stress).

The proportions in Table 4.11 were used as training data to fit a MaxEnt model with hidden structure, using Pater and Staubs (2015). The constraint set was based on that in Table 4.10, and is given in Table 4.12.

The constraint set contains two versions each of ALIGN-R and ALIGN-L - in each case one version targeting any stress and one version targeting specifically main stress. There are also three versions of NONFINALITY, one which demands that feet be nonfinal - that final syllables not be footed, one which demands that the final syllable be stressless but does not care whether that syllable is footed, and one which

Table 4.11. Proportions of each stress pattern occurring on each type of word, in the CMU pronouncing dictionary. 1=main stress, 2=secondary stress, 0=stressless. Parentheses indicate foot boundaries. All words three syllables long and longer which have main stress on one of the final three syllables are included. All proportions over 0.1 are in bold. The first column lists the surface form of a potential stress pattern, while the second column lists possible foot structures for each surface stress pattern.

| | | Last three syllables | | | | | | | |
| | | Light Penult | | | | Heavy Penult | | | |
Stress pattern	Structure	LLL	HLL	LLH	HLH	LHL	LHH	HHL	HHH
100	(10)0 (1)00	**0.57**	**0.62**	**0.17**	0.19	0.011	0	0.014	0
102	(10)(2) (1)0(2)	**0.049**	0.076	**0.67**	**0.70**	0.006	0.17	0.004	0.25
120	(1)(20) (1)(2)0	0.002	0.008	0	0	0.004	0	0.039	0
122	(1)(2)(2)	0	0	0	0	0	0	0	0
010	0(10) 0(1)0	**0.23**	**0.14**	0.066	0.015	**0.90**	0.51	**0.51**	0
012	0(1)(2)	0.001	0	0.002	0.004	0.007	0.20	0	0.19
210	(2)(10) (2)(1)0	0.007	**0.13**	0.002	0.030	0.032	0.024	**0.42**	0.38
212	(2)(1)(2)	0	0	0	0.007	0.001	0	0	0.063
001	00(1)	0.002	0	0.002	0	0.001	0	0	0
021	0(2)(1)	0	0	0	0	0.001	0	0	0
201	(20)(1) (2)0(1)	0.020	0.029	0.060	0.052	0.001	0.024	0.007	0.13
221	(2)(2)(1)	0	0.001	0	0	0.001	0.024	0.011	0
Pre-antepenult	main stress	**0.12**	0.005	0.033	0.004	0.040	0.049	0	0
	Total n:	*3120*	*1060*	*636*	*269*	*1087*	*41*	*281*	*16*

demands that main stress not be final, but which does not incur a violation for a final secondary stress.

Weights of these 9 constraints were fit in R using a batch optimization algorithm (Pater and Staubs, 2015). In the lexicon, each word has a single best output, but there is variation across lexical items. In order to reduce the processing power required to fit the data, individual words were collapsed into the weight categories in Table 4.11. Each weight pattern was treated as a single input which could be realized many different ways, with probabilities according to how often that weight pattern occurred

Table 4.12. Constraints used in modeling the data in Table 4.11. Constraints similar to these are used in modeling English stress by Pater (2000) and Alcántara (1998)

Align-R	Assign a violation for every syllable intervening between the right edge of a word and the rightmost foot
Align-Main-R	Assign a violation for every syllable intervening between the right edge of a word and the head foot
Align-L	Assign a violation for every syllable intervening between the left edge of a word and the leftmost foot
Align-Main-L	Assign a violation for every syllable intervening between the left edge of a word and the head foot
Foot-NonFin	Assign a violation if the final syllable of the word is parsed into a foot
Stress-NonFin	Assign a violation if the final syllable of the word bears stress
Main-Stress-NonFin	Assign a violation if the final syllable of the word bears primary stress
FtBin	Assign a violation for each foot that consists of a single light syllable
WTS	Assign a violation for each heavy syllable that does not bear main or secondary stress

with that stress pattern in the lexicon. Each weight pattern had as candidates the 18 possible stress patterns in Table 4.11 (excluding preantepenultimate stress). Because some candidates have the same surface form (e.g. (10)0 and (1)00), the 'hidden structure' option in the algorithm was used so that probabilites were fit to distinct surface forms rather than the hidden structure. A weak L2 regularization term was used, so that each constraint's weight had a Gaussian prior with a mean of 0 and a variance of 100,000[3].

Both left-alignment constraints receive 0 weight, indicating that they are not useful in improving the fit to the lexicon. Most of the work of right-alignment is being done by Align-R. This means that no extra predictive power is gained by a pressure for main stresses specifically to be right aligned. The Align-Main-R constraint might gain more weight if the training included candidates more than three syllables from

_____

[3]L1 regularization was also used (α=0.01) but results were very similar

137

Table 4.13. Weights fit using a batch optimization algorithm, and data from Table 4.11. An L2 regularization term with variance of 100,000 was used. Probabilities associated with the LLL and LHL forms (corresponding to the items in the experiment) are shown on the right. The probabilities from the Lexicon were the training data, model probabilities are those predicted by this set of constraints and weights, and experimental probabilities are those produced by participants in a production task (reported in Chapter 4). The sum squared error (SSE) for all training data is 0.619.

Constraint	Weight			Predictions for experimental data		
ALIGN-R	2.06					
ALIGN-MAIN-R	0.02		Stress	Lexicon	Model	*Exp*
ALIGN-L	0	LLL	100	0.63	0.61	*0.66*
ALIGN-MAIN-L	0		010	0.18	0.33	*0.34*
FOOT-NONFIN	2.79					
STRESS-NONFIN	0	LHL	100	0.01	0.15	*0.25*
MAIN-STRESS-NONFIN	1.78		010	0.91	0.77	*0.75*
FTBIN	2.91					
WTS	1.22			Lexicon SSE: 0.619		

the right edge of the word, since such stresses are rare in English. Both FOOT-NONFIN and MAIN-STRESS-NONFIN get some weight. Recall from Table 4.10 that the specific formulation of FOOT-NONFIN is necessary to require stress to sometimes be antepenultimate. If only STRESS-NONFIN (final syllables are stressless) were used, there would be no constraint preferring a candidate like (10)0 over a candidate like 0(10). Instead, the former would violate ALIGN-R while the latter would be violation-free. Since in the lexicon there is a preference for (10)0 over 0(10) in words with light penultimate syllables, the constraint FOOT-NONFIN is necessary, and in fact the version which is footing-agnostic, STRESS-NONFIN gets 0 weight.

FTBIN gets the highest weight of all the constraints, since it is essentially unviolated in the training data. Candidates with a foot consisting of a single light syllable violate FTBIN, and although there are surface forms in the training data which could be parsed as containing such a foot, these forms are parsed by the model as observing the constraint. For example, 100 stress on a LLL word could be parsed as (1)00,

violating FtBin, but the parse (10)0, not violating FtBin receives more probability. The constraint also rules out many rarely-attested candidates, such as 210 stress with a light initial syllable (210 can only be parsed with a monosyllabic foot on the initial syllable).

WTS also gets a high weight, and jointly with the other constraints predicts different probability of penultimate stress on words with light penults and words with heavy penults. WTS by itself targets all stressed syllables, and does not specifically demand stress on heavy penults. However, because of its interaction with the other constraints, penultimate syllables are particularly affected. Table 4.14 illustrates this. Each cell of the table corresponds to a set of word types which have either a heavy or a light syllable in a particular position. For example, the cell marked 'Antepenult' and 'L' corresponds to all word types with a light antepenultimate syllable: LLL, LHL, LLH, and LHH. The percentage of main stress on that syllable is recorded in the cell, so the number 0.35 indicates that word types with a light antepenult are predicted by the model to take main stress on the antepenult 35% of the time. Main stress on the antepenult could manifest as any of the four shapes: 100, 120, 102, or 122, so 0.35 is the average predicted percentage over all different word types with a light antepenult.

Table 4.14. Locations of main stress predicted by the model in Table 4.13. For each position, the percentage of time that position gets main stress when it is light (L) or heavy (H) is recorded. The model predicts that the weight of the penultimate syllable strongly affects how likely it is to take main stress, while the weight of the final syllables only weakly affects main stress placement, and the weight of the antepenult does not affect its likelihood of taking main stress at all.

		Percentage main stress in position		
		Antepenult	Penult	Final
Weight of	H	0.53	0.74	0.089
Position	L	0.41	0.23	0.003

Table 4.14 shows that the MaxEnt model makes different predictions for different positions in the word with respect to how their weight affects the stress pattern.

Antepenultimate syllables are predicted to be equally likely to take main stress when they are heavy and when they are light. Final syllables are predicted to virtually never take main stress when light, and to take main stress only 0.3% of the time when heavy. Penultimate syllables are predicted to take main stress 74% of the time when heavy but only 23% of the time when light - the biggest difference in the table. Even without a Weight-to-Stress constraint that specifically targets penultimate syllables, the set of constraints from Table 4.12 are able to capture the fact that the weight of the penultimate syllable is special with respect to main stress.

Although the current model does capture the basic fact about the lexicon that penult weight is more important than the weight of other syllables, it does not fit the actual percentages in the lexicon well. As illustrated in Table 4.13, the model underpredicts the rate of antepenultimate stress on LLL words, and underpredicts the rate of penultimate stress on LHL words, which in the lexicon is close to 100%. This is because the weight of the WTS constraint can only get so high without compromising the model's fit on words with heavy syllables in other positions. In order to fit the lexical data more closely, a version of the WTS constraint which specifically targets the penultimate position could be used.

