PHONOTACTIC GENERALIZATIONS AND THE METRICAL PARSE

by

PAUL OLEJARCZUK

A DISSERTATION

Presented to the Department of Linguistics
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2018

DISSERTATION APPROVAL PAGE

Student: Paul Olejarczuk

Title: Phonotactic Generalizations and the Metrical Parse

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Linguistics by:

Vsevolod Kapatsinski	Chairperson
Melissa A. Redford	Core Member
Melissa M. Baese-Berk	Core Member
Charlotte R. Vaughn	Core Member
Kaori Idemaru	Institutional Representative

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2018

DISSERTATION ABSTRACT

Paul Olejarczuk

Doctor of Philosophy

Department of Linguistics

September 2018

Title: Phonotactic Generalizations and the Metrical Parse

This dissertation explores the relationship between English *phonotactics* – sequential dependencies between adjacent segments – and the metrical parse, which relies on the division of words into syllables. Most current theories of syllabification operate under the assumption that the phonotactic restrictions which co-determine syllable boundaries are constrained by word edges. For example, a syllable can never begin with a consonant sequence that is not also attested as a word onset. This view of phonotactics as categorical is outdated: for several decades now, psycholinguistic research employing monosyllables has shown that phonotactic knowledge is gradient, and that this gradience is projected from the lexicon and possibly also based on differences in sonority among consonants located at word margins. This dissertation is an attempt to reconcile syllabification theory with this modern view of phonotactics.

In what follows, I propose and defend a gradient metrical parsing model which assigns English syllable boundaries as a probabilistic function of the well-formedness relations that obtain between potential syllable onsets and offsets. I argue that this well-formedness is subserved by the same sources already established in the phonotactic literature: probabilistic generalizations over the word edges as well as sonority. In support of my proposal, I provide experimental evidence from five sources: (1) a

pseudoword hyphenation experiment, (2) a reanalysis of a well-known, large-scale hyphenation study using real English words, (3) a forced-choice preference task employing nonwords presented as minimal stress pairs, (4) an online stress assignment experiment, and (5) a study of the speech errors committed by the participants of (4). The results of all studies converge in support of the gradient parsing model and correlate significantly with each other. Subsequent computer simulations suggest that the gradient model is preferred to the categorical alternative throughout all stages of lexical acquisition.

This dissertation contains co-authored material accepted for publication.

CURRICULUM VITAE

NAME OF AUTHOR: Paul Olejarczuk

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Northwestern University, Evanston, IL

DEGREES AWARDED:

Doctor of Philosophy, Linguistics, 2018 University of Oregon
Bachelor of Arts, Psychology, 2000, Northwestern University

AREAS OF SPECIAL INTEREST:

Laboratory phonology, learning theory, speech perception, categorization, second language acquisition, usage-based linguistics

PROFESSIONAL EXPERIENCE:

Graduate Teaching Employee, University of Oregon, 2011-2018

TEFL Instructor, Aichi Prefecture, Japan, 2006-2011

GRANTS, AWARDS, AND HONORS:

Dissertation Research Fellowship, University of Oregon College of Arts and Sciences, 2017-2018

Gary E. Smith Summer Professional Development Award, University of Oregon, 2015

Summer Institute Fellowship, Linguistic Institute of America, 2013

PUBLICATIONS:

Olejarczuk, P. & Kapatsinski, V. (to appear). The metrical parse is guided by gradient phonotactics. To appear in *Phonology*.

Olejarczuk, P., Kapatsinski, V. & Baayen, R.H. (to appear) Distributional learning is error driven: the role of surprise in the acquisition of phonetic categories. To appear in *Linguistics Vanguard*.

Kapatsinski, V., Olejarczuk, P. & Redford, M.A. (2017). Perceptual learning of intonation in adults and 9- to 11-year old children: Adults are more narrow- minded. *Cognitive Science.* doi: 10.1111/cogs.12345

Olejarczuk, P. & Kapatsinski, V. (2016). Attention allocation in phonetic category learning. *Proceedings of the 4th International Forum on Cognitive Modeling*, 148- 156.

Olejarczuk, P. & Redford, M.A. (2013). The relative contribution of rhythm, intonation and lexical information to the perception of prosodic disorder. *Proceedings of Meetings on Acoustics,* 19. doi: 10.1121/1.4800625

ACKNOWLEDGMENTS

Looking back over the years, I am indebted to a number of people without whom this project would not have come to fruition. All the words I can muster cannot express the depth of my gratitude, but they will have to suffice.

First and foremost, I would like to thank my adviser, Volya Kapatsinski, for providing invaluable guidance over the years and for encouraging me to pursue all of my ideas, no matter how half-formed they happened to be at the time. I am also grateful to my other committee members – Lisa Redford for her early support, intellectual rigor and professional advice, Melissa Baese-Berk for dispelling many doubts with her constant encouragement, Kaori Idemaru for the detailed notes on an earlier draft of this dissertation, and Charlotte Vaughn for reminding me to stop and hear the music every now and then.

I would also like to thank my fellow members of the Usage-Based Linguistics Lab – Zara Harmon, Amy Smolek and Hideko Teruya – for sharing the rollercoaster of successes and failures that is experimental linguistics, and for providing a much needed outside perspective on my work.

Early stages of this project (and all my other work) benefitted greatly from the feedback I received at Eric Pederson's Cognitive Linguistics Workgroup. I would like to thank all of the current and former group members over the years for their critical eyes and helpful advice, including Danielle Barth, Ted Bell, Wook-kyung Choe, Charlie Farrington, Jeff Kallay, Misaki Kato, Jason McLarty, Shahar Shirtz, Matt Stave, Amos Teo, Julia Trippe and Wan Vajrabhaya.

This has been a journey of personal as well as academic growth, and for that I am most thankful to my peers, whose friendships have sustained me through it all. I

would like to extend particular thanks to Manuel Otero for all the exploded cigars, the Taco Tuesdays (on Wednesdays), and for being the best cohort mate I could have asked for, to Becki Quick for sharing her home and for putting up with all the gentlemen's dinners, and to Ted Adamson, Wolfgang Barth, Krishna Boro, Brian Butler, Jaime Peña and Marie Pons for the various ways in which they've helped me get through the program.

Last but not least, I am thankful to my mother for crossing the Pond all those years ago and for making all of the sacrifices that made my life possible.

to the memory of my grandparents, Zofia and Kazimierz

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

1.1 The Metrical Parse

One of the hallmarks of human languages is hierarchical structure: elements combine to make larger units, which in turn form even larger constituents. For example, morphemes fuse to form words, words combine into phrases, and phrases can function as parts of larger phrases or clauses. Like morphosyntax, prosody has also been argued to feature hierarchical organization by a number of phonologists (Beckman & Pierrehumbert, 1986; Gussenhoven, 1992; Hayes, 1989a; Liberman, 1975; Nespor & Vogel, 1986; Selkirk, 1978). Consider the prosodic hierarchy illustrated in Figure 1.1, adapted from Shattuck-Hufnagel & Turk (1996).

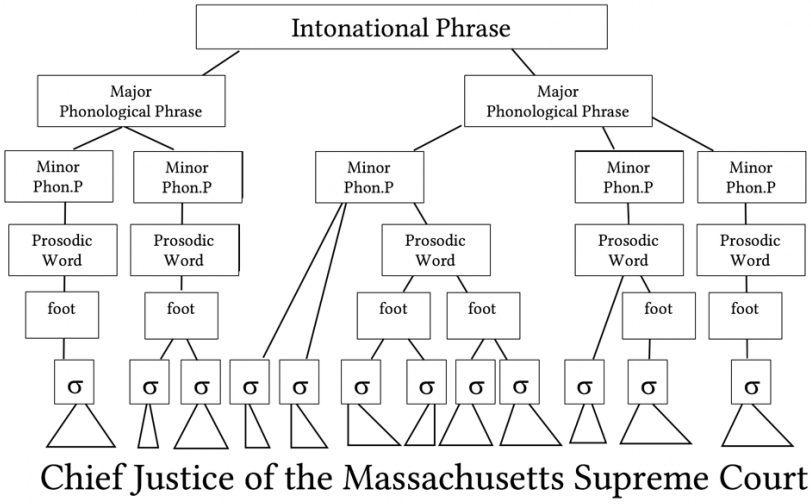

Figure 1.1. Composite prosodic hierarchy based on Beckman & Pierrehumbert (1986), Hayes (1989a), Nespor & Vogel (1986) and Selkirk (1978). Adapted from Figure 3 in Shattuck-Hufnagel (1996:206).

At the top of the hierarchy is the Intonational Phrase, which is the largest stretch of speech produced under a coherent intonational contour (intonation contours are identified by the presence of nuclear accents and boundary tones). The Intonational Phrase may be further subdivided into Major Phonological phrases, which are the domain of phrasal stress and tend to align with syntactic constituents (though, as Shattuck-Hufnagel & Turk (1996) emphasize, the syntax-prosody mapping is not isomorphic). Minor Phonological Phrases contain a single content word along with any cliticized function words. Immediately below this level is the Prosodic Word, which may correspond to either any lexical word or to content words only, depending on the theory (e.g. Hayes, 1989a; Inkelas & Zec, 1993). Prosodic words are made up of feet, which constitute the domain of lexical stress. Finally, each foot may contain one or two syllables, or groups of segments arranged around a single vocalic nucleus.

This dissertation is concerned with the bottom part of the prosodic hierarchy: syllables, and to a lesser extent, feet. Specifically, I investigate the way in which English speakers divide novel strings into syllables. This process is conventionally called *syllabification*; while I adhere to this convention when reviewing prior literature, I also refer to the process as the *metrical parse*. This terminological choice was motivated by the fact that the most compelling evidence I present in support of my argument comes from the metrical phenomenon of stress assignment; the label thus recognizes the syllable's role in the hierarchy of prosodic prominence.

1.2 Contribution of this Dissertation

The goal of this dissertation is to explore the relationship between the English metrical parse and *phonotactics* — sequential dependencies between adjacent segments. The vast majority of phonological theories recognize the syllable as the proper domain of phonotactic restrictions and acknowledge the role of syllable structure in metrical phenomena. This position is nicely summarized in Selkirk (1982):

> First, it can be argued that the most general and explanatory statement of phonotactic constraints in a language can be made only by reference to the syllable structure of an utterance [...] And third, it can be argued that an adequate treatment of suprasegmental phenomena such as stress and tone requires that segments be grouped into units which are the size of the syllable. (p.19)

To paraphrase this well-established view, phonotactics are involved in shaping syllable structure, which in turn determines the placement of stress and tone in a number of languages. The majority of phonologists thus recognize a relationship between phonotactics and the metrical parse.

What is the exact nature of this relationship? Virtually all prior syllabification theories assume a particular kind of phonotactic model which relies on categorical restrictions on word margins. All else being equal, this model constrains possible syllable edges to the set of attested word edges, so that a syllable nucleus cannot be surrounded by consonants (or consonant sequences) which do not also end and begin words in the language in question. Thus, the English word *atlas* invariably syllabifies as *at.las* and never as *a.tlas* or *atl.as* because /tl/ is not an attested word margin. On the other hand, the name *Austin* contains a medial cluster /st/ which is perfectly legal as a

word onset or offset (*st*one, *August*). Words like *Austin* motivate the inclusion of other, non-phonotactic influences on syllabification, leading to much disagreement among phonologists (see section 2.2.1 for a review of the relevant literature). Such words also elicit relatively high uncertainty in syllable division tasks performed by native speakers in laboratory settings (section 2.2.2).

Curiously, the last three decades have seen the rise of a different view of phonotactics, one which casts the well-formedness of a phoneme string not in terms of categorical prohibitions against certain sound sequences, but rather as a continuum projected from lexical statistics and other factors. Under this modern, granular view, nonsense strings beginning with unattested word onsets may nevertheless differ in relative grammaticality (e.g. *dlonk* may be more grammatical than *ldonk*), and the same holds for nonwords with attested onsets (*dronk* may be better than *dwonk*). Experimental support for the gradient view of phonotactics has been abundant (see section 2.3), leading to its widespread adoption. However, perhaps because much of this support has come from studies employing monosyllables, it has gone largely unnoticed by syllabification models. In other words, virtually all extant metrical parse theories operate under outdated phonotactic assumptions.

In this dissertation, I attempt to reconcile these two areas of phonology — syllabification and phonotactics — by proposing and defending a probabilistic parsing model. This metrical parser, operationalized as a multiple regression model, relies on gradient well-formedness relations that obtain between different syllable onsets and offsets. I argue that this well-formedness is subserved by the same sources already established in the phonotactic literature: probabilistic generalizations over the lexicon as well as certain phonetic properties. The model can handle words like *atlas* and

Austin under a unified phonotactic analysis, and accurately predict human parsing behavior.