Table 4.15 lists five potential versions of the WTS constraint, including the original version used above. First, the constraint is split up into three versions, each targeting one of the final three positions in the word: WTS-ANTEPENULT, WTS-PENULT, and WTS-FINAL. These constraints do not distinguish between main stress and secondary stress, but the weight of each can vary independently, meaning that the model has the ability to fit the different effects of weight in different positions. Second, a constraint specifically demanding main stress on heavy penults is used: WTS-MAIN-PENULT. This constraint will be used together with the general WTS constraint, for a model in which heavy penults specifically attract main stress, but

140

there is also pressure for all heavy syllables to have some stress, regardless of their position.

Table 4.15. Different versions of WEIGHT-TO-STRESS to be fit to the lexicon.

WTS	Assign a violation for each heavy syllable that does not bear main or secondary stress
WTS-ANTEPENULT	Assign a violation for a heavy antepenultimate syllable that does not bear main or secondary stress
WTS-PENULT	Assign a violation for a heavy penultimate syllable that does not bear main or secondary stress
WTS-FINAL	Assign a violation for a heavy final syllable that does not bear main or secondary stress
WTS-MAIN-PENULT	Assign a violation for a heavy penultimate syllable that does not bear main stress

Table 4.16 shows the outcome of fitting the model using the three different position-specific Weight-to-Stress constraints. Both WTS-PENULT and WTS-FINAL get weight; WTS-ANTEPENULT gets zero weight, indicating that it is not needed to capture the behavior of heavy syllables in antepenult position. Note that FTBIN gets a higher weight in this model than in the previous model - this constraint does not penalize unstressed heavy syllables, but it penalizes stressed light syllables (when they are alone in a foot). Words with a heavy syllable in some position will therefore have more probability on candidates with that position stressed than will words with a light syllable in that position.

The model fit in Table 4.16 is improved somewhat compared to the model with just a single WTS constraint - the sum squared error for the fit to the data in Table 4.11 is 0.555 rather than 0.619. An F-test comparing these two SSE terms finds that the gain in model fit is significant: $F(2,123)=7.09$, corresponding to a p of 1.2×10^{-03}. Degrees of freedom are calculated as follows: for the numerator, 11 parameters (constraints) in the model with three WTS constraints minus 9 parameters in the model with one. For the denominator, 134 data points (the number of data points in Table 4.11) minus 11 parameters in the full model.

Table 4.16. Weights fit to the three versions of WEIGHT-TO-STRESS, using a batch optimization algorithm, and data from Table 4.11. An L2 regularization term with variance of 100,000 was used. Probabilities associated with the LLL and LHL forms (corresponding to the items in the experiment) are shown on the right. The probabilities from the Lexicon were the training data, model probabilities are those predicted by this set of constraints and weights, and experimental probabilities are those produced by participants in a production task (reported in Chapter 4). The sum squared error (SSE) for all training data is 0.555.

Constraint	Weight			Stress	Lexicon	Model	*Exp*
ALIGN-R	0.33						
ALIGN-MAIN-R	1.02			Stress	Lexicon	Model	*Exp*
ALIGN-L	0.27		LLL	100	0.63	0.60	*0.66*
ALIGN-MAIN-L	0			010	0.18	0.35	*0.34*
FOOT-NONFIN	1.65						
STRESS-NONFIN	0		LHL	100	0.01	0.16	*0.25*
MAIN-STRESS-NONFIN	2.97			010	0.91	0.75	*0.75*
FTBIN	3.03						
WTS-ANTEPENULT	0				Lexicon SSE: 0.555		
WTS-PENULT	1.31						
WTS-FINAL	2.2						

Table 4.17 shows the results of fitting a model with both the general WTS constraint and a specific WTS constraint demanding that heavy penults take main stress. In this fit, both the general WTS constraint and the specific WTS-MAIN-PENULT get a high weight. The improvement in model fit relative to the model with just the general WTS constraint is significant: $F(1,124)=63.67$, $p=8.4 \times 10^{-13}$. The probabilities predicted by this model specifically for the forms tested in the experiment mirror the lexical distribution better than in the previous two models - and also start to diverge from the probability distributions found in the production experiment.

Both models with more specific WTS constraints fit the lexical training data better than the model with only the general WTS constraint, and the model with the specific constraint WTS-MAIN-PENULT, demanding specifically main stress on exactly heavy penultimate syllables matches the lexical distributions for LHL words particularly well. However, the model with the general WTS constraint generates

Table 4.17. Weights fit using a batch optimization algorithm, and data from Table 4.11, including WTS-MAIN-PENULT. An L2 regularization term with variance of 100,000 was used. Probabilities associated with the LLL and LHL forms (corresponding to the items in the experiment) are shown on the right. The probabilities from the Lexicon were the training data, model probabilities are those predicted by this set of constraints and weights, and experimental probabilities are those produced by participants in a production task (reported in Chapter 4). The sum squared error (SSE) for all training data is 0.409.

Constraint	Weight			Stress	Lexicon	Model	*Exp*
ALIGN-R	0.16						
ALIGN-MAIN-R	1.59			Stress	Lexicon	Model	*Exp*
ALIGN-L	0		LLL	100	0.63	0.59	*0.66*
ALIGN-MAIN-L	2.41			010	0.18	0.29	*0.34*
FOOT-NONFIN	0						
STRESS-NONFIN	0		LHL	100	0.01	0.06	*0.25*
MAIN-STRESS-NONFIN	1.08			010	0.91	0.87	*0.75*
FTBIN	2.87						
WTS	2.32				Lexicon SSE: 0.409		
WTS-MAIN-PENULT	0.9						

The header "Predictions for experimental data" spans the right portion of the table.

predicted probabilities on LLL and LHL words which fit participants' behavior in the production experiment, while the model with WTS-MAIN-PENULT underpredict participants' rate of producing 100 stress on LHL words. Participants' undermatch of the lexical distribution observed in Experiment 4 can be understood in terms of the constraints that participants have learned - although a WTS constraint specifically targeting penultimate syllables would produce a better match to the training data (the lexicon of English), this constraint is not available to learners. This could be because the constraint is not part of the universal constraint set available to learners, or it could be that the complexity of such a constraint makes it too costly to incorporate into the phonological grammar.

4.5.3 Scaling up

The training data in the previous section consisted of a single input for each type of word (LLL, LHL, HLL, . . .), with a probability distribution over candidate outputs

which matched the lexical distribution over stress patterns on words of that type. Two fundamental pieces of information are missing from this set of training data: first, that each word has a single correct pronunciation and does not vary, and second that the various types of words occur differently often. For example, LLL words are about ten times as common as LHL words.

In this section, the previous model is scaled up so that the entire lexicon of English serves as training data. A set of tableaux were created via a python script - one for each entry in the CMU dataset described in section 4.5.1, for a total of 11765 inputs. Because the number of tableaux was so large, the MaxEnt Grammar Tool (Wilson and George, 2009) was used instead of the HGR optimizer. Because the MaxEnt Grammar Tool has no way to handle hidden structure, only surface strings were used as candidates. Candidate stress patterns were generated for each word based on its length. Every possible combination of 1, 0, and 2 stress which had exactly one main (1) stress was included in the candidate set. Thus, for two syllable words the candidate set was (10,01,12,21). For purposes of assigning constraint violations, each candidate was assumed to have foot structure such that each stress corresponded to a trochee consisting of itself and any following 0's. A candidate of shape 200120 would have the structure (20)0(1)(20). Candidates with the same surface form did not compete with each other: 010 stress was only given the footing 0(10), not 0(1)0.

Because not all footing candidates were included in the set, the constraint definitions were adjusted accordingly. The new statements of each constraint are given in Table 4.18. FOOT-NONFIN previously served to prefer antepenultimate stress to penultimate stress in forms with a light penult. This constraint is cast here instead as HAVEFINALLAPSE, a constraint demanding that a word end in two stressless syllables. This constraint will prefer (10)0 stress to 0(10), just as FOOT-NONFIN did previously. Unlike FOOT-NONFIN, it will always be violated by penultimate main stress.

144

Table 4.18. Constraints used in modeling the entire lexicon (CMU) with Wilson and George (2009)

Align-R	Assign a violation if the final syllable is not stressed
Align-Main-R	Assign a violation for every syllable intervening between the right edge of a word and the head foot
Align-L	Assign a violation if the initial syllable is not stressed
Align-Main-L	Assign a violation for every syllable intervening between the left edge of a word and the head foot
HaveFinalLapse	Assign a violation if the final two syllables are not stressless
Stress-NonFin	Assign a violation if the final syllable of the word bears stress
Main-Stress-NonFin	Assign a violation if the final syllable of the word bears primary stress
FtBin	Assign a violation for each light stressed syllable not followed by a stressless syllable
WTS	Assign a violation for each heavy syllable that does not bear main or secondary stress

The weights learned by the MaxEnt Grammar Tool are reported in 4.19, as are the model's predictions for the items tested in the production experiment. This model's predictions fit the experimental data well - antepenultimate stress is preferred when the penultimate syllable is light and penultimate stress is preferred when the penultimate syllable is heavy, but antepenultimate stress when the penult is heavy still gets substantial probability.

Although this model provides a reasonably good fit to the lexicon as a whole, and predicts participants' data well, it has no way to distinguish between different words with the same phonological shape, like those shown in Table 4.20. An extra mechanism is needed to get each lexical item correct, which will be discussed in the next section.

Table 4.19. Weights fit using the MaxEnt Grammar Tool, with the entire lexicon as training data. An L2 regularization term with variance of 100,000 was used. Predicted probabilities for with LLL and LHL forms (corresponding to the items in the experiment in chapter 4) are shown on the right.

Constraint	Weight	Predictions for experimental data				
			Stress	Lexicon	Model	*Exp*
ALIGN-R	0					
ALIGN-MAIN-R	0.70		Stress	Lexicon	Model	*Exp*
ALIGN-L	0	LLL	100	0.63	0.62	*0.66*
ALIGN-MAIN-L	0		010	0.18	0.34	*0.34*
HAVEFINALLAPSE	1.29					
STRESS-NONFIN	0.11	LHL	100	0.01	0.24	*0.25*
MAIN-STRESS-NONFIN	1.85		010	0.91	0.67	*0.75*
FTBIN	3.14					
WTS	1.6					

Table 4.20. The model's predictions for lexical items with the same weight pattern but different correct stresses

	gálaxy (LHL)	enígma (LHL)	récipe (LLL)	spaghétti (LLL)
100	0.24	0.24	0.62	0.62
010	0.67	0.67	0.34	0.34
210	0.03	0.03	0.01	0.01
012	0.03	0.03	0	0

4.6   Conclusion

In a productivity experiment, the Latin Stress Rule was found to be productive, but participants mismatched the near-categorical trend in the lexicon, producing main stress on a heavy penult about 75% of the time rather than the 96% of the time observed in the lexicon. This is a case of mismatching, or underlearning, of lexical statistics similar to those found by Hayes et al. (2009); Becker et al. (2011), and others. They argue that learners either fail to learn or underlearn generalizations in the lexicon when those generalizations exceed the grammatical system's capacity to either learn or represent complex generalizations. Such mismatching would not be expected if learners could freely induce constraints of arbitrary complexity. I provide a model of participants' undermatching of the lexical trend using Maximum Entropy Grammar.

The general constraint WEIGHT-TO-STRESS can, when acting together with the other constraints involved in the English stress system, single out penultimate syllables to attract main stress. However, the system which uses WEIGHT-TO-STRESS cannot precisely match the lexicon. It predicts a rate of penultimate main stress on heavy penults similar to that actually observed in the experiment. A closer match to the lexicon can be achieved by using a more specific constraint which specifically demands main stress on a heavy penult, and ignores both secondary stress and other positions in the word. I argue that participants mismatch the lexicon the way they do because their internal grammar contains the general constraints WEIGHT-TO-STRESS, which does not include in its structural description a statement about which syllable is most relevant, or about which type of stress a heavy syllable should attract. This constraint is structurally simpler than that constraint which produces a closer match to the lexicon.

CHAPTER 5

CONCLUSION

This dissertation investigates the cognitive representation of probabilistic phonotactic patterns, arguing that speakers' phonological grammar contains abstract representations which are probabilistic in nature. Those probabilistic representations are used not only during production and judgment of novel words, but also in the perception of known words. Humans' ability to generalize probabilistic trends in their lexicon to known words has been extensively studied previously, for example in Ernestus and Baayen (2003); Hayes et al. (2009); Becker et al. (2011). One contribution of this dissertation is the direct comparison of a 'grammatical' account of this ability to generalize with alternative accounts.

In Experiment 3, a grammatical model of a probabilistic trend in the English stress system is compared to an analogical account. Participants choose a stress pattern for novel words, and additionally provide possible analogical bases for each nonword. They successfully extended a trend in the lexicon to nonwords in production. If they used a process of analogy to do so, we would expect the stress pattern of the provided analogical bases to predict the stress pattern a participant chose in production. No such correlation is observed in the experiment. Additionally, the aggregate stress properties of all of the provided analogical bases for an item do not predict participants' behavior on that item. Participants productively apply a trend in the lexicon of English to novel words, but do not use analogy to do so. Instead, I argue that abstract grammatical generalizations are used.

Nonwords in Experiment 3 were specifically chosen to be phonologically very distinct from any actual words of English. That is, while they were all phonotactically licit, they did not have any near neighbors in the lexicon. In past similar experiments, Guion et al. (2003); Baker and Smith (1976) have found that properties of participant-provided analogical bases did at least partially predict participants' productions. In both of these experiments, nonword items did have relevant lexical neighbors. Potentially, participants are able to analogize to actual words that are sufficiently similar to a given nonword, but have difficulty finding an appropriate analogical base for a nonword with no near lexical neighbors. If so, Experiment 3 demonstrated that in the absence of tempting clear neighbors to analogize to, participants have another mechanism available to them to generalize probabilistic trends to nonwords. I argue that this mechanism is the phonological grammar.

In Experiment 4, chapter 4, I investigate the Latin Stress Rule, a near-categorical trend in the lexicon of English. Longer words of English with a heavy penultimate syllable nearly always take penultimate main stress. I find that speakers can extend this trend to novel words, but do not match the lexical statistics exactly. Rather, they under-match, producing an approximately 70%-30% distribution of forms, while the distribution in the lexicon is approximately 95%-5%. I compare two potential grammatical models of the trend. One model predicts a close match to the lexical statistics but uses a complex constraint that singles out a particular location in the word (the penultimate syllable), and demands that it take main stress when heavy. Using this constraint, it is possible to closely match the lexical distribution of 95%-5%. However, a grammar with a simpler and more general version of this constraint - the well-known WEIGHT-TO-STRESS constraint - can still predict a preference for main stress to be on heavy penults, but it undermatches the lexical distribution, predicting a distribution much closer to that produced by participants in the experiment. I argue

that participants undermatch the lexical statistics because they have this simpler, more general constraint rather than the more complex one.

This finding is an instance of a much more broadly studied phenomenon of simplicity bias in categorization tasks, both in language and in other cognitive systems (see Moreton and Pater, 2012 for an overview of the linguistic literature). Generalizations involving fewer features seem to be consistently easier to learn in the laboratory than generalizations involving more features. The finding of Experiment 4, that a simplicity bias could cause a learner to undermatch the statistics available in the training data, raises an interesting set of questions: Under what circumstances will a gain in simplicity lead a learner to mismatch the training data? How much mismatch will be tolerated before the learner will resort to a more complex generalization?

Experiments 1 and 2, in chapter 2, investigate the effects on lexical processing of probabilistic phonotactics. I find that the same probabilistic trend studied in experiments 3-4 affects the process of accessing an already known word. Specifically, actual words of English which violate the trend are processed differently from actual words which observe the trend. This could be due to a difference in representation of the lexical form of violators and observers, or it could be the effect of a violation of expectations imposed by the grammar on lexical items. An explanation of the former type would follow in the footsteps of previous literature on the English stress system in particular, and inter-word variability in general, which proposes some type of lexical marking of exceptional forms. In the latter case, the trend-violating status of exceptional words would not be marked on their lexical entry in any way. Rather, the (probabilistic) grammar would impose certain expectations on the lexical access process, so that when an exceptional item which violated those expectations is encountered, more effort must be expended by the processing system to resolve the conflict.

Probabilistic generalizations like the ones studied here, which are statistics across different words of the lexicon rather than across different instances of the same word, are interesting in that they are not strictly necessary for a learner to acquire. A speaker of English would in principle have no trouble correctly perceiving and producing all the words of their language without knowing the statistical trends across them. Furthermore, knowing the trend does not save on memory usage in the lexicon. Each individual word's behavior must be memorized, since each word behaves the same way each time (the same word's stress pattern does not vary from utterance to utterance for example). Despite this, humans form abstract generalizations about probabilistic trends in their lexicon and use those trends not just on novel words, but also in the everyday perception of words they already know.

APPENDIX: WORDS USED IN EXPERIMENTS AND MODELING

Lists of words

For each weight category in the table in chapter 4, and for each stress pattern, the table below lists the words belonging to that category present in the corpus used for training. Syllables designated as L (light) have a short nucleus and no coda when they are word-internal, and may have a single consonant coda when they are the final syllable in the word (under the assumption that word-final consonants are extrametrical, Chomsky and Halle, 1968; Hayes, 1982). Syllables designated as H (heavy) have either a long nucleus, a coda consonant, or both. Diphthongs [ẽɪ, ãɪ, õʊ, ãʊ, ɔ̃ɪ] were counted as long nuclei, while the non-diphthongs [ɑ, a, æ, ɛ, ɪ, i, ʊ, u, ɹ̩, m̩, n̩, ŋ̩, l̩] were counted as short. Stress transcription is taken directly from the corpus used, Hayes (2012), a version of the CMU pronouncing dictionary Weide (1994). Lists of words for cells of the table containing less than 100 items are exhaustive. For cells containing more than 100 items, 100 words are randomly selected. See Section 4.5.1 for a much more in-depth discussion of the corpus used.