The basic idea is that syllable boundaries are not deterministically assigned with reference to categorical phonotactics. Instead, the parser is stochastic: the probability of a boundary location in a VC(C)V sequence is modeled as a function of the cumulative well-formedness of the different candidate onsets and codas produced under alternative parses. Support for the model is provided in five different experimental studies employing a range of methods, including hyphenation and stress assignment.

As will be made clear in the next chapter, proper understanding of syllabification has profound consequences for phonological theory because the syllable has played a central role in accounting for allophone distributions, metrical phenomena and many other phonological processes. It also has consequences for psycholinguistic models of speech production and perception, many of which incorporate the syllable as a unit of representation. The findings presented in this dissertation will demonstrate that gradient phonotactics influences intuitions about sublexical units, that they matter during online speech perception and production of stress, and that they have consequences for nonword production accuracy. It will be argued that syllables are best understood as emergent, probabilistic generalizations over word-edges (guided also by certain phonetic properties), that the phonological grammar itself is a system of interacting and competing generalizations over the lexicon, and that, consistent with the modern view of phonotactics, gradient phonotactic knowledge permeates many aspects of linguistic behavior.

1.3 Overview of this Dissertation

This dissertation is organized as follows. In chapter 2, I introduce the basic notion of the syllable, summarize the major theoretical and experimental arguments with respect to syllable division, review the current state of phonotactic theory, and make explicit the gradient metrical parse hypothesis. Chapter 3 follows with a brief overview of the experiments, a description of the database used to calculate lexical statistics, as well as descriptions of the stimuli used in the studies and the predictors included in the statistical models. Chapter 4 contains two hyphenation studies: an original experiment employing pseudowords and a reanalysis of Eddington et al. (2013a,b), which used real English words and a slightly different method. The findings consistently support the gradient phonotactic parser over a categorical alternative. In chapter 5, I present three studies which rely on a novel method of inferring syllable boundaries from stress placement. The tasks involve well-formedness judgments, online stress assignment and production errors. Once again, the results are in favor of the gradient parsing model. Chapter 6 offers further evidence by presenting various correlations between the results of the five studies. At the level of syllabifying unique intervocalic consonants and clusters, all correlations are statistically significant, with the correlation coefficients reaching as high as .86. In chapter 7, I explore the viability of the gradient parser as a learnable model by simulating its acquisition. The results suggest that, in spite of its somewhat greater complexity, unbiased learners should prefer the gradient model to the categorical model regardless of vocabulary size. Chapter 8 offers some concluding remarks and directions for future research. Portions

of the data presented in chapters 4, 5 and 7 will appear in a journal article coauthored by Vsevolod Kapatsinski.

CHAPTER II

THEORETICAL BACKGROUND

2.1 The Many Faces of the Syllable

The syllable is at once one of the oldest ideas in linguistics and one of the most controversial. In modern phonological theory, the sources of controversy are two-fold. The first concerns internal structure: most scholars agree that a syllable consists of some arrangement of consonantal elements around a single vocalic peak, but the exact nature of the arrangement varies widely among the theories (Clements & Keyser, 1983; Davis, 1985; Fudge, 1969; Hayes, 1989b; Hyman, 1985; Kahn, 1976; Pike & Pike, 1947; Selkirk, 1982; Yi, 1999, *inter alia*). The details of this debate are beyond the present scope (see van der Hulst & Ritter, 1999 for a comprehensive survey); a few of the more influential proposals for the internal structure of a CVC sequence are illustrated in (2.1).

(2.1) a. $[C\ V\ C]_{\sigma}$ b. $[C_{onset}\ [V_{nucleus}\ C_{coda}]_{rime}]_{\sigma}$

c. $[[C_{onset}\ V_{nucleus}\ ]_{body}\ C_{coda}\ ]_{\sigma}$ d. $[\ C\ [V]_{\mu}\ [C]_{\mu}\ ]_{\sigma}$

The flat view seen in (a), which connects all elements directly to the syllable node, is assumed in Kahn (1976) and supported in Davis (1985). Of the remaining, hierarchical views, the onset-rime model in (b) is widely accepted for English (Fudge, 1969; Kapatsinski, 2009; Treiman, 1983), the body-coda model (c) has been proposed for Korean (Lee, 2006; Yi, 1999), and the hybrid mora (μ) model (d) has been influential in

accounting for weight-sensitivity in tone and stress systems (Hyman, 1985; Hayes, 1989b; see section 5.1 for more details).

In this dissertation, I am ambivalent about the question of internal structure, focusing instead on the second, related controversy: that of syllabification. The division of words into syllables — in particular, the affiliation of intervocalic consonants — has long been an area of dispute among phonologists and psycholinguists (Eddington et al., 2013a,b; Fujimura & Lovins, 1977; Gussenhoven, 1986; Hammond, 1999; Hoard, 1971; Kahn, 1976; Pulgram, 1970; Redford & Randall, 2005; Selkirk, 1982; Treiman & Danis, 1988; Vennemann, 1972). In section 2.2, I survey the theoretical positions and review experimental evidence that bears on this question. In the remainder of this section, I briefly discuss the syllable's importance in phonological theory, its elusive phonetic correlates, and its controversial status in psycholinguistics.

2.1.1 The Syllable's Utility in Phonological Theory

Although phonologists argue about its nature, the majority would agree that the notion of the syllable makes their job easier. As a sublexical constituent, the syllable has proved a useful tool in the description of a number of phonological processes and phenomena otherwise difficult to capture in a formally elegant way. A complete survey of these phenomena is beyond the scope of this dissertation; here, I briefly mention three areas of phonology where the syllable's utility is perhaps most recognized. These areas are phonotactic restrictions, the distribution of allophones, and metrical phenomena.

In phonotactic theory, a major goal is to capture generalizations about possible sound combinations occurring within words. In his groundbreaking dissertation, Kahn (1976) argues that constraints on English medial consonant sequences are best understood in terms of combinations of possible syllable edges. For Kahn, a form like **atktin* is not a possible English word because the sequence /tkt/ cannot exhaustively syllabify into a valid onset and coda (Kahn, 1976:57). Under his theory, valid syllable edges are constrained by attested word edges: since /tk/ cannot end a word and /kt/ cannot begin one, they cannot appear as margins of internal syllables (see also Kuryłowicz, 1948; Pulgram, 1970). This analysis is more elegant than a syllable-free alternative, which would have to posit constraints against /tk/ in the context of a following /t/ or a word boundary — two environments that do not form a natural class. Furthermore, the relationship between syllable edges and word edges captures the well-formedness of medial sequences in unattested but possible words like *atklin* and *atquin*. Kahn argues that a syllable-free analysis would find such cases to be accidental. This view of the syllable as the domain of phonotactic restrictions is not universally held — for example, Steriade (1999) and Blevins (2003) argue for phonotactics as purely sequential constraints — but it remains the dominant view in phonology (see Goldsmith, 2011).

In addition to constraining phoneme sequences, Kahn (1976) pointed out that syllable structure appears to condition the distribution of English allophones: sounds are often pronounced differently depending on whether they occur in the onset or the coda. For example, the stops /ptk/ tend to be aspirated in syllable-initial position ([ə.ˈpʰiɹ], [tʰə.ˈmɔ.ɹoʊ̯], [ə.ˈkʰɔɹd]) but may be unreleased or glottalized in syllable final

position (['ʃɹæpˀ.nəl], [ætˀ.ləs], [ækˀ.ni]; see also Gussenhoven 1986; Hall, 2004; Pike, 1947).

Finally, metrical phenomena like tone and stress have also been argued to be best understood with reference to the syllable (Gordon, 1999; Hayes, 1980, 1995; Selkirk, 1982; Watkins, 1984). In languages with both level and contour tones, the former are often less restricted while the latter might only fall on syllables which pass a certain size threshold (Zhang, 2002). In languages with quantity-sensitive stress, the location of stress is likewise dependent on syllable structure (Gordon, 1999). For example, the Dutch stress system has been analyzed as differentiating between closed and open syllables: the former count as heavy and attract stress, while the latter count as light and typically do not (van der Hulst, 1984; Kager, 1989). This dependence of stress on syllable structure entails a directionality: stress assignment requires a metrical parse, the first step of which involves syllabification (see section 5.1 for details). Directionality is inherently captured by derivational phonology, where surface forms are taken to be outputs of ordered rules. It can also be captured in constraint-based approaches that allow sequential processing, such as Harmonic Serialism (McCarthy, 2010). The assumption that stress assignment is preceded by syllabification is a staple of metrical theories that focus on weight sensitivity (e.g. Hayes, 1995) and is adopted in this dissertation.

Taken together, sequential constraints, allophone distributions and stress patterns join many other phenomena in arguing for the inclusion of the syllable into the system of abstract, formal representations in phonology. That said, ever since Linguistics declared itself a branch of Cognitive Science in the 1950s, any theoretical claim about language in essence became a claim about the nature of the human mind. In

other words, formal representations like the syllable were assumed to have mental analogues. Searching for behavioral evidence for these representations became (and continues to be) a major goal of the psycholinguistic enterprise. Here, the psychological reality of the syllable — along with its role in mediating various linguistic behaviors — has been somewhat more controversial. In the remainder of this section, I review a small portion of the work on the syllable's role in speech production and processing.

2.1.2 The Syllable as a Unit of Articulatory Organization and Planning

One of the earliest definitions of the syllable was articulatory in nature. Goldsmith (2011) credits Whitney (1874) for introducing what later became known as the *sonority* approach to the syllable — the idea that speech is organized as a series of amplitude peaks and valleys which roughly correspond to the degrees of vocal tract stricture imposed by the movements of the jaw and tongue. For Whitney, a syllable was defined as a sublexical chunk that was produced by a 'single effort or impulse of the voice' (1874:291). This idea was much later taken up by Stetson (1951), who argued that the effort involved was pulmonary — the production of each syllable was hypothesized to be independently controlled by the intercostal muscles, resulting in pulses of forced expiration (see also Pike, 1947). This view did not survive long; subsequent work found no correlation between muscle activity and syllable production (Draper, Ladefoged & Whitteridge, 1959). In fact, much of the early work found little evidence for articulators conspiring to effect clear, observable boundaries at the sub-lexical level. For example, anticipatory lip rounding has been observed to occur across syllable and even word boundaries (Daniloff & Moll, 1968). Such findings led to general pessimism about the

syllable as a physiological unit, and to the emergence of the view that speech is not simply a concatenation of discrete, syllable-sized motor plans.

In subsequent work, the search for discrete boundary events was abandoned in favor of a more holistic approach which sought to associate syllabic position with different intra- and inter-articulator patterns during consonant production. Here, the findings appear to be more promising. In her extensive review of the relevant literature, Krakow (1999) offers the generalization that, relative to codas, segments in onset position tend to be hyperarticulated — produced with tighter degree of constriction and less articulatory variability. For example, in their X-ray imaging study of American English /l/, Giles & Moll (1975) found that the allophone in initial position featured a tighter palatal constriction than the coda variant. Looking at the relative timing of the tongue tip and dorsum during the production of initial and final /l/, Browman & Goldstein (1995) found dorsum retraction to be synchronized with the end of tip raising in initial /l/ and with the beginning of the tip gesture in final /l/. Thus, to the extent that the syllable is involved in organizing speech, its effects may be subtle and indirect. Further confounding the interpretation of these findings is the fact that the majority of the studies used monosyllabic words as stimuli, making it difficult to disentangle syllable-level from word-level effects.

In addition to articulatory investigations, some evidence for the role of the syllable in speech planning comes from various psycholinguistic paradigms. Tip-of-the-tongue phenomena have shown that, even when speakers are unable to access an intended word form, they are nevertheless aware of the number of the syllables it contains (Brown, 1991). In her classic survey of speech errors, Fromkin (1971) argued that segmental exchanges respect syllable position — onsets are swapped for other

onsets and codas for other codas (*stress and pitch* ⟶ *piss and stretch*), but cross-swapping between these constituents is rarely attested. Fromkin suggested that such errors provide evidence for the psychological reality of units like syllables and their internal constituents. However, Shattuck-Hufnagel (1992) noted that the vast majority of the reported exchange errors occur between monosyllabic words. After conducting a series of experiments investigating the influence of word position, syllable position and stress on speech errors, she concludes that it is the word rather than the syllable that provides the frame for serial ordering of segments during production.