LLL   *100 (count:1784)*

electricity, originality, rational, regional, coventry, feasibility, elegant, inanimate, sesame, numeral, cyclical, topography, geography, psychological, insecurity, impartiality, cannibal, globular, reversible, artillery, obsidian, significance, preferable, leniency, conductivity, charleston, siberia, fidelity, corridor, pornographer, burial, impossible, fugitive, silicon, mendacity,

subtlety, clarity, hypothetical, papua, omnivorous, furrier, victoria, morality, chemistry, detriment, soluble, volatility, acropolis, felony, respectability, libyan, progenitor, privilege, graphical, depravity, opulence, memorial, severity, tammany, milliner, capitol, remedial, obedient, municipal, paleontology, obscenity, possible, indicative, crematorium, geophysical, predominance, sociology, periphery, syphilis, theological, anomalous, bicentennial, facility, administer, pharmacology, sheraton, genealogy, extremity, passenger, bedouin, irresistible, lazarus, glossary, calorie, orifice, tricycle, jeopardy, requiem, banister, funeral, dominica, nitroglycerine, effusive, daiquiri, absenteeism

102 (count:152)

turpitude, sycamore, panama, solicitude, anagram, hemisphere, certitude, gemini, milligram, epitaph, parachute, commissar, catapult, parapet, travelogue, juggernaut, kilogram, destitute, attitude, resolute, epithet, telegram, hercules, parakeet, aniline, ionosphere, stevedore, hypocrite, lunatic, habitat, caviar, taciturn, monologue, acetylene, guillotine, methanol, monolith, lollipop, astrakhan, pedicure, gratitude, decibel, paragon, matador, carousel, aerosol, troubadour, commodore, posterior, ecuador, camouflage, afghanistan, ballyhoo, monogram, vestibule, aristocrat, phenomenon, metaphor, epigram, sabotage, polygon, humidor, epilogue, minuscule, mastodon, catalogue, apogee, ziggurat, stratosphere, molecule, magazine, boomerang, mackintosh, rubicon, democrat, registrar, damocles, synonym, politic, synagogue, isosceles, demagogue, astronaut, residue, caravan, acrobat, paragraph, bivouac, philippine, cameroon, repertoire, echelon, socrates, caribou, diplomat, analogue, internecine, platypus, latitude, sassafras

120 (count:7)

cucumber, beautician, autism, hospital, tourism fascism, eleventh

010 (count:721)

pantheistic, aggression, analytic, dissolution, resilient, madagascar, democratic, concurrence, ebullient, nirvana, magician, acquisition, antagonistic, cyrillic, metallic, cinematic, professor, tequila, psychiatric, disaster, oceanic, imprimatur, debacle, electrocution, toboggan, delicatessen, elicit, rebellion, european, superstitious, academic, subpoena, distribute, masonic, repression, exposition, lilliputian, repellent, attrition, panorama, thesaurus, paprika, convergent, republic, persecution, congested, religion, recurrence, atavistic, psychokinesis, catatonic, apparent, regressive, suspicion, civilian, concertina, emblematic, intuition, photographic, polemic, prestigious, burundi, erudition, enamel, aggressive, assistant, parasitic, disparage, vicuna, apparition, circumference, semitic, narcissistic, expedition, consultant, demolition, brazilian, cathedral, aromatic, abolish, consider, ceramic, pornographic, ritualistic, condition, desertion, antibiotic, indecision, melancholic, riviera, periodic, acquittal, emission, possession, fruition, curmudgeon, paraplegic, iridescent, tahiti, nicaraguan

012 (count:4)

djibouti, decathlon, attribute, devalue

210 (count:23)

quadruple, autistic, senora, ennui, audition, athletic, recession, neurotic, umbrella, sumatra, auspicious, already, proficient, mascara, laconic, traumatic, autumnal, swahili, hysteric, eureka, tunisia, fluorescent, horrific

212 (count:0)

001 (count:7)

mademoiselle, senegalese, aquamarine, catamaran, gobbledygook, hullabaloo, entrepreneur

021 (count:0)

201 (count:63)

immature, souvenir, suffragette, chevalier, persevere, cavalier, millionaire, profiteer, nominee, reproduce, opportune, baronet, baccarat, repartee, guaranty, pirouette, domineer, puppeteer, nepalese, brigadier, repossess, rationale, clarinet, sudanese, figurine, statuette, saboteur, intervene, reappear, halloween, seventeen, interfere, restaurateur, jubilee, esplanade, solitaire, javanese, connoisseur, kitchenette, minuet, acquiesce, minaret, commandeer, personnel, coronet, buccaneer, guarantor, supersede, mutineer, comedienne, debonair, volunteer, musketeer, racketeer, guarantee, inopportune, silhouette, japanese, cigarette, referee, senegal, reminisce, gelatine

221 (count:0)

Pre-antepenult (count:336)

eligible, controversy, eroticism, military, criticism, impracticable, expeditionary, fanaticism, stoicism, auditory, admirable, seminary, obligatory, mandatory, expediency, presidency, narcissism, exploratory, technicolor, deviancy, bolshevism, palatable, memorable, burgomaster, navigable, judaism, depository, spiritual, magnetism, romanticism, conciliatory, cautionary, operable, enthusiasm, commissary, dysentery, dormitory, inseparable, constituency, preparatory, reformatory, stationary, hypnotism, constabulary, angelica, confucianism, inaccuracy, corollary, botulism, misogyny, euphemism, observatory, ancillary, monastery, actionable, solitary, signatory, relevancy, cataclysm, amicable, momentary, mortuary, veterinary, considerable, retaliatory, degenerative, mimeograph, hereditary, impressionism, category, figurative, inoperative, evolutionary, operative, agriculture, antagonism, derogatory, catholicism, altruism, terrorism, legendary, elitism, adversary, depilatory, candidacy, efficacy, dilatory, incorrigible,

mercenary, perishable, illiteracy, irrevocable, incalculable, mannerism, chauvinism, lavatory, fragmentary, ordinary, cumulative, confectionery

HLL   *100 (count:655)*

powdery, available, admiral, excelsior, pancreas, conversational, sexual, symphony, titanium, gargantuan, copious, fictional, alias, intelligentsia, expectancy, deformity, sensory, infamous, icicle, responsible, rotary, advisory, vasectomy, radium, radius, particle, conceptual, accident, fortunate, irresponsible, pregnancy, stadium, incompetent, calcium, hosiery, peremptory, corporate, rancorous, scandinavia, vacancy, diocese, myopia, tournament, eventual, indigent, quantity, claustrophobia, indolent, melodious, odious, ignorant, harmonious, intransigent, pendulous, laxative, temperate, tendency, accordion, entitlement, bipartisan, insomnia, symposium, amnesty, formula, pilferage, patronage, impractical, corruptible, ascendancy, argument, constable, macedonia, romanian, vagrancy, maniacal, liable, farcical, directional, article, buoyancy, testimonial, ambient, binary, instrument, structural, malignancy, satisfactory, dialectical, industry, obdurate, cadmium, recalcitrance, campion, omnibus, cranium, raphael, imperceptible, petroleum, normative, larceny

102 (count:81)

absolute, diagram, pliocene, alcohol, bayonet, mountaineer, altitude, cardiac, singapore, mandolin, nymphomaniac, dinosaur, neanderthal, albatross, pantheon, bolshevik, atmosphere, cosmonaut, maniac, centrifuge, intercom, portuguese, octopus, hexagon, dialogue, arithmetic, constitute, magnitude, technocrat, photograph, quantifiable, marchioness, cyclotron, carnivore, interlude, phonograph, pleistocene, diaphragm, protocol, centipede, handicap, almanac, institute, olympiad, pentagon, lexicon, insomniac, marmoset, barbecue, aptitude, subterfuge, fortitude, handker-

chief, ineptitude, electrocute, amplitude, rendezvous, infidel, interview, caesarean, soviet, antifreeze, kleptomaniac, substitute, longitude, execute, extrovert, zodiac, parmesan, diadem, caveat, argentine, nincompoop, hypochondriac, rectitude, apricot, orangutan, counterfeit, manganese, octagon, laundromat

120 (count:8)

escapism, orgasm, ancestry, library, ancestor, primary, sarcasm, racism

010 (count:145)

succession, excretion, chinchilla, encourage, exchequer, imprudent, adherence, exertion, intrusion, insistence, inhuman, polaris, nomadic, envelop, fictitious, expression, quixotic, coercion, imperil, internal, indifference, admission, indonesian, impostor, intrepid, invalid, enigmatic, embarrass, opossum, rotunda, ingenious, injustice, coercive, coherence, infertile, imprison, recognition, infusion, adolescence, endeavour, infrequent, embody, coherent, inhibit, diagnostic, insurgent, inclusive, cohesion, explicit, acknowledge, inversion, indecent, successive, charismatic, linguistic, external, poetic, subversion, inclusion, stigmata, obstetric, indonesia, koala, expulsion, submission, insistent, exhaustive, embezzle, subversive, exclusion, imbalance, obsessive, exhibit, infernal, adherent, evolution, homeric, interpret, suggestive, subsistence, incessant, inherit, adhesive, neoclassic, indulgence, ignition, novella, prohibit, observance, examine, exotic, quintessence, existence, incoherent, indebted, indulgent, incision, esoteric, intrusive, admittance

012 (count:0)

210 (count:134)