Nevertheless, some of the most influential models of speech production have incorporated the syllable (e.g. Dell, 1986; Levelt, 1989). For example, Levelt (1989; 1992) employs the notion of a mental syllabary introduced in Crompton (1982) — a repository of motor programs which can be retrieved during the production of frequent syllables in the speaker's language. Evidence for the syllabary comes mainly from naming latency studies. For instance, Levelt & Wheedon (1994) showed that, after controlling for overall word frequency, words consisting of frequent syllables were repeated faster by Dutch speakers than words made up of rare syllables. Since online computation of syllables should be insensitive to frequency effects, this finding was interpreted as evidence for the retrieval of stored gestural scores (see also Cholin, Levelt & Schiller, 2005; Cholin & Levelt, 2009).

Support for the syllable in speech planning appears to vary with the language under investigation. On one hand, Ferrand & Segui (1998, Experiment 2) report a robust naming latency effect in French: after reading a series of 'inductor' words with uniform syllable structure, speakers respond faster when the name of the subsequent picture shares this structure than when it does not. On the other hand, Croot & Rastle (2004)

found very limited evidence for syllable frequency effects in English. Upon reviewing the literature on frequency effects in production, Shattuck-Hufnagel (2011) speculates that the lack of robust priming effects in English may be due to the 'blurry' nature of English syllable boundaries, and the idea that it is the foot rather than the syllable that might be the relevant unit in this language. This notion of blurry boundaries is central to the present dissertation.

2.1.3 The Syllable as a Unit of Perception and Processing

There is also some psycholinguistic evidence for the role of the syllable in speech perception and spoken word processing. Like in production, however, the findings are mixed and controversial, and seem to depend on the experimental task and the language under investigation. One form of evidence in favor of the syllable as a unit of perception comes from illusory vowels reported by Japanese listeners in a study by Dupoux et al. (1999). Japanese syllables are mostly restricted to CV structure; when presented with pseudowords like *ebzo*, Japanese listeners reported hearing an epenthetic /u/ between the two medial consonants at much higher rates than French listeners, whose native language features many more closed syllables. In my own unpublished work, I have found a related effect in English. When presented with sCV sequences where the second consonant is a voiced stop (e.g. [sbɛ]), listeners often reported hearing [spɛ], perceptually repairing the sequence to conform with English phonotactics. The effect disappeared when these same strings were prepended with vowels: voiced stops in sequences like [ɛsbɛ] were perceived veridically. One way to interpret this finding is to say that the voicing mismatch in the longer sequences cued a

syllable boundary ([ɛs.bɛ]), which obviated the need for perceptual repair (English allows voicing dissimilation of this sort, although it is rarely attested within morphemes).

A great deal of research has investigated the role of the syllable in pre-lexical segmentation of spoken words. The effort was initiated by Mehler et al. (1981), who discovered that, when asked to detect sound sequences inside words, French listeners were faster when those sequences corresponded to syllables in the words (e.g. given the word *balance*, *ba* was identified faster than *bal*, and the opposite held for the word *balcony*). This finding prompted the authors to suggest that the syllable was a unit of processing important for lexical access. Subsequent work employing other tasks like phoneme detection has also found a robust syllable effect in French (Dupoux, 1994; Pallier et al., 1993). However, the effect was much less robust for listeners of other languages (see Frauenfelder & Kearns, 1996). In particular, Cutler et al. (1986) failed to replicate the Mehler et al. (1981) results with English, as did many subsequent studies (see Cutler, 1997 for a review of this work). This failure has prompted Cutler et al. (1986) to hypothesize that syllabic segmentation is inefficient in stress languages like English, where stress-based cues to segmentation are easier to learn than the cues provided by the relatively large inventory of syllable structures.

A different perspective to that of Cutler et al. (1986) is offered in Bruck, Caravolas & Treiman (1995). These authors used a comparison task where participants were presented with pairs of nonwords and asked to determine whether the two pair members began with the same sequence of sounds. Restricting the sequence length to three phonemes across all trials, the responses were faster when the initial sounds formed a complete syllable ([kɪp.kæst] ~ [kɪp.bɛld]) than when they formed only part of

a syllable ([flɪg.mɪl] ~ [flɪk.boz]). This lead Bruck et al. (1995) to suggest that the participants were comparing syllabified representations of the nonwords. Under this strategy, items sharing entire syllables benefitted because the syllable was hypothesized to constitute a processing unit, speeding up the comparison. To explain the disparity of their results and those of Cutler et al. (1986), the authors further argued that, unlike the monitoring task, the nonword comparison task placed a burden on phonological memory because the first pair member had to be retained for comparison. The authors suggested that the storage and maintenance of nonwords in working memory may differ from the activation of real words in the lexicon (as in the monitoring task), with the former process relying more heavily on syllabic representations.

Kapatsinski & Radicke (2009) suggest a possible methodological reason why Cutler et al. (1986) were unable to find a syllable effect in English. Namely, many of the stimuli used in that study featured postvocalic sonorants near the putative syllable boundary (as in *balance*, *balcony*, etc.). The syllabic affiliation of English sonorants is less clear (see section 2.2 below), possibly making it difficult to identify boundaries during processing. In this dissertation, I will demonstrate that probabilistic syllabification applies to *all* intervocalic consonants, not only sonorants. The findings will provide a plausible explanation for the inconsistent results of syllable monitoring tasks.

2.1.4 Summary

To sum up, the syllable has proven to be both indispensable and controversial among researchers interested in understanding the mental representation of sound

structure. On the one hand, many phonologists rely on it to explain within-language sound processes and typological generalizations. On the other, phoneticians have struggled with discovering clear acoustic and articulatory correlates, while psycholinguists have had difficulty defining its exact role in production and perception. To many early phonologists in the generative tradition, phonetic and psycholinguistic evidence was irrelevant because they viewed the syllable as an abstract unit whose existence is entirely justified on phonological grounds. Kahn (1976) falls into this camp, claiming that it is unfair to ask the phonologist for physiological proof of the syllable because speech production necessarily obscures underlying phonological units. More recently however, the phonological landscape has shifted toward informing theory with experimental evidence. For instance, Hammond (1995) notes that, "All else being equal, we would hope that the syllables manipulated in processing to be the same as those motivated on linguistic grounds" (p. 9). In other words, theoretical phonologists have begun to take psycholinguistic studies more seriously.

This dissertation follows the latter tradition, where behavioral patterns observed in the laboratory must have theoretical consequences. In this case, the behaviors in question consist of performance on various experimental tasks that are subserved by the metrical parse, and the consequences entail modifying metrical theory to accommodate probabilistic syllabification. As stated above, this modification will help explain the controversial status of the syllable in production and perception. In the following section, I review prior theoretical work on syllabification, most of which assumed that syllable boundaries are assigned deterministically rather than variably.

2.2 Syllabification

2.2.1 Syllable Division: Theoretical Views

The nature of syllable division has been one of the most controversial areas of phonology. While most researchers agree that each syllable contains a nucleus (usually a vowel or sometimes a sonorant), the affiliation of intervocalic consonants has been hotly disputed. Broadly speaking, there are two major classes of syllabification theories: those which deterministically assign each segment to one syllable only, and those which allow for segments to be *ambisyllabic* — that is, to belong to more than one syllable. Within each type of theory, there are many disagreements; here, I briefly highlight a few of the more influential views that have a bearing on the present study. All of these have been mainly motivated by phonological evidence of the kind discussed in section 2.1.1 (allophone distributions, etc.).

Pulgram (1970) is an early, influential theory where syllable assignment proceeds in a series of ordered rules which essentially constrain syllable margins to attested word margins. Briefly stated, the initial boundaries are placed immediately after each vowel in a string. If the vowel cannot appear in word-final position or the postvocalic consonant(s) cannot begin a word, the boundary is incrementally shifted to the right until the maximal possible word onset is achieved. In the event that it is impossible to achieve both a well-formed onset and a well-formed coda, the latter must bear the irregularity. Pulgram's system is thus a deterministic parsing theory which relies heavily on the 'identity of word-terminal and syllable-terminal phonotactics' (1970:309). This relationship between syllable edges and word edges is assumed in some

form by most subsequent theories. For example, Vennemann (1972) reformulates it into the Law of Initials and Law of Finals (like Pulgram, Vennemann gives priority to the former). In some cases, however, priority is given to the coda; for example, Hammond (1999) syllabifies English intervocalic /l+stop/ sequences with the preceding vowel (e.g. Vlt.V) because the same restrictions on the VlC sequence hold word-medially and word-finally.

It is important to point out that the relationship between English word and syllable margins is asymmetrical in another sense: while most researchers agree that medial onsets and codas must be attested at word edges, it is not the case that all consonant sequences permitted at the ends and beginnings of words can be associated with syllables (Fujimura & Lovins, 1977; Kaye, Lowenstamm & Vergnaud, 1990). For example, complex word offsets such as those in *desks* and *strengths* do not appear word medially, and the coronal obstruents in *homes* and *jumped* are phonetically less coarticulated with the preceding vowel than other consonants. Generally speaking, the inventory of attested medial clusters in English is grossly underpredicted by the cross-product of attested word onsets and offsets (Pierrehumbert, 1994). In addition, there are arguments against /s+stop/ clusters, which begin many English words, as constituting sub-syllabic constituents (e.g. Kaye et al., 1990). Like the final coronals in the examples above, the initial /s/ in these words is sometimes seen as an 'appendix' which attaches directly to the word node rather than to the intermediate onset node. Nevertheless, the so-called Legality Principle preventing unattested word edges from constituting legal syllable edges holds for most scholars.

In addition to word-edge phonotactics, a number of other influences on syllabification have been proposed. By far the most influential of these is the notion of

sonority. Sonority is often defined as an abstract, scalar property of segments that roughly correlates with loudness (Parker, 2002). Generally speaking, vowels feature the highest sonority, followed by glides, liquids, nasals and obstruents (Clements, 1990). Cross-linguistically, syllables tend to rise in sonority from edge to nucleus, with rises preferred through onsets and falls favored through codas. For example, in languages that permit complex onsets, obstruents are generally featured on the periphery, with sonorants closer to the vowel. This typological generalization has been formalized as the Sonority Sequencing Principle (SSP; Jespersen, 1904; Selkirk, 1982; Sievers, 1881). According to the SSP, rising-sonority onsets are universally preferred over falling-sonority onsets. Accordingly, a number of theories rely primarily on the SSP in building the syllable, augmenting it with language-specific constraints (including word-edge phonotactics) to handle sonority violations (e.g. Clements, 1990; Hooper, 1976; Kiparsky, 1979; Murray & Vennemann, 1983). The nature and psychological reality of sonority are controversial. Some researchers propose that the SSP is innate and synchronically active, directly involved in adjudicating the relative well-formedness of unattested syllable onsets (Berent et al., 2007; 2009). Others claim that sonority is phonetically grounded in perception or production (Parker, 2002; Redford, 2008; Wright, 2004). Daland et al. (2011) argue that sonority-based preferences can be viewed as another case of lexical support, at least for English speakers: as long as the learner is allowed to generalize over phonological features and the feature system explicitly represents sonority, relevant similarities between natural classes will be captured and well-formedness asymmetries will fall out from the lexicon. In this dissertation, I adopt the epiphenomenal/lexicalist view of sonority.

Another hypothesized influence on syllabification is stress. In some theories, stress and phonotactics entirely determine the placement of syllable boundaries. For example, Hoard (1971) argues for maximizing the legal onsets of stressed syllables only. For others, stress leads to adjustments of boundaries previously determined on phonotactic and/or sonority grounds (e.g. Hooper, 1978; Kahn, 1976; Selkirk, 1982). For example, Selkirk (1982) argues that intervocalic consonants are initially (at the level of 'deep structure') syllabified into onsets but may be resyllabified (at the 'surface level') as codas if the preceding vowel is stressed. The relationship between stress and syllabification is thus complicated. On the one hand, stress has been argued to determine or shift syllable boundaries. Evidence for this view comes from *perception* experiments employing or studies using real words: when the stress pattern is perceived or known, it can influence judgments of boundary locations (e.g. Eddington et al., 2013a,b; Redford, 2008; see section 4.3.1 for details). On the other, recall from section 2.1.1 that, in *production*, weight-sensitivity requires syllable structure to precede stress assignment. This idea is supported by a number of production experiments or studies employing pseudowords, whose stress patterns are not stored in the lexicon (see section 5.1 for a review). With the exception of Study 2 (a re-analysis of Eddington et al., 2013a,b), this dissertation employs pseudoword stimuli and focuses largely on weight-sensitivity. Given this design, stress is treated here as the outcome of (rather than an influence on) syllabification[1], and will be argued to constitute a major piece of evidence for the emergent nature of syllable-like units.

[1] Nevertheless, it seems clear that, at least in cases where weight-sensitivity is irrelevant and stress information is present in the signal, English listeners use stress as a boundary cue.

Researchers also differ about the interaction of phonological and morphological influences on syllabification. For Pulgram (1970), syllable division is strictly phonological, so that onsets are maximized even across morphemes (see also Kahn, 1976). In contrast, Selkirk's (1982) account requires the final stage of the derivation to align syllable boundaries with morpheme boundaries.

A number of researchers allow for ambisyllabicity of intervocalic consonants. For some, ambisyllabicity is conditioned by stress; for others, it's a part of core syllabification. For example, Trager & Bloch (1941) argue that, in English VCV sequences with stress on the first vowel (as in *hitting*, *pudding*, etc.), the intervocalic consonant belongs to both syllables (or the boundary is inside the segment). For Kuryłowicz (1948), medial consonant sequences can be shared by both vowels to the extent that they form legal word onsets and offsets; the only exception is the final consonant in the sequence, which belongs exclusively to the following vowel. Anderson & Jones (1974) also allow for overlap whenever permitted by word-edge phonotactics. Like Pulgram (1970), Kahn (1976) maximizes onsets on the first pass but allows ambisyllabicity to arise due to subsequent adjustments based on stress and speech rate. For example, the medial consonant in *city* is initially an onset to the second syllable, but the stressed first syllable forms an ambisyllabic association (Kahn takes the flap allophone of /t/ as evidence of its syllabic overlap). In fast speech, Kahn (1976) also allows resyllabification across word boundaries, so that vowel-initial words can gain onsets by sharing the final consonant of the preceding word. Extending Kahn's work, Gussenhoven (1986) relies heavily on ambisyllabicity to account for a number of allophones of British and American English stops.

2.2.2 Syllable Division in English: Experimental Evidence

A long line of research has probed the psychological reality of various theoretical claims about syllable structure by examining how speakers chunk words into smaller units (Berg & Niemi, 2000; Content, Kearns, & Frauenfelder, 2001; Eddington, Treiman, & Elzinga, 2013a; Fallows, 1981; Goslin & Frauenfelder, 2001; Pierrehumbert & Nair, 1995; Redford, 2008; Redford & Randall, 2005). The methods employed in these studies can be roughly divided into two categories: metalinguistic and implicit. Implicit methods will be briefly discussed in section 4.3.4; here I focus on metalinguistic studies.

Among metalinguistic syllabication tasks, there are both written and oral variants. The written tasks usually ask subjects to divide orthographic forms by inserting slashes or hyphens, or else to choose from among pre-syllabified alternatives (e.g. *lemon* → *le*|*mon* or *lem*|*on*?). Oral tasks consist of various word games that require the subjects to manipulate an aurally-presented form in some way, or else to indicate their preference for competing outputs of a manipulation. For example, participants might be asked to break a disyllabic word by inserting a pause (*lemon* → *le...mon, lem...on*, etc.), permute the order of syllables (*monle, monlem, onlem*), repeat either the first or second part (*le, lem, mon, on*), or reduplicate one of the elements (*le-lemon, lem-lemon, lemon-mon, lemon-on*). A thorough review of these tasks in provided in Côté & Kharlamov (2011).

With respect to English, the results of this body of research are somewhat mixed. On the one hand, the studies generally agree that unattested CC word onsets are almost always split when in medial position. For example, Fallows (1981) reported that,

across two oral reduplication tasks using real, disyllabic words, children treated such clusters as heterosyllabic about 98% of the time. Similarly, Treiman & Zukowski (1990) found that, when provided with pre-hyphenated alternatives, adults chose the heterosyllabic option 99% of the trials where the words contained illegal clusters. More recently, Redford & Randall (2005) reported nearly 97% split rates in nonsense disyllables while Eddington et al. (2013b) found illegal clusters to be split 91% of the time.

On the other hand, there is also evidence suggesting that word-edge legality does not guarantee tautosyllabic treatment. For one, attested CC word onsets are quite likely to be split, in apparent violation of the Maximal Onset Principle. This is especially true of sC clusters, where the initial /s/ has been argued to be extrasyllabic on theoretical grounds (e.g. Kaye et al., 1990). For example, Treiman, Gross & Cwikiel-Glavin (1992) found that, in a hyphenation and partial repetition task, sC clusters were split nearly 66% of the time. Similar rates were reported in Eddington et al. (2013b) and Redford & Randall (2005), though in the latter study the boundary judgments were modulated by a number of additional phonetic factors.

Interestingly, non-categorical parsing behavior does not appear to be confined to sC clusters. Other, legal CC word onsets are often split in both written and oral tasks, sometimes at rates over 50% (Eddginton et al., 2013b; Redford & Randall, 2005; Treiman et al., 1992; Treiman & Zukowski, 1990). A similar degree of uncertainty is exhibited with respect to intervocalic singletons. Despite the fact that some classical phonological theories usually require onsets to be filled (e.g. Itô, 1989), empirical parsing studies find that singletons are often affiliated with the preceding vowel. This is especially true if that vowel is lax and/or stressed, and if the segment is a sonorant (Eddington et al.,

2013a; Fallows, 1981; Treiman, Straub & Lavery, 1994; Treiman, Bowey & Bourassa, 2002; Treiman & Danis, 1988).

Redford & Randall (2005) appealed to gradient phonetics as explanation for variable hyphenations of phonotactically permissible onsets. In that study, native English listeners heard nonsense disyllables produced by different speakers, then wrote down and syllabified the forms. As mentioned above, medial sequences unattested in word-initial position were almost always split, and first-syllable stress was also a near-categorical cue for a heterosyllabic parse. However, the variability in the treatment of phonotactically viable CC onsets in items with second-syllable stress was well-captured by acoustic cues that characterized the different productions. Specifically, C1:C2 duration ratios correlated positively with the likelihood of the participants syllabifying the cluster as a complex onset (see section 5.3.4 for more discussion of this study). Redford and Randall (2005) argued for a two-step model wherein boundary judgments carried out by listeners are influenced first by categorical phonological factors (deterministic phonotactics and stress) and subsequently by gradient perceptual cues in the signal.

2.2.3 Summary

Although the theoretical views of syllabification reviewed in section 2.2.1 are characterized by a great deal of controversy and disagreement, one common thread runs through all of them. Namely, almost every account acknowledges some relationship between word and syllable margins. As noted above, this relation is asymmetrical, with the set of possible syllable edges being constrained by (but not

coextensive with) the set of attested word edges. This reflects a particular, classical view of phonotactics as *categorical restrictions* on sound sequencing. Curiously, the experimental findings summarized above exhibit more variability than a categorical view of phonotactics might allow. Most of this variability applies to attested word onsets, but even illegal onsets are not always split. A common explanation for non-categorical responses is that they are the product of competition between categorical syllabification principles. For example, the pressure to close lax vowels might compete with onset maximization, yielding variable parsing judgments (Fallows, 1981). In other words, variability in behavior reflects the ambisyllabic status of medial consonants and clusters. Note, however, that the notion of ambisyllabicity is qualitative rather than quantitative — simply saying that a segment is ambisyllabic does not confer enough precision to explain the variance in responses (i.e. to predict the probability of boundary placement).

To the extent that metalinguistic parsing behavior relies at least in part on grammatical knowledge (a view adopted here), the correct view of the grammar must accommodate stochastic parsing behavior. As noted above, Redford & Randall (2005) suggest a model where the locus of variability is in the signal. While such a model may go a long way toward explaining behavior in *perception*-based tasks, its applicability to *production* is less clear. A syllabification theory whose predictions generalize across different tasks and modalities would be more desirable[2].

One promising direction in developing such a theory lies in re-examining the phonotactic model assumed by classical syllabification theories. In the next section, I

[2] This is not to say that acoustic juncture cues are irrelevant; see section 5.3.4 for the suggestion that such cues might compete with phonotactics in cuing boundary locations in perception-based studies.

review evidence arguing that this model is outdated, and show that it has been replaced by a modern, stochastic view of phonotactics.

2.3 The Gradient Nature of Phonotactic Knowledge

A well-established finding in experimental phonology is that wordlikeness judgments are gradient: when evaluating the phonological acceptability of made-up words, people systematically exhibit fine-grained preferences for some strings over others (Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Hay, Pierrehumbert & Beckman, 2003; Vitevitch et al., 1997). In many cases, these preferences have been attributed to the composition of onset clusters: given a set of monosyllables like {blick, dwick, bnick, lbick}, English speakers do not make a binary distinction between the accidentally absent and the completely impossible (*blick, dwick* ≻ **bnick, *lbick*), as predicted by traditional phonological theory (e.g. Halle, 1959; Hooper, 1972; Prince & Smolensky, 1993/2004). Instead, their judgments tend to fall on a continuum such that *blick* ≻ *dwick* ≻ *bnick* ≻ *lbick* (e.g. Daland et al., 2011; Scholes, 1966). These judgments are generally taken to reflect the speakers' phonotactic grammar — the part of their phonological knowledge concerned with sound sequencing patterns. Fine-grained sensitivity to these patterns is difficult to capture by classical models that cast phonotactics in terms of absolute restrictions, leading to the alternative view that phonotactic knowledge is gradient rather than categorical. This view has received support from a variety of psycholinguistic studies, which repeatedly show gradient processing asymmetries related to phonological structure (Berent et al., 2007; Luce &

Pisoni, 1998; Pitt & McQueen, 1998; Vitevitch et al., 1997). Recent modeling efforts have been aimed at capturing this gradience by imputing a stochastic component to the grammar (e.g. Albright, 2009; Berent et al., 2009; Boersma & Hayes, 2001; Coetzee, 2009; Coleman & Pierrehumbert, 1997; Hammond, 2004; Hayes & Wilson, 2008).

Two kinds of factors have been implicated in the gradient well-formedness of nonce forms. The first is the influence of the lexicon: novel forms elicit favorable responses and enjoy certain processing advantages to the extent that they receive lexical support. One way to operationalize this support is in terms of frequencies, transitional probabilities, and other statistics accumulated over sublexical units such as segments, syllables, and sub-syllabic constituents. For example, Bailey & Hahn (2001) reported that nonce forms featuring highly probable bigrams were judged as better than those featuring low-probability sequences. Coleman & Pierrehumbert (1997) modeled acceptability scores of nonce words as a function of the cumulative probability of their subparts as estimated from the lexicon. In addition to being judged as better, Frisch, Large & Pisoni (2001) found that nonwords with higher probability constituents were remembered more accurately, and Hay, Pierrehumbert & Beckman (2003) showed that such forms were less likely to be misperceived. In production, Vitevitch et al. (1997) found that pseudowords consisting of high-frequency syllables were repeated facter than those made up of low-frequency syllables. Taken together, these studies suggest that phonotactic knowledge is 'projected from the lexicon' in the sense of being extracted from linguistic experience via the mechanism of statistical learning (see Saffran, Aslin & Newport, 1996 for experimental evidence of statistical learning of phonotactics in infants and Dell et al., 2000, Onishi, Chambers & Fisher, 2002, Warker &

Dell, 2006, and Whalen & Dell, 2006 for evidence that adults require relatively little exposure in order to learn certain novel phonotactic patterns).

Aside from sublexical statistics, another way to measure lexical support is in terms of similarity to real words. A common similarity metric is edit distance, defined as the number of phoneme additions, deletions or substitutions required to change one string into another (Levenshtein, 1966). Words within one edit from an item are said to comprise that item's phonological neighborhood (Luce & Pisoni, 1998); the size of this neighborhood correlates with well-formedness ratings and production accuracy[3] (Arnold, Conture & Ohde, 2005; Bailey & Hahn, 2001; Hammond, 2004). For the monosyllables *blick* and *dwick*, both of which feature attested onsets, the well-formedness asymmetry is transparently projected from the lexicon: *blick* features 11 phonological neighbors to *dwick's* two, and [bl] is about 13 times more likely than [dw] to begin a word.[4]

In addition to common measures of lexical support, the second factor often associated with well-formedness of a monosyllable is the sonority profile of its onset. Several sonority scales varying in granularity have been proposed in the literature (see Baertsch, 2012 for a review); a representative, coarse scale from Clements (1990) is shown in (2.2), with natural classes increasing in sonority from left to right:

(2.2) *obstruents* < *nasals* < *liquids* < *glides* < *vowels*

[3] The influence of lexical neighborhoods has been argued to be separate from that of phonotactics, possibly affecting processing at different stages (Bailey & Hahn, 2001; Vitevitch & Luce, 1998; Storkel, Armbrüster & Hogan, 2006).

[4] Calculation based on a pre-processed CMU pronouncing dictionary (Weide, 1994). See following chapter for details.

As noted above, the SSP favors rising-sonority onsets and falling-sonority codas.