157

narcotic, antacid, angelic, nocturnal, emphatic, myopic, ambitious, amoral, clitoris, diurnal, honduras, honduran, monsignor, pentathlon, italic, adhesion, artistic, volcanic, electrician, pineal, gardenia, transmission, titanic, orchestral, triumphal, transparent, alpaca, magnetic, electronic, norwegian, ancestral, digression, electromagnetic, idiosyncratic, idyllic, philharmonic, transition, agnostic, aerodynamic, antenna, cosmetic, chaotic, transference, albeit, translucent, hyena, manhattan, magnesia, malfeasance, climatic, harmonic, poinsettia, admonish, technician, ecstatic, dynamic, fantastic, composite, vendetta, reptilian, optician, angora, armada, bengali, encircle, sarcastic, spasmodic, bitumen, afghani, scintilla, tortilla, arthritic, transgressor, psychotic, entreaty, iraqi, idea, amnesia, mortician, phantasm, galvanic, franciscan, existent, iguana, asthmatic, transfusion, combative, marsala, encompass, partition, symphonic, asbestos, transgression, sardonic, ambition, envision, pugnacious, triumphant, lymphatic, abhorrent

212 (count:0)

001 (count:0)

021 (count:0)

201 (count:31)

clientele, engineer, incomplete, violin, trampoline, marketeer, entourage, chandelier, bombardier, amputee, questionnaire, auctioneer, expertise, bangladesh, congolese, tangerine, propylene, intercede, gondolier, convalesce, pioneer, vietnamese, indiscreet, submarine, obsolete, marguerite, tambourine, potpourri, siamese, coalesce, kangaroo

221 (count:1)

denouement

Pre-antepenult (count:5)

egoism, heroism, embolism, ignominy, ecosystem

LLH   *100 (count:109)*

specialist, neurologist, meteorologist, radiologist, strategist, classicist, entomologist, asterisk, illusionist, supremacist, lobbyist, dramatist, patio, video, adagio, pertinent, impressionist, eucharist, antagonist, oregano, psychotherapist, audio, horticulturist, anarchist, expressionist, amethyst, pessimist, geologist, ecologist, humanist, fetishist, anatomist, ophthalmologist, economist, astrophysicist, protagonist, dermatologist, publicist, terrorist, biologist, vertigo, earliest, cheerio, pianist, pugilist, herbalist, impresario, stereo, penitent, individualist, zoologist, paleontologist, physiologist, empiricist, analyst, pathologist, moralist, generalissimo, piccolo, therapist, revisionist, buffalo, medico, eskimo, archaeologist, criminologist, sociologist, calico, populist, physio, intaglio, anthropologist, communist, accompanist, journalist, abolitionist, novelist, domino, satirist, nuncio, studio, catalyst, physicist, scenario, curio, pacifist, gigolo, humorist, botanist, exhibitionist, feminist, columnist, dissident, colonist, geneticist, pistachio, psychologist, methodist, politico, seismologist

102 (count:423)

stipulate, elaborate, personify, brutalize, glorify, commemorate, capricorn, ameliorate, estimate, urinate, episode, participate, flagellate, eradicate, penalize, retrospect, cellophane, animate, rheumatoid, pontificate, canonize, desecrate, circumcise, crocodile, avalanche, educate, rehabilitate, horrify, relegate, celebrate, humanize, classify, edify, initiate, boulevard, dissipate, cicerone, aspirate, energize, infuriate, assimilate, perpetuate, hermaphrodite, paradox, manifold, assassinate, immolate, situate, indi-

vidualize, apologize, attenuate, reprimand, columbine, resume, automate, domesticate, speculate, document, terrify, qualify, jeopardize, justify, exterminate, agonize, desolate, baritone, dividend, calamine, surrogate, occupy, exonerate, creosote, cultivate, copulate, deviate, casserole, duplicate, catacomb, derelict, realize, colonize, manuscript, suburbanite, amortize, terminate, legitimize, aureole, fascinate, diversify, parasite, cellulose, pacify, exaggerate, revolutionize, clarify, nauseate, appreciate, stimulate, precipitate, classified

120 (count:)

010 (count:42)

filipino, tumescent, morocco, defeatist, allegro, colorado, desperado, casino, tomato, elitist, avocado, violinist, fiasco, chicano, tomorrow, mikado, balloonist, merino, histrionics, gestapo, maraschino, tabasco, bonito, deforest, stiletto, palomino, tobacco, mulatto, bravado, opportunist, libido, vibrato, libretto, armadillo, placebo, manifesto, mosquito, reservist, soprano, aficionado, incommunicado, staccato

012 (count:1)

apostate

210 (count:1)

espresso

212 (count:0)

001 (count:1)

legerdemain

021 (count:0)

201 (count:38)

interrupt, recreate, presuppose, correspond, redirect, debutante, redeploy, disappoint, decompose, lemonade, interstate, attache, seventeenth, promenade, commandant, apropos, communique, circumvent, recommend, renaissance, represent, serenade, dossier, cabaret, intercept, resurrect, afrikaans, untoward, entertain, intersect, interact, statuesque, polonaise, recollect, prearrange, ascertain, apprehend, masquerade

221 (count:0)

Pre-antepenult (count:21)

memorialize, personalize, heliotrope, nationalist, cannibalize, manicurist, deteriorate, oxygenate, cuneiform, serialize, capitalist, naturalist, materialize, meteorite, categorize, hospitalize, generalize, capitalize, rationalize, visualize, trivialize

HLH *100 (count:50)*

pharmacist, loyalist, psychiatrist, environmentalist, receptionist, chauvinist, competence, optimist, ratio, scientist, competent, fatalist, isolationist, philanthropist, socialist, fundamentalist, picturesque, perfectionist, activist, rodeo, atheist, pragmatist, polio, calvinist, braggadocio, exorcist, interest, motorist, mexico, finalist, alchemist, abortionist, organist, conservationist, narcissist, arsonist, contortionist, extortionist, hypnotist, evangelist, indigo, scorpio, zionist, vocalist, radio, satanist, embryo, dynamo, portico, fellatio

102 (count:189)

contretemps, symbolize, carbonate, procreate, matriarch, simplify, consummate, carmelite, alternate, desensitize, contemplate, hibernate, indicate, artifact, vindicate, intoxicate, concentrate, prototype, diatribe, authenticate, intimate, antidote, amplify, indemnify, infanticide, quantify,

concubine, sympathize, conjugate, compliment, socialite, potentate, appropriate, insecticide, sauerkraut, extricate, mortify, confiscate, antelope, artichoke, incarcerate, dissociate, magnify, alkali, idolize, advocate, disintegrate, plagiarize, alienate, organize, compromise, contraband, accentuate, grandiose, dignify, archetype, compensate, captivate, envelope, expedite, counteract, consecrate, extradite, dictaphone, imbecile, orchestrate, asphyxiate, compartmentalize, frankincense, coordinate, electrify, corduroy, vaccinate, insulate, xylophone, comatose, anglophile, intensify, bifurcate, romanticize, harmonize, palpitate, porcupine, normalize, rectify, anthracite, maximize, congregate, chromosome, calculate, immobilize, fortify, instigate, synthesize, radiate, integrate, microscope, podiatrist, tantamount, centigrade

120 (count:0)

010 (count:4)

inferno, extremist, bolero, imperfect

012 (count:1)

promulgate

210 (count:8)

sombrero, bordello, cartoonist, tuxedo, falsetto, bambino, incognito, torpedo

212 (count:2)

zimbabwe, transvestite

001 (count:0)

021 (count:0)

201 (count:14)

improvise, inhumane, juxtapose, nonchalant, comprehend, interject, romanesque, contradict, indirect, impolite, nonchalance, condescend, indistinct, antitank

221 (count:0)

Pre-antepenult (count:1)

characterize

LHL   *100 (count:12)*

emotion, imagery, messenger, character, cylinder, heroin, heroine, hanover, government, wellington, galaxy, adjective

102 (count:6)

autobahn, bureaucrat, reservoir, yugoslav, discotheque, turbojet

120 (count:4)

prospector, autopsy, monarchy, prefecture

010 (count:974)

appropriation, pretended, contagious, ornamental, insinuation, cancellation, classification, activation, modulation, despondent, effective, concoction, detective, restriction, calibration, renunciation, component, enormous, appliance, rhododendron, pneumonia, corespondent, benighted, affirmation, organization, prevention, appraisal, reactor, maturation, avenger, require, consummation, exhalation, indecisive, defiance, dissociation, mitigation, arachnid, congratulation, equator, demotion, insubordination, elevation, confection, vigilante, desire, continuation, creative, defensive, insurrection, fascination, incorporation, delegation, circumstantial, equalization, decapitation, acacia, addictive, disruption, defamation, cicada, irrigation, machination, amplification, consolidation, cataclysmic, crustacean, speculation, prediction, adulation, anaconda, predictor, amalgam,

redemption, ineffective, resuscitation, assimilation, invocation, preponderance, abortion, reportage, incremental, provincial, sclerosis, attraction, suspension, coincidental, syncopation, coagulation, profanation, investigation, ascendant, remainder, presumptive, appellation, conduction, saliva, verification, aberration, rapacious

012 (count:8)

aforethought, paranoiac, elephantine, labyrinthine, immobile, electron, aforesaid, benedictine

210 (count:35)

escarpment, causation, cessation, prostration, authentic, protestation, existential, australia, oration, forensic, castration, narration, duration, manifestation, glaucoma, retire, gestation, torrential, anarchic, exploration, horrendous, uganda, molestation, neurosis, discovery, exhumation, australian, infestation, voracious, caucasian, eurasian, nostalgia, nostalgic, gastritis, audacious