The SSP appears to be a useful generalization in that it predicts not only wordlikeness judgments but also performance in several perception and production tasks. For example, among unattested word onsets, those with falling sonority profiles are more likely to be misperceived with an epenthetic schwa than sonority plateaus, which in turn induce perceptual epenthesis at rates higher than rises (p([ləbɪf] | [lbɪf]) > p([bədɪf] | [bdɪf]) > p([bənɪf] | [bnɪf]; see Berent et al., 2007). This effect appears to hold even for speakers of languages which prohibit complex onsets altogether (Berent et al., 2008). In children's productions, cluster reduction patterns appear to be motivated by the preservation of the best sonority profile available (Ohala, 1999).

As noted in section 2.2.1, the cognitive status of sonority is controversial, with nativist, naturalist and lexicalist accounts characterizing the debate. The position taken in this dissertation is a combination of the latter two viewpoints. That is, sonority itself is seen as a cover term for a number of articulatory properties (e.g. jaw displacement, degree of stricture, etc.; see Redford, 2008) and their perceptual correlates (namely loudness, either maximal or integrated over duration; Parker, 2002; Wright, 2004). The SSP as a typological generalization is understood here as an epiphenomenon of these phonetic properties exerting soft pressure on the evolution of lexicons across languages. Following Daland et al (2011), I also assume that SSP effects are projected from the English lexicon in that the treatment of unattested onsets can be modeled as a function of feature-based similarity to attested onsets, as long as the model is capable of expressing sonority relations. Thus, in what follows, the use of the term 'sonority' is to be understood as a label of convenience covering phonetically-grounded properties of

segments, and the term 'SSP' as a sequencing preference that is largely recoverable from English lexical statistics.

In summary, people's sensitivity to sound sequences clearly goes beyond categorical phonotactic distinctions. In some cases, the performance is captured by a straightforward projection of lexical statistics; in others, sonority (understood as stated above) appears to be a useful cover term. Given this sensitivity to gradience, an interesting question arises regarding the relationship between phonotactics with the rest of phonology. Namely, how is fine-grained phonotactic knowledge deployed by the grammar? To the extent that other phonological processes interface with this knowledge, what is the relevant level of detail? Does all of phonology respond to gradient phonotactics, or are there processes which rely on more coarse-grained phonotactic generalizations? In this dissertation, I argue that syllabification — or, what I call the *metrical parse* — is a phonological process that, contrary to the classical assumptions reviewed in section 2.2.1, relies on fine-grained rather than categorical phonotactics. Thus, a main contribution of this thesis is to incorporate a modern phonotactic model into theories of syllabification. Following the work outlined above (Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Frisch et al., 2000; Hay et al., 2003; Vitevitch et al., 1997, *inter alia*), the source of phonotactic knowledge – including knowledge related to sonority sequencing – is assumed to be the lexicon. The gradient metrical parse hypothesis is made explicit in the next section.

2.4 The Gradient Metrical Parser Hypothesis

Consider the set of pseudowords discussed above, this time prepended with the sequence *vata*, in order to place the onsets in medial position: {*vatablick, vatadwick, vatabnick, vatalbick*}. What is the appropriate metrical parse of each medial cluster? Table 2.1 summarizes four logical possibilities.

Table 2.1. Four parsing hypotheses.

Parsing model:	lower ← P(C.C parse) → higher			
H1: CATEGORICAL	*vatablick, vatadwick*			*vatabnick, vatalbick*
H2: GRADIENT, LEXICON-BASED	*vatablick*	*vatadwick*		*vatabnick, vatalbick*
H3: GRADIENT, SONORITY-BASED	*vatablick, vatadwick*		*vatabnick*	*vatalbick*
H4: FULLY GRADIENT	*vatablick*	*vatadwick*	*vatabnick*	*vatalbick*

In *H1*, the parser is phonotactically coarse-grained; all else being equal, syllable boundaries are predicted by the Legality Principle so that /bl/ and /dw/ remain tautosyllabic while /bn/ and /lb/ are split. Alternatively, the parse may be gradient, relying on fine-grained word-edge statistics calculated over segments (*H2*), fine-grained sonority (*H3*), or both (*H4*). In this dissertation, I test these four hypotheses in a number of experiments that probe the relationship between phonotactic and metrical knowledge from different angles. All experiments utilize the same set of stimuli — trisyllabic nonce forms with embedded clusters and singleton consonants, similar in shape to the example items in Table 2.1.

The four hypotheses above can be formally described with equal success in a number of ways, using either rule-based or constraint-based frameworks. In the remainder of this section, I briefly discuss the relationship between phonotactics and syllabifications in terms of a variant of Optimality Theory (OT) (Prince & Smolensky, 1993/2004). In OT, grammatical well-formedness is decided with reference to a hierarchy of ranked constraints which push for the preservation of lexical contrasts or militate against specific structures. The choice to employ Optimality Theory as an expository device was motivated by the fact that (a) this framework has largely supplanted derivational phonology and is thus preferred by most phonologists, and (b) OT-based accounts of gradience are accessible and amenable to visualization (see below).

As discussed above, the mainstream view in phonology assumes the model listed under *H1* in Table 2.1, where syllable boundaries are determined with reference to rather coarse-grained phonotactics. A theory of this sort must reconcile the categorical phonotactic parser with gradient, lexicon-based phonotactic effects observed in perception, production and well-formedness judgments. There are two ways of achieving this within OT. One is to assume that different processes interact with the constraint hierarchy in different ways. For example, the parser might be driven by the relative constraint ranking of of LOI ≫ NoCoda when selecting the output, where LOI is a constraint militating against all unattested onsets (named after Vennemann's (1972) Law of Initials, see Raffelsiefen, 1999) and NoCoda is a constraint banning syllable codas. Under classical OT which features strict ranking, outputs violating low-ranked constraints are selected over competitors which violate highly-ranked constraints; ranking LOI over NoCoda ensures that the input *vatabnick* will always surface as

va.tab.nick and never as **va.ta.bnick*. At the same time, constraints banning individual onsets might be ranked on a continuum **.lb* ≫ **.bn* ≫ **.dw* ≫ **.bl*, which is established as learners become attuned to lexical statistics and/or sonority. This continuum would be invisible to the parser but not to processing tasks and well-formedness judgments, giving speakers the ability to judge the relative harmony of losing candidates. Such hybrid grammar proposals have been advanced to account for the differences in task sensitivity to OCP violations (Berent & Shimron, 1997; Berent et al., 2001; Coetzee, 2009).

The other possibility is to model categoricity as extreme probability. For instance, NoCoda might have an extremely low probability of outranking individual constraints militating against unattested onsets, but a very high probability of outranking those banning attested onsets. This would yield a nearly categorical parser without the need for LOI, while at the same time preserving the relative rankings of the individual markedness constraints. Several existing, stochastic OT models could easily incorporate such a parser because they were designed to accommodate variation (e.g. Boersma & Hayes, 2001; Hayes & Wilson, 2008). All that is required is some mechanism for probabilistically ranking or weighting the constraints militating against alternative parses. For example, consider a grammar that operationalizes variable constraint ranking in the form of probability distributions over a continuous ranking scale. A toy version of such a grammar, using the Gradual Learning Algorithm (GLA; Boersma, 1997; Boersma & Hayes, 2001) is illustrated in Figure 2.1. The horizontal axis in each panel represents the weight scale; the further left a constraint is positioned, the higher its weight. Constraint weights are transformed into rankings at the moment of production using the distributions represented by the normal curves. Each distribution

corresponds to a different constraint and is centered on the weight of the constraint. Its variance represents noise in the evaluation process and is assumed to be constant across all constraints. The height of the curve at a given point along the scale therefore represents the probability of the constraint being ranked at that point. To the extent that two distributions overlap, their relative ranking is variable, potentially resulting in observable variation in the output.

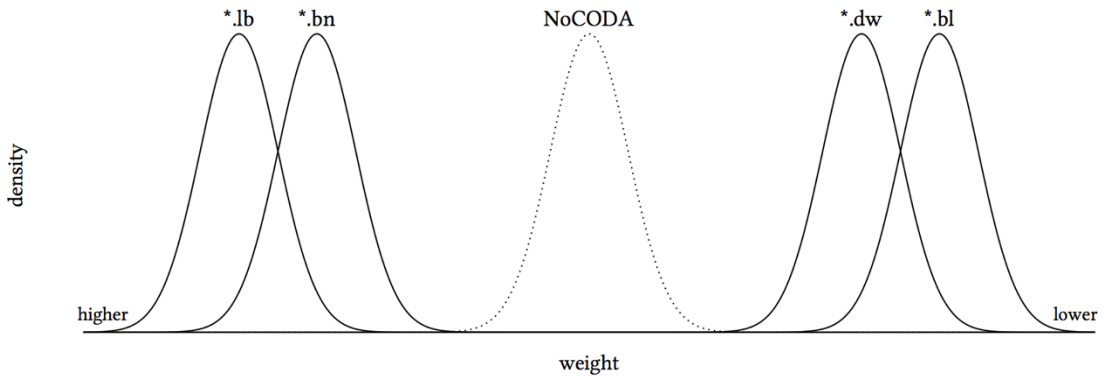

Figure 2.1. Reconciling a categorical parser (*H1*) with gradient well-formedness in a stochastic OT grammar based on the GLA (Boersma, 1997; Boersma & Hayes, 2001).

The probability distributions plotted with solid lines correspond to the markedness constraints militating against individual syllable onsets. Their ordering represents the well-formedness gradient, which I assume to be estimated from the lexical statistics of word edges, as well as sonority profiles (although it may be the case that sonority is itself projected from the lexicon, see Daland et al., 2011).

The parsing preference of the toy grammar is represented by the NoCoda constraint (plotted with a dotted line for clarity), which prefers complex onsets to split clusters. Of course, a single constraint is a gross oversimplification of the parsing system (see e.g. Hall, 2004 for a representative constraint set), but it is sufficient for the

purpose of illustrating the interaction between phonotactics and syllabification. The position of the NoCoda distribution along the axis (chosen arbitrarily for this illustration) establishes a well-formedness threshold of sorts: clusters banned by the constraints whose curves lie to the left of NoCoda are likely to be split, while those to the right are likely to be preserved.

In this particular example, the markedness constraints are loosely arranged into two groups, with those banning initially-unattested CC onsets (/lb, bn/) outranking those that militate against attested onsets (/dw, bl/). There is substantial overlap within each group, allowing for the emergence of gradience in a number of behavioral outcomes, including well-formedness judgments, processing speed, perceptual repairs, and speech errors. That is, *lbick* will usually but not always be judged as worse than *bnick*, and the same advantage will hold for *blick* over *dwick*. At the same time, the gap between the two groups is wide enough so that the unattested onsets will almost never be judged as better than the attested onsets. Crucially, there is virtually no overlap between the two groups of markedness constraints and the NoCoda distribution positioned between them. This arrangement virtually guarantees that unattested onsets will be split, and attested onsets preserved.

The other possibility is that the metrical parse is gradient rather than categorical. There is some empirical evidence suggesting such a model. For one, probabilistic, sonority-based parsing strategies have been reported in word segmentation and phonotactic learning studies. Ettlinger, Finn & Hudson Kam (2011) trained native English listeners on an artificial speech stream that contained novel CC clusters with fixed transitional probabilities and varying sonority profiles. After training, SSP-violating clusters were more likely to cue a word boundary between the

two consonants than SSP-preserving clusters. However, it is not clear whether this sonority preference would operate on medial syllables. Better evidence is provided in Redford (2008), where native English-speaking adults listened to disyllabic nonce words with novel onsets of either rising or flat sonority (e.g. *tlevat* or *bdevat*). Following training, the subjects performed a written hyphenation task on items containing the same clusters in intervocalic position (*vatlet* or *vabdet*). The group that trained on rising word onsets showed better generalization to medial position, producing a higher rate of V.CCV parses than the flat onset group. Finally, Kharlamov (2009) asked Russian speakers to judge the well-formedness of initial and medial onsets on a Likert scale (the stimuli were orthographically presented, pre-syllabified nonwords so that medial onsets were preceded by a dash). The results indicated some influence of word-edge statistics on medial onset judgments.

A gradient metrical parser would also fall out naturally from a stochastic grammar like the one assumed by the GLA. This is illustrated in Figure 2.2. The order of the markedness constraints is the same as in Figure 2.1, but the two distribution groups are close enough that they overlap with NoCoda. This overlap is what ensures gradient parsing outcomes: the larger the overlap, the higher the probability of a ranking reversal so that even *vatalbick* has some chance of syllabifying as *va.ta.lbick*. Note also that, in this example, all of the markedness distributions overlap with each other, indicating a non-zero probability of an unattested onset being judged as better than an attested onset.

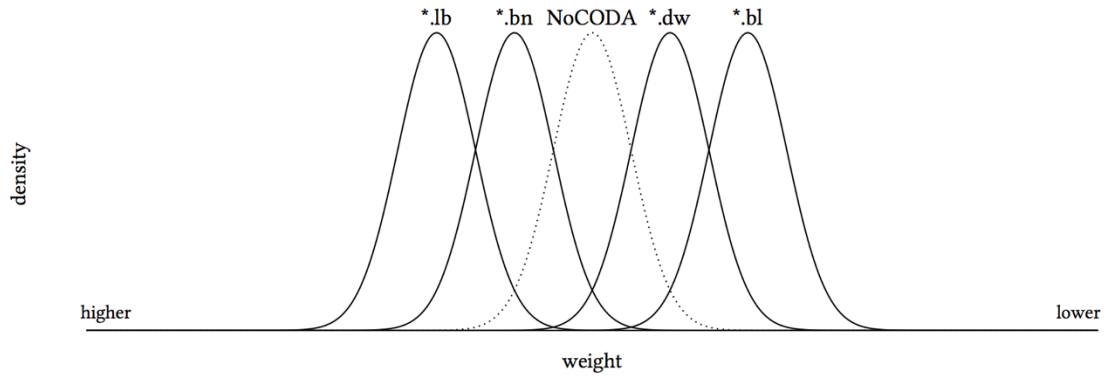

Figure 2.2. Fully gradient parser (*H4*) as a stochastic OT grammar based on the GLA (Boersma, 1997; Boersma & Hayes, 2001).

The toy parser shown in Figure 2.2 illustrates the fully gradient model (*H4* in Table 2.1 above). The assumption is that the sources of gradience that govern the parse are the same as those reflected in phonotactic judgments and other processing tasks. However, this is an empirical question; in principle, the influences of sonority and lexical support could be approached as orthogonal (though see Daland, et al., 2011). For instance, the parser might be sensitive to the statistics of word edges: given medial C1C2 clusters, word-initially common sequences might prefer to syllabify as complex onsets, while C1 segments frequent in word offset position might push for a split parse. This possibility (*H2* in Table 2.1) can be easily visualized by shifting the two leftmost curves further to the left, as shown in Figure 2.3.

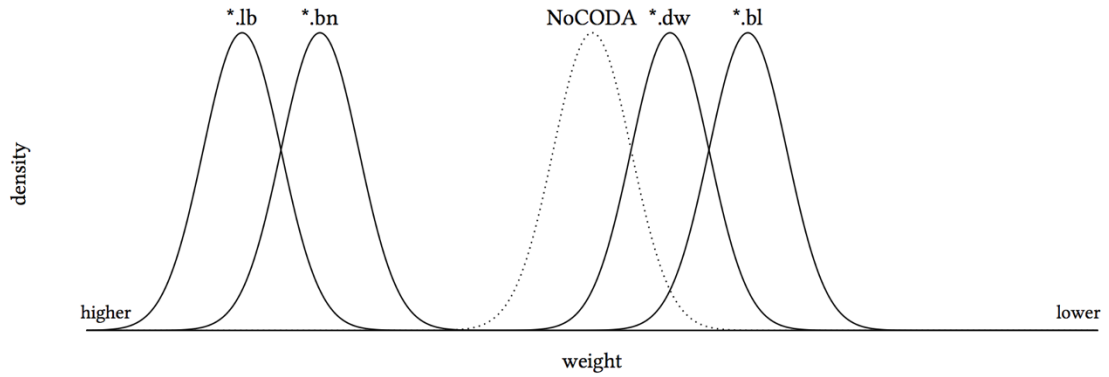

Figure 2.3. Lexicon-based, gradient parser (*H2*) as a stochastic OT grammar based on the GLA (Boersma, 1997; Boersma & Hayes, 2001).

Here, only the distributions banning legal onsets overlap with NoCoda. This model, which predicts that initially unattested onsets are always split but attested onsets are not always maximized, seems to be consistent with the bulk of the experimental syllabification studies reviewed in section 2.2.2. Independently of this, syllabification might be guided by sonority, with the probability of a heterosyllabic parse rising as the sonority slope across the cluster grows more negative. Such a relationship would not only be consistent with the SSP, which disprefers falling onsets as discussed above, but also with the Syllable Contact Law (Vennemann, 1988), which prefers sonority falls across syllable boundaries. This model, *H3* in Table 2.1 above, is shown in Figure 2.4. The two initially-attested onsets are always maximized (since they both feature rising profiles), whereas among the unattested onsets, the one with a rising sonority profile has a greater chance of being preserved.

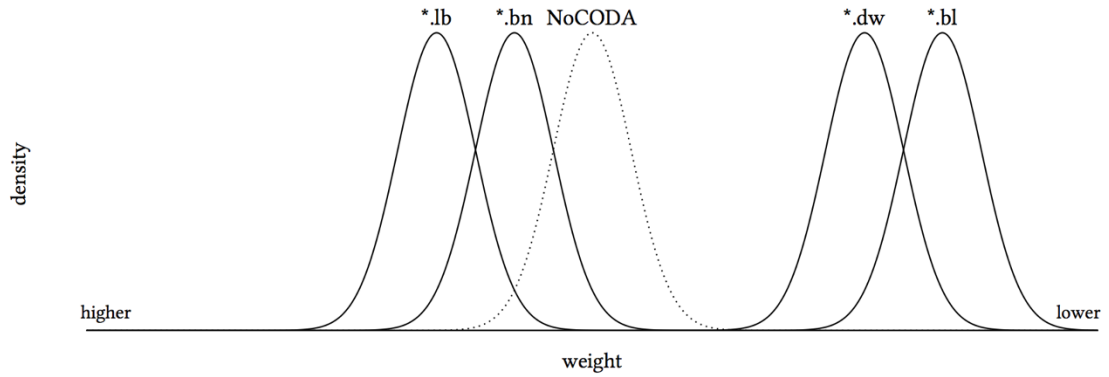

Figure 2.4. Sonority-based, gradient parser (*H3*) as a stochastic OT grammar based on the GLA (Boersma, 1997; Boersma & Hayes, 2001).

In this dissertation I will argue against the categorical parser (*H1*) in favor of a gradient model along the lines of *H2* and *H4*. Across a number of production, perception and metalinguistic experiments, I will demonstrate that syllable boundaries are assigned stochastically. In some cases, there will be clear evidence for *H4*, with gradience within both attested and unattested word onsets. In others, attested onsets will exhibit variability based on their word-edge frequencies, but the contributions of sonority to predicting behavior on unattested onsets will be more modest (*H2*). The overall picture that will emerge is one which sees syllables as emergent from probabilistic generalizations over the lexicon (specifically, over word edges) rather than as deterministic products of categorical rules or fixed constraint rankings.

CHAPTER III

METHODOLOGICAL PRELIMINARIES

3.1 Overview of the Experiments

The bulk of this dissertation is composed of five different studies which represent different ways of addressing the question of phonotactic granularity involved in syllabification. Four of the studies rely on essentially the same set of trisyllabic pseudoword stimuli. The stimuli contained medial singletons and clusters of differing phonotactic properties (see section 3.3 below). Study 1 is a written hyphenation task where participants syllabified the orthographically presented nonwords by inserting slashes between graphemes. Study 2 is a reanalysis of Eddington et al. (2013a,b), where participants indicated their preference for pre-syllabified alternatives of disyllabic English words. Studies 3 and 4 infer the location of syllable boundaries from stress assignment; Study 3 is a binary preference task and Study 4 is an online stress assignment task. Finally, Study 5 is an analysis of the speech errors produced by the participants in the stress assignment study.

In all five studies, the categorical parsing model (*H1*) was compared to the gradient parsing model (*H4*). The same set of independent variables were used to predict responses across the studies. For the categorical models, the predictor was legality of the medial consonant sequence in word-initial position. In the gradient models, predictors included two measures of lexical support — word onset-frequency and word-offset frequency of the medial sequence — as well as sonority slope. Models

in the stress-based studies also included nuisance factors. All of the predictors are described in detail in section 3.4.

3.2 The Lexicon

All of the measures of lexical support used as predictors were calculated over the same database of English words. This approximation of the lexicon (henceforth: 'the lexicon') was assembled by filtering the CMU pronouncing dictionary (Weide, 1994) through the SUBTLEXus corpus of film and television subtitles (Brysbaert & New, 2009). In this section, I describe the assembly process in some detail.

The CMU pronouncing dictionary is a machine-readable database developed with the purpose of aiding automatic speech recognition research. The dictionary contains over 134,000 phone-level transcriptions of word forms intended to reflect North American English pronunciations (it is not clear which dialect is taken as the standard, though many words are listed with several pronunciation variants). The transcription system employs 39 phones and marks three levels of stress: main, secondary and unstressed.

Because the CMU dictionary was designed to provide maximum coverage, it contains a large number of proper names, borrowings and other rare forms (for instance, according to the documentation, the dictionary contains over 53,000 synthesizer-generated, unproofed proper names). While necessary to a robust speech recognition system, such forms are extremely unlikely to be encountered by a typical native speaker. Furthermore, their preponderance might skew the lexical support

measures of interest, misrepresenting the phonological generalizations available to ordinary human learners. For this reason, the lexicon was constrained to those CMU entries which also appeared in the SUBTLEXus corpus at least once (see Moore-Cantwell, 2016 for the same approach). The SUBTLEXus corpus contains some 51 million words harvested from the subtitles of US-produced films and television series. Frequencies based on SUBTLEXus have been shown to be very effective in predicting lexical decision accuracies and reaction times (Brysbaert & New, 2009), outperforming counts based on Kučera & Francis (1967) as well as the CELEX corpus. This makes SUBTLEXus one of the best available sources for studying token frequency effects in contemporary American English speakers.

After filtering the CMU dictionary through SUBTLEXus, the lexicon was checked by hand and further refined. Acronyms and abbreviations were removed, as were any errors in stress placement. In addition, two types of pronunciation variants were removed. The first contained schwas which I judged to be epenthetic in the sense that they were likely to be produced only in slow, emphatic speech such as when the speaker is trying to sound out the letters in the word. For example, *chronically* was transcribed as both [kɹanɪkli] and [kɹanɪkəli]; the latter variant was judged to contain an epenthetic schwa and was therefore deleted. The other type of pronunciation variant removed from the lexicon contained initial [hw] clusters in words like *wet* (listed as both [wɛt] and [hwɛt]).

Thus refined, the lexicon contained a total of 48,951 word forms. Further details about syllabification, morphology and stress are presented in Section 5.2.

3.3 The Stimuli

Ever since Berko's (1958) ground-breaking work on morphological productivity, nonsense words have become an indispensable tool for probing the nature of linguistic knowledge. Alternatively referred to as 'nonwords', 'pseudowords', 'nonce probes', or 'wugs' (the last after one of Jean Berko's original stimuli), these meaningless phoneme or grapheme strings are typically designed to test specific hypotheses related to phonological structure. Because the processing of unfamiliar forms cannot involve wholesale recall and must therefore be mediated by grammatical knowledge, pseudowords represent an ideal test case for competing theories of grammar. In experimental phonology, they have been employed to investigate a number of phenomena, including phonotactics (Scholes, 1966; Redford, 2008), sonority (Berent et al., 2007, 2008), voicing alternations (Becker, Ketrez & Nevins, 2011), palatalization (Kapatsinski, 2013; Wilson, 2006), syllable weight (Ryan, 2011a), stress (Baker & Smith, 1976; Carpenter, 2010; Guion et al., 2003), saltatory alternations (White, 2017), vowel assimilation (Moreton, 2008), pitch accent (Shport, 2011), and many others.

In this dissertation, I employ pseudowords to probe the granularity of the metrical parse. Four experiments draw their stimuli from the same set of 170 nonsense probes. These items, listed in Appendix A, were specifically designed to focus on the effect of phonotactics on syllabification. To achieve this focus, the design was constrained by a number of criteria. First, to limit the number of nuisance factors, the stimuli had to be consistent in size, CV shape and locus of the phonotactic interactions of interest. Second, because three of the experiments relied on Latin Stress as a window into the parse, the words had to be long enough to carry this stress pattern (i.e.

trisyllabic or longer). Finally, it was important to discourage analogical processing (comparing the nonce probes to similar lexical neighbors). For this reason, the probes could not resemble real English words in any obvious way.

These three constraints gave rise to a set of nonsense trisyllables that all shared the same CVCVC(C)VC template. The underlined portion between the penultimate and final vowel represents the embedding site for various *inserts*, while the remainder of the pseudoword will be referred to as the *context frame*. The inserts consisted of singletons and biconsonantal clusters chosen to vary along a number of dimensions, including word-initial legality and frequency, sonority profile, and word-final frequency of the initial consonant (see section 3.4.1 for description of the measures). In other words, they instantiated the phonotactic generalizations of interest. A total of 75 inserts were chosen; 12 singletons, 28 clusters attested as word onsets at least once in the lexicon (as defined in the preceding section), and 35 initially unattested CC sequences. The complete set of inserts are listed in Table 3.1.