212 (count:1)

jujitsu

001 (count:1)

cabriolet

021 (count:1)

vietnam

201 (count:1)

reimburse

221 (count:1)

raconteur

Pre-antepenult (count:42)

escalator, quantitative, necromancy, commentator, apoplexy, epilepsy, supervisor, oleander, accumulative, dandelion, legislature, legislative, alligator, elevator, matrimony, dedicated, acrimony, miscellany, manipulative, commemorative, benefactor, chiropractor, oligarchy, meditative, authoritative, nomenclature, alimony, refrigerator, patriarchy, administrative, kindergarten, gladiator, qualitative, orthodoxy, ceremony, ventilator, interloper, helicopter, architecture, instigator, necromancer, testimony

LHH   *100 (count:0)*

102 (count:7)

sacrosanct, aerospace, recognize, reprobate, innovate, designate, anecdote

120 (count:0)

010 (count:21)

rococo, appendix, alarmist, flamenco, machismo, akimbo, innuendo, concerto, virtuoso, archipelago, flamingo, crescendo, potato, calypso, esperanto, majordomo, conformist, propagandist, proviso, memento, commando

012 (count:8)

percentile, electrode, divorcee, dehydrate, projectile, apartheid, monoxide, peroxide

210 (count:1)

lumbago

212 (count:0)

001 (count:0)

021 (count:0)

201 (count:1)

inexact

221 (count:1)

fiance

Pre-antepenult (count:2)

saxophonist, telephoto

HHL   *100 (count:4)*

parkinson, howitzer, harbinger, bankruptcy

102 (count:1)

introvert

120 (count:11)

bonfire, rectangle, hierarchy, badminton, exponent, triangle, tormentor, anchovy, archetypal, andiron, spectator

010 (count:143)

example, hypnosis, extortion, insubstantial, developmental, extraction, infraction, olympus, momentous, inventor, israeli, fixation, superintendent, introspection, empower, november, retroactive, designation, introduction, adaptation, acceptance, exponential, location, electromagnet, abduction, insolvent, rotation, impartial, governmental, explosive, entitle, informant, exposure, resignation, induction, prosaic, absorbent, omega, explosion, carbohydrate, exemption, chiropractic, vocation, extinguish, objective, incisor, impromptu, exhaustion, anecdotal, momentum, alexandrine, indenture, invasion, mosaic, substantial, implosion, ignoble, pyrotechnic, injection, objection, inspection, infliction, dictation, introspective, intrinsic, importance, ingestion, quotation, inexpensive, indignation, expansion, evocation, subjective, diagnosis, obstruction, encounter, impious, ovation,

166

enclosure, venezuela, jurisdiction, hydrocarbon, informal, exception, radioactive, olympic, suggestion, incisive, objector, flotation, extensive, egocentric, commingle, important, extinction, obnoxious, excepted, embroider, expensive, excited, intensive, entire, engender, subscription, invective, incentive, romantic, invasive, inspector, absorption, inviting, notation, venezuelan

012 (count:0)

210 (count:117)

abstraction, angola, foundation, icelandic, forsaken, expectation, privation, clairvoyance, vacation, almighty, dissection, transpire, departmental, migration, abstention, archaic, payola, ornamentation, entangle, bilingual, ambrosia, magnolia, alsatian, filtration, regimentation, embankment, antarctic, iota, bronchitis, implementation, damnation, didactic, clairvoyant, indentation, aorta, foreclosure, taxation, angina, dictator, orientation, laotian, hiatus, mendacious, documentation, deformation, transcendence, translation, stagnation, endanger, dilation, salvation, instrumentation transcendent, antisocial, eccentric, transaction, abnormal, gondola, tangential, plantation, translator, deportation, exploitation, mastoiditis, elongation, gigantic, prognosis, alfalfa, climactic, flamboyant, citation, affectation, ensemble, thrombosis, portentous, sensation, annexation, baptismal, vindictive, symbiosis, tendentious, october, incarnation, transplantation expiry, augmentation, incantation, modernization, arthritis, mongolia, formation, tanzania, temptation, licentious, mongolian, carnation, enlighten, vibration, dalmatian, starvation, exhortation, tanzanian, osmosis, representation, digestion

212 (count:0)

001 (count:0)

021 (count:0)

201 (count:2)

introduce, insincere

221 (count:3)

financier, mozambique, chimpanzee

Pre-antepenult (count:0)

HHH   *100 (count:0)*

102 (count:4)

beaujolais, infiltrate, endocrine, nightingale

120 (count:0)

010 (count:0)

012 (count:3)

incarnate, impregnate, reincarnate

210 (count:6)

volcano, fandango, albino, barbados, embargo, tornado

212 (count:1)

andante

001 (count:0)

021 (count:0)

201 (count:2)

coincide, diagnose

221 (count:0)

Pre-antepenult (count:0)

Experimental Items

The table below lists the items used in Experiments 3, and 4. For the (much longer) list of items used in Experiments 1 and 2, see the author's website. For real words used and their mis-stressed counterparts, see the methods section of Experiments 1 and 3.

The first column contains the IPA transcription of the three syllables presented individually to participants. Participants each heard either the i-final version or the ə-final version of each item. The second and third columns contain the IPA transcription of each stress version which participants heard, and chose between in the forced choice task. The last column contains for each item the consonant that was added to the penultimate syllable to make that syllable heavy in the Heavy condition. Consonants were added as codas to the penultimate syllable. For example, [butʃəli] became [butʃədli] were chosen so that they could not form a legal onset with the following consonant. Items with no listed heavy penult consonant did not appear in Experiment 4.

Nonword	Antepenultimate stress	Penultimate stress	Heavy Penult Consonant
[bu] [ʧu] [li/ə]	[buʧəli/ə]	[bəʧuli/ə]	d
[bɛ] [vɛ] [di/ə]	[bɛvədi/ə]	[bəvɛdi/ə]	k
[ʧɑ] [tɑ] [ri/ə]	[ʧɑtəri/ə]	[ʧətɑri/ə]	v
[dæ] [kæ] [θi/ə]	[dækəθi/ə]	[dəkæθi/ə]	s
[dɛ] [lɛ] [si/ə]	[dɛləsi/ə]	[dəlɛsi/ə]	k
[fɛ] [sɛ] [li/ə]	[fɛsəli/ə]	[fəsɛli/ə]	v
[fæ] [tæ] [si/ə]	[fætəsi/ə]	[fətæsi/ə]	
[kɛ] [bɛ] [li/ə]	[kɛbəli/ə]	[kəbɛli/ə]	
[ɛ] [kɛ] [mi/ə]	[ɛkəmi/ə]	[əkɛmi/ə]	
[æ] [kɹæ] [θi/ə]	[ækɹəθi/ə]	[əkɹæθi/ə]	d
[læ] [dæ] [si/ə]	[lædəsi/ə]	[lədæsi/ə]	p
[læ] [mæ] [si/ə]	[læməsi/ə]	[ləmæsi/ə]	f
[mæ] [ʃæ] [bi/ə]	[mæʃəbi/ə]	[məʃæbi/ə]	d
[pæ] [ʧæ] [li/ə]	[pæʧəli/ə]	[pəʧæli/ə]	d
[pɑ] [dɑ] [mi/ə]	[pɑdəmi/ə]	[pədɑmi/ə]	z
[pæ] [jæ] [ni/ə]	[pæjəni/ə]	[pəjæni/ə]	s
[pæ] [kæ] [ɹi/ə]	[pækəɹi/ə]	[pəkæɹi/ə]	s
[pæ] [læ] [ki/ə]	[pæləki/ə]	[pəlæki/ə]	z
[pæ] [mæ] [di/ə]	[pæmədi/ə]	[pəmædi/ə]	f

Nonword	Antepenultimate stress	Penultimate stress	Heavy Penult Consonant
[ɹæ] [mæ] [ki/ə]	[ɹæməki/ə]	[ɹəmæki/ə]	f
[ɹæ] [mæ] [ni/ə]	[ɹæməni/ə]	[ɹəmæni/ə]	f
[ɹɪ] [nɪ] [mi/ə]	[ɹɪnəmi/ə]	[ɹənɪmi/ə]	f
[æ] [ɹæ] [spi/ə]	[æɹəspi/ə]	[əɹæspi/ə]	f
[ɹɪ] [vɪ] [si/ə]	[ɹɪvəsi/ə]	[ɹəvɪsi/ə]	p
[sɛ] [fɛ] [ni/ə]	[sɛfəni/ə]	[səfɛni/ə]	p
[sɛ] [lɛ] [ki/ə]	[sɛləki/ə]	[səlɛki/ə]	
[sæ] [pæ] [ɹi/ə]	[sæpəɹi/ə]	[səpæɹi/ə]	z
[θæ] [mæ] [ɹi/ə]	[θæməɹi/ə]	[θəmæɹi/ə]	v
[θæ] [næ] [zi/ə]	[θænəzi/ə]	[θənæzi/ə]	f
[tæ] [mæ] [pi/ə]	[tæməpi/ə]	[təmæpi/ə]	g
[tɛ] [pɛ] [di/ə]	[tɛpədi/ə]	[təpɛdi/ə]	f
[vɪ] [zɪ] [ni/ə]	[vɪzəni/ə]	[vəzɪni/ə]	k

Items which appeared in Experiment 4, and not in Experiment 3:

Nonword	Antepenultimate stress	Penultimate stress	Heavy Penult Consonant
[fɑ] [mɑ] [vi/ə]	[fɑməvi/ə]	[fəmɑvi/ə]	k
[nɑ] [dɑ] [vi/ə]	[nɑdəvi/ə]	[nədɑvi/ə]	k
[næ] [gæ] [si/ə]	[nægəsi/ə]	[nəgæsi/ə]	f
[ɹɪ] [lɪ] [ki/ə]	[ɹɪləki/ə]	[ɹəlɪki/ə]	f
[æ] [ɹæ] [spi/ə]	[æɹəspi/ə]	[əɹæspi/ə]	f
[ɹæ] [mæ] [ni/ə]	[ɹæməni/ə]	[ɹəmæni/ə]	f
[ɹɪ] [nɪ] [mi/ə]	[ɹɪnəmi/ə]	[ɹənɪmi/ə]	f

BIBLIOGRAPHY

Adam Albright and Bruce Hayes. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90:119–161, 2003. doi: 10.1016/S0010-0277(03)00146-X.