Table 3.1. Set of inserts used in pseudoword construction (orthographic representation).

Type	Natural Classes	Insert
singleton	obstruent	*p, t, k, b, d, g, f, v, th, s, z, sh*
attested	obstruent + sonorant	*pr, pl, tr, tw, kr, kw, br, bl, dr, dw, gr, gl, fr, fl, thr, sl, sm, sn, shr*
unattested (rising sonority)	obstruent + sonorant	*pm, pn, tl, tn, kn, bn, bw, dl, dm, gm, gn, fm, vr, vl, thl, sr, shn, zr, zl*
unattested (falling sonority)	sonorant + obstruent	*lp, lt, lb, lf, lv, lth, ls, rb, rz, mp, md, mg, mf, nt, nk, nb, ng, ns, nsh*

A small number of the inserts in the 'attested' category (namely, {*shn, tl, vl, vr, zl*}) have been treated as 'unattested' or 'marginal' in prior work (e.g. Daland et al, 2011). This choice is usually justified by the intuition that, because these onsets are instantiated in a very small set of rare borrowings, they constitute exceptions that are processed differently from other legal onsets. Here, I take a different, data-driven approach: as long as a word onset appeared in the SUBTLEXus corpus, it was counted as attested. This approach has the benefit of objectivity in that the line between borrowings and native vocabulary is often difficult to draw. That said, in order to forestall objections, the relevant analyses were also conducted with these inserts reclassified or excluded (the details are spelled out below where necessary).

The inserts were distributed across 44 unique CVCV__VC context frames. The vowel graphemes used to construct them were limited to {*a, e, i*} as these were thought least likely to be interpreted as phonologically tense, an undesired complication that would affect stress placement (see section 5.4.2.4 for details). There were no *a priori* constraints on the frame consonants. Each CC cluster was embedded into two different frames while the singletons were placed in 2-5 contexts. The frames were distributed such that each one covered a similar sonority range. For example, the frames *daka___uth* and *shepi___oph* took the same set of inserts, producing the following pseudowords:

daka<u>d</u>uth, *shepi<u>d</u>oph* (singleton)

daka<u>dw</u>uth, *shepi<u>dw</u>oph* (attested/rising sonority)

daka<u>dm</u>uth, *shepi<u>dm</u>oph* (unattested/rising sonority)

daka<u>md</u>uth, *shepi<u>md</u>oph* (unattested/falling sonority)

This arrangement constrained variability among the 170 test items and emphasized the contrastive role of the inserts.

Obvious similarity to real words was avoided by making sure that most of the test items did not contain any substrings that could be parsed out as common English affixes[5]. Furthermore, the pseudowords were compared to the lexicon of English trisyllables using a measure of orthographic edit distance. Edit distance is a common similarity metric intended to quantify the density of a probe's lexical neighborhood (the subset of the real words that passes some pre-defined similarity threshold relative to the probe). Orthographic edit distance is defined as the number of grapheme additions, deletions or substitutions required to change one string into another. The standard definition of a lexical neighborhood encompassed lexical items within one edit of the probe (Luce, 1986). However, the pseudowords used in this dissertation intentionally had no neighbors under this definition, necessitating a different approach. Namely, lexical similarity was operationalized as the average orthographic edit distance to 10 nearest neighbors (see Keuleers, 2013 for R implementation). A similar measure based on 20 neighbors has been found to outperform the standard definition of neighborhood as a predictor of lexical decision speed and nonword production accuracy (see also Suárez et al., 2011; Yarkoni, Balota & Yap, 2008). On average, the nonce probes were nearly 5 edits away from their 10 nearest neighbors, confirming the intuition that they did not resemble real words in an immediately obvious way. Nevertheless, this did not rule out the potential role of analogical processing. The next section describes a

[5] A few frames contained initial *be-*, *de-*, *re-* and final *-ish*, all of which are valid affixes. Removing these items from the analyses had no effect on the findings.

statistical measure intended to control for a potential cofound between analogy and phonotactics.

3.4 Predictors

This section describes those properties of the nonce probes which were examined as potential factors in the metrical parse. They divide into two sets: the phonotactic predictors were of theoretical interest, representing generalizations over sequential dependencies in the lexicon. The 'nuisance predictors' controlled for potential confounds in a subset of the studies. The by-item and by-insert values of the predictors are tabulated in Appendix A and Appendix B, respectively.

3.4.1 Phonotactic Predictors

Insert Status

This predictor (sometimes shortened to 'status' in what follows) was based on the word-initial legality of the inserts, as established by checking the entire word form lexicon described in section 3.2 (see also section 3.4 for discussion of the criteria). The predictor divided the inserts into three levels: *singleton*, *attested* and *unattested*. All of the singleton inserts were initially legal (i.e. the segment [ŋ] was not included among them). Coda legality was not variable because the initial consonant of every insert was attested word-finally. Therefore, initial status was the only measure of categorical phonotactics.

Word Onset Frequency

This predictor was a segment-based, gradient measure of lexical support computed over word onsets in the entire lexicon of 48,951 word forms. Figure 3.1 shows a histogram of the values across the 170 stimuli.

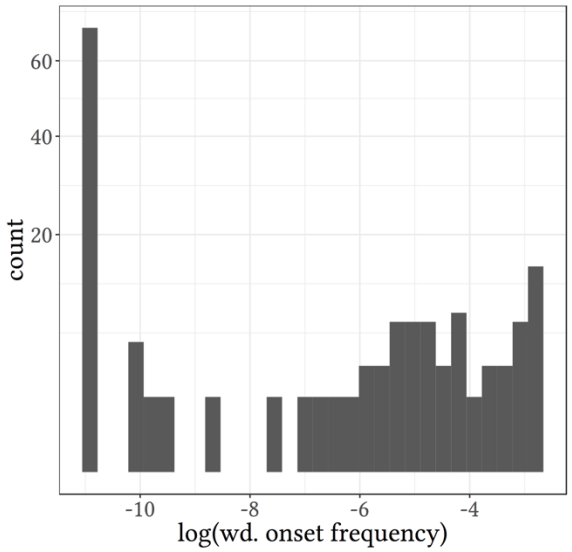

Figure 3.1. Histogram of the log frequencies of the inserts in word initial position (170 nonce probes). The leftmost spike represents unattested onsets, which were assigned a count of 1 in order to enable the log transformation.

The distribution is roughly bimodal, with a spike on the left representing unattested items (all of which were assigned the identical score of -10.8), a large mass on the right depicting frequent C and CC word onsets, and a sparsely-populated region of marginal onsets in between.

Word Offset Frequency

The affinity of consonants to parse into codas was approximated by measuring the word-final log frequency of the initial segment of each insert (in the case of

singletons, the only segment). The formula was the same as for onset frequency. The distribution of scores in plotted in Figure 3.2.

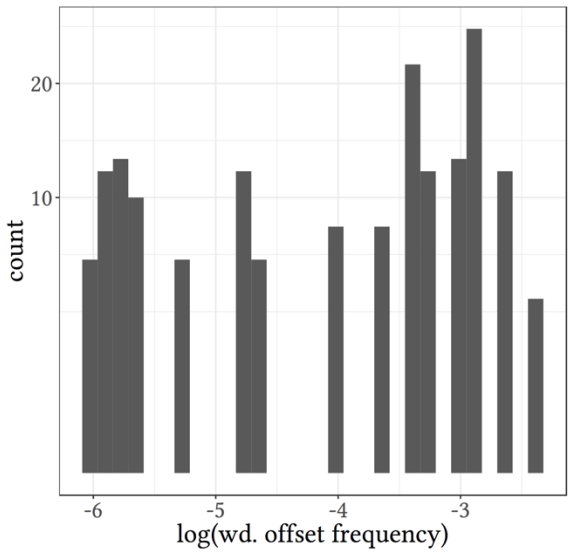

Figure 3.2. Histogram of the log frequencies of the inserts' initial consonants in word final position (170 nonce probes).

Because several inserts shared the same initial consonant, items containing singletons, attested and unattested clusters were collapsed across as long as their insert began with the same segment (e.g. /bn/, /bl/ and /b/ all received the same value on the measure).

Sonority Slope

The sonority slope predictor captured both the direction and magnitude of each insert's sonority profile. The measure was based on Jespersen's (1904) fine-grained sonority hierarchy, recapitulated in Table 3.2

Table 3.2. Sonority values used to calculate insert sonority profiles.

natural class	vowel	glide	rhotic	lateral	nasal	vd. fricative	vcls. fricative	vd. stop	vcls. stop
sonority	9	8	7	6	5	4	3	2	1

For CC inserts, sonority slope was calculated by subtracting the value of the first consonant from that of the second. For example, the values for *pr*, *lv*, and *lp* were 6, -2 and -5, reflecting a steep rise, shallow fall and steep fall, respectively. For singleton inserts, the sonority values were subtracted from 9, the value of a vowel. Figure 3.3 shows the histogram of sonority slopes across the pseudowords.

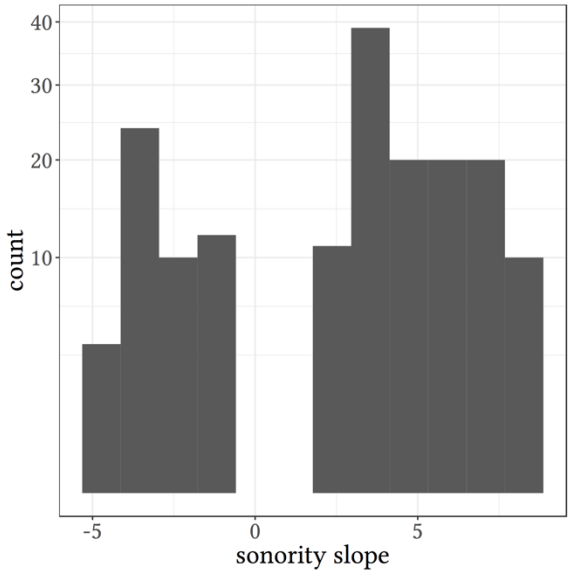

Figure 3.3. Histogram of the sonority slope values of each insert (170 nonce probes).

The fine granularity of the Jespersen scale yielded no flat-sonority profiles because no inserts were made up of two segments that agreed in manner and voicing. The closest were [s+stop] clusters, which were assigned a score of -1 (they would be treated as flat under Clements, 1990). Besides [s+stop] sequences, English has no word onsets with falling sonority; all other negative profiles thus corresponded to initially

unattested clusters. Positive values were distributed across singleton, attested and unattested onsets.

Sonority slope was found to be correlated with both word onset frequency (r = .74) and word offset frequency (r = -.45): clusters with steeper sonority rises tend to be rather frequent in word-initial position but rare in word-final position. For this reason, sonority was residualized against the two frequency measures before it was entered into multivariate models. This procedure effectively eliminated the collinearity and was justified on conceptual grounds: since both frequency measures are based on experience with the lexicon and thus reflect positive evidence for syllable boundaries, I deemed it appropriate that they account for all of the variance shared with sonority. The residuals can be understood as phonetic substance constraints operating on unattested onsets (see sections 2.2.1 and 2.3).

3.4.2 Nuisance Predictors

Two additional predictors were included in Studies 3 and 4, which relied on the relationship between English stress and syllable structure to infer the metrical parse. Both reflect legitimate influences of the lexicon on processing and thus constitute potential factors in stress assignment (see chapter VIII). However, neither relates explicitly to sequential dependencies between segments. Since the aim of this dissertation is to investigate the relationship between phonotactics and syllable structure (rather than develop a comprehensive model of stress assignment), these predictors were treated as 'nuisance variables' and added to the models as statistical controls for the lexicon-based, non-phonotactic influences on stress placement.

Edit Distance Bias

Previous research has argued that, when faced with the task of assigning stress to a novel form, one available strategy is to proceed on the basis of similarity to known words. The definition of similarity has differed depending on the study. Baker & Smith (1976) created nonwords by altering real lexical items by one or two graphemes. Guion et al. (2003) and Moore-Cantwell (2016) simply asked their participants to produce the closest lexical neighbor for each test probe. In their study of Dutch stress errors, Gillis, Daelemans & Durieux (2000) calculated similarity as the degree of overlap between segments occupying the same syllabic positions. In each of these studies, the assumption was that stress is assigned by reference to the single closest neighbor.