Jonathan B. Alcántara. *The architecture of the English lexicon*. PhD thesis, Cornell University, 1998.

Arto Anttila. Deriving variation from grammar. In Frans Hinskens, Roeland van Hout, and W. Leo Wetzels, editors, *Variation, Change, and Phonological Theory*, pages 35–68. John Benjamins, Amsterdam, 1997.

R Harald Baayen, Richard Piepenbrock, and Rijn van H. The CELEX lexical data base on CD-ROM. 1993.

Todd M. Bailey and Ulrike Hahn. Determinants of wordlikeness: Phonotactics or lexical neighbors? *Journal of Memory and Language*, 44:568–591, 2001. doi: 10.1006/jmla.2000.2756.

Robert G. Baker and Philip T. Smith. A psycholinguistics study of English stress assignment rules. *Language and Speech*, 19:9–27, 1976.

Robert Bayley. The quantitative paradigm. In J.K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, editors, *The handbook of language variation and change*, pages 117–141. Blackwell, 2002.

Michael Becker. *Phonological trends in the lexicon: the role of constraints*. PhD thesis, University of Massachusetts Amherst, 2009.

Michael Becker and Jonathan Levine. Experigen - an online experiment platform.

Michael Becker, Lauren Eby Clemens, and Andrew Nevins. A richer model is not always more accurate: the case of French and Portuguese plurals.

Michael Becker, Nihan Ketrez, and Andrew Nevins. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, 87(1):84–125, March 2011. doi: 10.1353/lan.2011.0016.

Jill N Beckman. Positional faithfulness, positional neutralisation and shona vowel harmony. *Phonology*, 14(01):1–46, 1997.

Mary E Beckman. *Stress and non-stress accent*, volume 7 of *Netherlands Phonetics Archives*. Walter de Gruyter, Dordrecht, The Netherlands, 1986.

Iris Berent, V. Vaknin, and G. Marcus. Roots, stems, and the universality of lexical representations: Evidence from hebrew. *Cognition*, 104:254–286, 2007.

Iris Berent, Tracy Lennertz, Paul Smolensky, and Vered Vaknin-Nusbaum. Listeners' knowledge of phonological universals: evidence from nasal clusters. *Phonology*, 26: 75–108, 2009.

Jean Berko. The child's learning of English morphology. *Word*, 14:150–77, 1958.

Koen B.E. Böcker, Marcel C.M. Bastiaansen, Jean Vroomen, Cornelis H.M. Brunia, and Beatrice De Gelder. An erp correlate of metrical stress in spoken word recognition. *Psychophysiology*, 36:706–720, 1999. doi: 10.1017/S0048577299971767.

Paul Boersma and Bruce Hayes. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32:45–86, 2001. doi: 10.1162/002438901554586.

Paul Boersma and David Weenink. Praat: doing phonetics by computer [computer program]. In *http://www.praat.org/*, 2011.

Mara Breen, John Kingston, and Lisa D. Sanders. Perceptual representations of phonotactically illegal syllables. *Attention, Perception, and Psychophysics*, 75(1): 101–120, 2013.

Marc Brysbaert and Boris New. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4): 977–990, November 2009. doi: 10.3758/BRM.41.4.977.

Luigi Burzio. *Principles of English stress*, volume 72 of *Cambridge studies in Linguistics*. Cambridge University Press, Melbourne, Australia, 1994.

Henrietta J. Cedergren and David Sankoff. Variable rules: Performance as a statistical reflection of competence. *Language*, 50:333–355, 1974.

Noam Chomsky. *Syntactic Structures*. Mouton de Gruyter, the Hague, 1957.

Noam Chomsky and Morris Halle. *The sound pattern of English*. Harper and Row, New York, Evanston, and London, 1968.

Andries W. Coetzee and Shigeto Kawahara. Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, 31:47–89, 2013.

Andries W. Coetzee and Joe Pater. The place of variation in phonological theory. In John A. Goldsmith, Jason Riggle, and Alan C. Yu, editors, *Handbook of Phonological theory*, pages 401–434. Wiley-Blackwell, 2 edition, 2009.

Sarah Collie. English stress preservation: the case for 'fake cyclicity'. *English Language and Linguistics*, 12(3):505–532, 2008.

Anne Cutler and Sally Butterfield. Rhythmic cues to speech segmentation: evidence from juncture misperception. *Journal of Memory and Language*, 31:218–236, 1992.

Anne Cutler and David M. Carter. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2(3-4):133–142, September-December 1987. doi: 10.1016/0885-2308(87)90004-0.

Anne Cutler and Dennis Norris. The role of strong sylables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14:113–121, 1988.

Walter Daelemans, Steven Gillis, and Gert Durieux. The acquisition of stress: A data-oriented approach. *Computational Linguistics*, 20(3):421–451, 1994.

Ghislaine Dehaene-Lambertz, Emmanuel Dupoux, and A. Gout. Electrophysiological correlates of phonological processing: A cross-linguistic study. *Journal of Cognitive Neuroscience*, 12(4):635–647, 2000.

Ulrike Domahs, W. Kehrein, J. Kraus, R. Wiese, and M. Schlesewsky. Event-related potentials reflecting the processing of phonological constraint violations. *Language and Speech*, 52(4):415–435, December 2009. doi: 10.1177/0023830909336581.

Ulrike Domahs, Ingo Plag, and Rebecca Carroll. Word stress assignment in german, english and dutch: Quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics*, pages 1–38, 2014.

Emmanuel Dupoux, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, and Jacques Mehler. Epenthetic vowels in japanese: A perceptual illusion? *Journal of experimental psychology: human perception and performance*, 25(6):1568, 1999.

David Eddington. Spanish stress assignment within the analogical modeling of language. *Language*, 76(1):92–109, March 2000.

Mirjam Ernestus and Harald Baayen. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79(1):5–38, 2003.

Amy S. Finn and Carla L. Hudson Kam. The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, 108:477–499, 2008.

Edward Garrett. Minimal words aren't minimal feet. In *UCLA Working papers in Linguistics*, volume 1, pages 68–105. UCLA, 1999.

M. Goldrick and M. Larson. Constraints on the acquisition of variation. *Laboratory phonology*, 10:285–310, 2010.

Sharon Goldwater and Mark Johnson. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120, 2003.

Barry Griner. Productivity of japanese verb tense inflection: A case study. Master's thesis, UCLA, 2001.

Susan G. Guion, J.J. Clark, Tetsuo Harada, and Ratree P. Wayland. Factors affecting stress placement for English nonwords include syllabic structure, lexical class, and stress patterns of phonologically similar words. *Language and Speech*, 46(4):403–427, December 2003. doi: 10.1177/00238309030460040301.

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

Morris Halle and Jean-Roger Vergnaud. *An essay on stress*. The MIT press, Cambridge, Massachusetts, 1987.

Bruce Hayes. Extrametricality and English stress. *Linguistic Inquiry*, 13(2):227–276, 1982.

Bruce Hayes. *Metrical stress theory: principles and case studies*. University of Chicago press, 1995.

Bruce Hayes. Blick - a phonotactic probability calculator. Project page at www.linguistics.ucla.edu/people/hayes/BLICK/, 2012. URL http://www.linguistics.ucla.edu/people/hayes/BLICK/CMUDictionaryEditedWithCELEXF

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

Bruce Hayes, Kie Zuraw, Péter Siptár, and Zsuzsa Londe. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85(4):822–863, December 2009.

Bruce Philip Hayes. *A metrical theory of stress rules*. PhD thesis, Massachusetts Institute of Technology, 1980.

Karen Henrich, Kai Alter, Richard Wiese, and Ulrike Domahs. The relevance of rhythmical alternation in language processing: an ERP study on English compounds. *Brain and Language*, 136:19–30, 2014.

Leonard Horowitz, Margeret White, and Douglas Atwood. Word fragments as aids to recall. *Journal of Experimental Psychology*, 76(2):219–226, 1968.

Derek M. Houston, Lynn M. Santelmann, and Peter W. Jusczyk. English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes*, 19(1):97–136, 2004.

Linda D. Jarmulowicz. English derivational suffix frequency and children's stress judgments. *Brain and Language*, 81:192–204, 2002. doi: 10.1006/brln.2001.2517.

Elizabeth K. Johnson and Peter W. Jusczyk. Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, 44:548–567, 2001.

Peter W. Jusczyk, Anne Cutler, and Nancy J. Redanz. Infants' preference for the predominant stress patterns of English words. *Child development*, 64(3):675–687, 1993.

René Kager. *A metrical theory of stress and destressing in English and Dutch.* Foris, Dordrecht - Holland/Providence RI - USA, 1989.

Rene Kager. *Optimality Theory.* Cambridge Textbooks in Linguistics. Cambridge University Press, 1999.

René Kager and Joe Pater. Phonotactics as phonology: knowledge of a complex restriction in Dutch. *Phonology*, 29:81–111, 2012.

Daniel Kahn. *Syllable-based generalizations in English phonology*, volume 156. Indiana University Linguistics Club Bloomington, 1976.

Shigeto Kawahara and Kazuko Shinohara. Phonetic and psycholinguistic prominences in pun formation: Experimental evidence for positional faithfulness. In *Japanese/Korean Linguistics*, volume 18, pages 177–188. Stanford: CSLI, 2011.

Michael H. Kelly and J. Kathryn Bock. Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):389–403, 1988.

Marta Kutas and Steven A. Hillyard. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, January 1980. doi: 10.1126/science.7350657.

William Labov. contraction, deletion, and inherent variability of the English copula. *Language*, 45(4):715–762, 1969.

Willem JM Levelt. *Speaking: From intention to articulation*, volume 1. MIT press, 1993.

Mark Liberman and Alan Prince. On stress and linguistic rhythm. *Linguistic Inquiry*, 8(2):249–336, 1977.

G.F. Marcus and I. Berent. Are there limits to statistical learning? *Science*, 300 (5616):53–55, 2003.

Dominic W. Massaro and Michael M. Cohen. Phonological context in speech perception. *Perception and Psychophysics*, 34(4):338–348, 1983.

John J. McCarthy. Prosodic morphology. ms. University of Massachusetts, Amherst, 1986.

James L. McClelland and Jeffrey L. Elman. The TRACE model of speech perception. *Cognitive Psychology*, 18:1–86, 1986. doi: 10.1016/0010-0285(86)90015-0.

Claire Moore-Cantwell. Over-and under-generalization in morphological learning. In Stefan Keine and Shayne Sloggett, editors, *The proceedings of the 42nd meeting of the North East Linguistics society*, volume 2, pages 41–54, 2012.

Claire Moore-Cantwell and Joe Pater. Gradient exceptionality in maximum entropy grammar with lexically specifc constraints. ms. Yale and University of Massachusetts Amherst, 2015. URL https://drive.google.com/file/d/0BzcIp6QV9VxleER5UG1VekEyM28/view.

Elliott Moreton. Structural constraints in the perception of English stop-sonorant clusters. *Cognition*, 84:55–71, 2002.

Elliott Moreton and Joe Pater. Structure and substance in artificial-phonology learning. *Language and Linguistics Compass*, 6(11):686–718, 2012.

Risto Näätänen. *Attention and brain function*. Psychology Press, 1992.

R.C. Nakisa, K. Plunkett, and Ulrike Hahn. A cross-linguistic comparison of single and dual-route models of inflectional morphology. In P. Broeder and J. Murre, editors, *Models of language acquisition: inductive and deductive approaches.* MIT press, 2001.

Seib Nooteboom. Lexical retrieval from fragments of spken words: Beginnings vs. endings. *Journal of Phonetics*, 9:407–424, 1981.

Robert M. Nosofsky. relations between exemplar-similarity and likelihood models of classification. *journal of mathematical psychology*, 34:393–418, 1990.

Paul Olejarczuk. The productivity and stability of competing generalizations in stress assignment. In *Laboratory Phonology*, Tachikawa, Japan, 2014.

Joe Pater. Against the underlying specification of an'exceptional'english stress pattern. *Toronto Working Papers in Linguistics*, 13, 1994.

Joe Pater. Non-uniformity in english secondary stress: the role of ranked and lexically specific constraints. *Phonology*, 17(2):237–274, 2000.

Joe Pater. Learning a stratified grammar. In Alejna Brugos, Manuella R. Clark-Cotton, and Seungwan Ha, editors, *Proceedings of the 29th Boston University Conference on Languega Development*, pages 482–492, Somerville, MA, 2005. Cascadilla Press.

Joe Pater. Gradual learning and convergence. *Linguistic Inquiry*, 39(2):334–345, 2008.

Joe Pater and Robert Staubs. Harmonic grammar in r, 2015. URL `http://blogs.umass.edu/hgr/hg-in-r/`.

Janet Pierrehumbert. Why phonological constraints are so coarse-grained. This is published in a journal. Look up where, 2001.

Steven Pinker. *Words and Rules: the ingredients of language.* William Morrow, New York, 1999.

Steven Pinker and Alan Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1): 73–193, 1988.

Ilona. Pitkanen. *Electrophysical investigations of second language word learning, attrition, and bilingual processing.* PhD thesis, University of Washington, 2010.

Sandeep Prasada and Steven Pinker. Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8:1–56, 1993.

Alan Prince. Relating to the grid. *Linguistic Inquiry*, 14:19–100, 1983.

Alan Prince and Paul Smolensky. *Optimality Theory: constraint interaction in generative grammar.* Blackwell, Malden, MA, 1993/2004.

Michael D. Rugg. Event-related brain potentials dissociate repetition effects of high- and low- frequency words. *Memory and Cognition*, 18(4):367–379, 1990. doi: 10.3758/BF03197126.

Michael D. Rugg and Margaret E. Nagy. lexical contribution to nonword-repetition effects: evidence from event related potentials. *Memory and Cognition*, 15(6):473–481, 1987. doi: 10.3758/BF03198381.

D. E. Rumelhart and J. L. McClelland. *On Learning the past tenses of English verbs*, chapter 18, pages 216–271. MIT press, 1986.

Bertrand Russell. Knowledge by acquaintaince and knowledge by description. *Proceedings of the Aristotelian Society*, 11:108–128, 1910.

Lisa D. Sanders and Helen J. Neville. Lexical, syntactic, and stress-pattern cues for speech segmentation. *Journal of speech, language, and hearing research*, 43(6): 1301–1321, 2000.

M.S. Seidenberg, J.L. Elman, et al. Networks are not 'hidden rules'. *Trends in Cognitive Sciences*, 3(8):288–289, 1999.

M.S. Seidenberg, M.C. MacDonald, and J.R. Saffran. Does grammar start where statistics stop? *Science*, 298(5593):553–554, 2002.

Elisabeth Selkirk. *Syntax and Phonology: The relation between sound and structure.* MIT Press, Cambridge, Massachusetts, 1984.

Joan A Sereno and Allard Jongman. Acoustic correlates of grammatical class. *Language and Speech*, 3(1):57–76, 1995.

Royal Skousen. *Analogical modeling of language.* Springer, 1989.

Morgan Sonderegger and Partha Niyogi. Variation and change in English noun/verb pair stress: Data, dynamical systems models, and their interaction. In A.C.L. Yu, editor, *Origins of sound patterns: approaches to phonologization*, pages 262–284. Oxford University Press, 2013.

Robert Staubs. *Computational modelingof learning biases in stress typology.* PhD thesis, University of Massachusetts Amherst, 2014.

Richard V. Tescner and M. Stanley Whitley. *Pronouncing English: A stress-based approach with CD-ROM.* Georgetown University Press, Washington, DC, 2004.

Erik D. Thiessen and Jenny R. Saffran. When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental psychology*, 39(4):706–716, 2003.

Sam Tilsen. Utterance preparation and stress clash: Planning prosodic alternations. *Fuchs, S., Perrier, P., Weirich, M. & Pape, D.(Eds.) Speech production and perception: Planning and dynamics*, 2012.

Andrea Tyler and William Nagy. the acquisition of English derivational morphology. *Journal of Memory and Language*, 28:649–667, 1989.

K.I. Vaden and H.R. Hickok. Irvine phonotactic online dictionary, version 2.0. URL `http://www.ipod.com`.

Robert L. Weide. CMU pronouncing dictionary, 1994. URL `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

Christine A. Whalen and Gary S. Dell. Speaking outside the box: learning of non-native phonotactic constraints is revealed in speech errors. In *Proceedings of the 28th annual conference of the cognitive science society*, pages 2371–2374, 2006.

Colin Wilson and Benjamin George. The maxent grammar tool. Department of Cognitive Science, Johns Hopkins University and Department of Linguistics, UCLA (http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/), 2009.

Hideki Zamma. *Patterns and categories in English suffixation and stress placement: a theoretical and quantitative study.* PhD thesis, University of Tsukuba, 2012.

Eugene B. Zechmeister, Andrea M. Chronis, WIlliam L. Cull, Catherine A. D'Anna, and Noreen A Healy. Growth of a functionally important lexicon. *journal of reading behavior*, 7(2):201–212, 1995.

Jie Zhang, Yuwen Lai, and Craig Turnbull-Sailor. Wug-testing the "tone circle" in taiwanese. In Donald Baumer, David Montero, and Michael Scanlon, editors, *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pages 453–461, Somerville, MA, 2006. Cascadilla Press.

Kie Zuraw. *Patterned Exceptions in Phonology*. PhD thesis, University of California, Los Angeles, 2000.

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