Here, I take a somewhat different approach to analogy. Recall from section 3.3 that the pseudowords had no immediate lexical neighbors and differed from the 10 closest words by an average of 5 edits. At this distance, a probe is likely to have more than a single nearest neighbor. Furthermore, it is not clear that a 5-edit neighbor should face no competition from a 6-edit neighbor. Indeed, the superiority of average edit distance over single-edit neighborhoods in predicting lexical decision tasks (Suárez et al., 2011; Yarkoni et al., 2008) suggests that an aggregate measure of phonological similarity may be more appropriate. Following this logic, I relied on a measure based on the mean orthographic edit distance to ten nearest neighbors (see section 3.3). First, I divided the database of trisyllabic word forms into penult- and antepenult- stressed words and calculated the average orthographic edit distance from each nonce probe to the ten nearest neighbors from each set. Having obtained two distance scores for each test item — one for antepenult-stressed, one for penult-stressed words — I subtracted

the former from the latter, yielding the predictor: an analogical measure of antepenult stress bias. Figure 3.4 displays the distribution of the bias scores across the test probes.

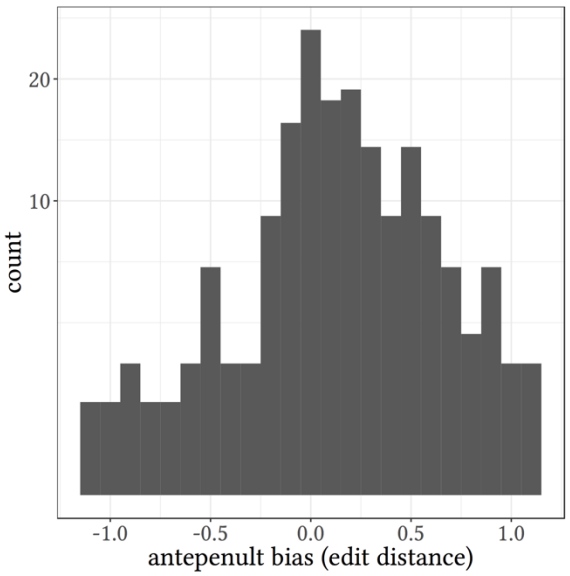

Figure 3.4. Histogram of the edit distance-based analogical bias measure (170 nonce probes). Positive values indicate test probes closer to the ten nearest antepenult- than penult-stressed lexical items.

Embedded Words

Although effort was made to minimize the embedding of shorter words in the stimuli, this could not be entirely avoided due to the large number of monosyllabic words in English.[6] Because spoken word recognition may involve activation of competing embedded forms (McQueen, 2004), there was a potential for such forms to influence stress placement strategies. For example, the word *mad* embedded in the test probe *madaplaz* might favor stressing the antepenult, whereas *gap* in *shigapleff* might push for penult stress. To control for this possibility, I counted the total number of

[6] Embedded words are a general property of the English lexicon, with the vast majority of polysyllabic word forms containing shorter words (Cutler et al., 2002).

lexical items contained by each nonword and subtracted the number of embeddings that favored antepenult stress from that of penult-stress cuing words. This procedure produced a measure of embedded word bias for penultimate stress, plotted in Figure 3.5.

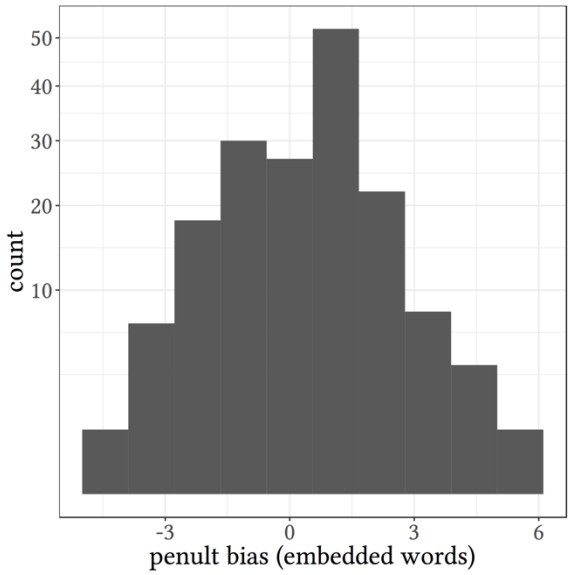

Figure 3.5. Histogram of the bias measure based on embedded words (170 nonce probes). Positive values indicate test probes for which more embeddings favored penultimate over antepenultimate stress.

The inclusion of phonotactic and nuisance predictors in the same set of models (as opposed to comparing the performance of phonotactics-only vs. nuisance-only models) assumes the position that these diverse sources of lexicon-based knowledge – phonotactic and otherwise – might compete with each other to influence behavior. This assumption is justified by the aforementioned findings suggesting that phonotactic and similarity-based metrics have independent effects on processing (e.g. Bailey & Hahn, 2001; Vitevitch & Luce, 1998; Storkel, Armbrüster & Hogan, 2006). Both edit distance and embedded words were thus featured in several models in chapter 5, which model stress assignment in perception and production. In the next chapter, however, I present

the results of two hyphenation studies where the nuisance predictors were not included.

CHAPTER IV

HYPHENATION STUDIES

Portions of the work presented in this chapter will be published as a coauthored article: Olejarczuk, P. & Kapatsinski, V. The metrical parse is guided by gradient phonotactics. To appear in *Phonology*.

4.1 Background

The two studies described in this chapter are both hyphenation tasks. Hyphenation was chosen because, unlike many of the word games reviewed in section 2.2.2, it does not isolate or otherwise transpose word parts in ways that expose them to word-edge effects and various other biases unrelated to syllable structure (see Côté & Kharlamov, 2011 for a review of the issues associated with these tasks). For example, partial repetition might bias speakers to produce closed syllables (at least in some instances) because English words must at minimum have two moras (see McCarthy & Prince, 1986). That said, hyphenation studies suffer from their own interpretation problems. These will be addressed in section 4.2.4, and chapter 5 will provide converging evidence from implicit tasks which are arguably more reliant on grammatical knowledge.

4.2 Study 1: Hyphenation of Pseudowords

4.2.1 Overview

Study 1 was a pen-and-paper variant of the hyphenation task, wherein participants syllabified pseudowords by inserting two slashes in between graphemes (see Redford & Randall, 2005 for a similar method).

4.2.2 Method

4.2.2.1 Participants

Forty-nine undergraduates participated in the study. All self-reported as monolingual, native speakers of American English with no reading difficulties and no prolonged exposure to another language.

4.2.2.2 Materials

The stimuli consisted of the 170 pseudowords described in section 3.3 (see also Appendix A). The items were presented orthographically, printed in 14- point, lower-case serif font on a sheet of paper.

4.2.2.3 Procedure

The experiment was administered individually in a laboratory setting. Each participant was given the sheet of paper containing a uniquely randomized list of all

170 test items. The participants were instructed to insert 2 slashes in each pseudoword with a pen, dividing it into 3 parts. No overt mention of syllables was made; the instructions simply asked for a division that seemed most 'natural' to the participants. The task took approximately 15 minutes to complete.

4.2.2.4 Data Pre-Processing

There were 8,330 responses in total (49 participants × 170 items). Of these, 283 (3.4%) were discarded because they constituted deviant parses, defined as yielding syllables with multiple vowels (.VCV. or .VCCV.) or with obstruents for nuclei (.C. or .CC.). An additional 254 responses (3% of total) parsed the embedded clusters entirely into the penult coda; because I was interested in complex onsets vs. splits, these responses were also excluded. The remaining 7,793 responses (93.6% of total) were included in the analysis.

4.2.2.5 Statistical Analysis

The dependent variable was the syllabification of the inserts located between the penultimate and final vowels (.CC vs C.C for clusters, and .C vs. C. for singletons). The predictors included word-initial insert status (singleton/attested CC/unattested CC), sonority slope, word-initial frequency of the insert, and word-final frequency of the singleton/C1 of the cluster.

All analyses were performed in R, using mixed-effects, logistic regression models constructed with the lme4 package (Bates et al., 2014). The models were fit by

the glmer() function, which uses the Laplace approximation and derives *p*-values from the normal distribution. In all multiple regressions, the continuous predictors were centered and scaled, enabling direct comparisons of the standardized coefficients. All mixed models featured maximal random effects (Barr et al., 2013); unless otherwise specified, this meant random intercepts for participant and frame, and random by-participant and by-frame slopes for all nested predictors. Additional details about individual model specifications are presented when necessary in the Results section.

4.2.3 Results

4.2.3.1 Coarse-Grained Phonotactics

I begin by examining the influence of coarse-grained phonotactics — namely, word-initial legality and onset maximization — on parsing intuitions. Of the items with singletons embedded between the second and third vowels, approximately 41% were parsed with the singleton belonging to the penult coda. For attested CC word onsets, this number increased to about 71%, while unattested CC onsets were split at a rate of 94%. These differences are summarized in Figure 4.1.

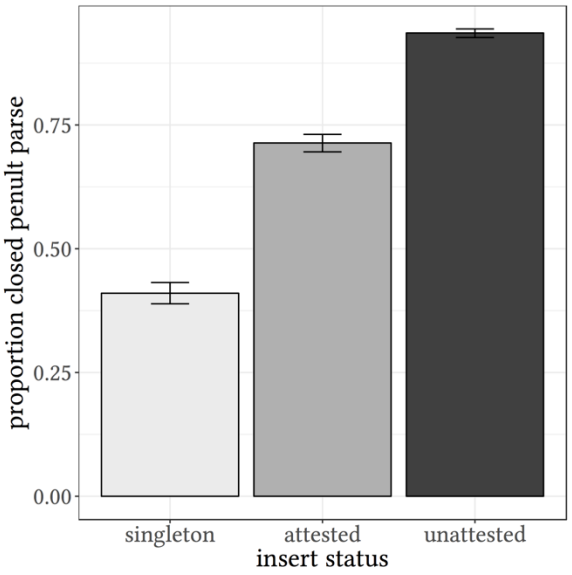

Figure 4.1. Closed penults by insert status. Error bars are 95% confidence intervals based on the proportion test.

To test for the significance of the pattern seen in the figure, a mixed-effects logistic regression predicting the penult rime structure (V vs. VC) from insert status was fit to the data. The predictor contrasts were coded using the treatment scheme, with singleton set to the reference level. The model featured by-participant and by-frame random slopes for insert status, as well as random intercepts for participant and CVCV__VC frame. A likelihood ratio test revealed that this model significantly outperformed a null version containing only random effects ($\chi^2(2) = 87.3$, $p < .001$). The model output is shown in Table 4.1.

As seen in the table, items with both attested and unattested clusters featured significantly higher rates of closed penults than did pseudowords with embedded singletons. For words with attested clusters, the odds of closing the penult were higher by a factor of 6.17 relative to words with embedded singletons. For the unattested:singleton pair, the odds-ratio was 96.02.

Table 4.1. Categorical model output (hyphenation task).

	Estimate (Std. Error)
Intercept (Status = singleton)	-0.535 (0.265)*
Status = attested	1.820 (0.208)***
Status = unattested	4.565 (0.328)***
Observations	7,793
Log Likelihood	-2,837.928
Bayesian Inf. Crit.	5,810.270
Note:	*p<0.05; **p<0.01; ***p<0.001

In order to test the difference between the two cluster types, a planned comparison was performed via another mixed-effects logistic regression. The model revealed that initially unattested clusters were indeed significantly more likely to be split than attested clusters, with the odds increasing by a factor of 15.86 (β = 2.76, *S.E.* = .26, p < .001).

These results support the long line of research arguing that word-edge phonotactics play some role in determining syllable boundaries: the finding that initially unattested clusters were much more likely than attested clusters to be split is consistent with the prior research reviewed above. At the same time, it is far from clear that the phonotactic generalizations which guide the parser are consistent with all assumptions of classical phonology. First, singletons were much more likely than attested clusters to be parsed as onsets of the final syllable. This finding, consistent with prior empirical work (see Eddington et al., 2013a,b *inter alia*), argues that onset maximization is not prioritized by the grammar nearly to the extent assumed by Pulgram (1970). Second and relatedly, the rate of closed penults among singleton items is surprisingly high — in spite of the requirement for filled onsets assumed in traditional theory (e.g. Clements & Keyser, 1983; Itô, 1989), over 40% of these inserts were parsed into the coda. This number is especially high given that all of the

singletons were obstruents, and thus should make much better onsets than codas according to the SSP.

The behavior of both singleton and attested CC inserts therefore shows more variability than expected under categorical assumptions. The unattested clusters were treated more uniformly by the participants, but strictly speaking their syllabification was not categorical either: about 6% were parsed as tautosyllabic onsets. I now turn to the question of whether any of the variability seen in the results can be explained by fine-grained phonotactic generalizations.

4.2.3.2 Fine-Grained Phonotactics

I begin by visualizing the correlations between each gradient predictor and the likelihood of a closed penult parse. Figure 4.2 shows the effect of word-initial frequency, with the data aggregated by insert. In order to avoid confounding frequency with phonotactic legality (all unattested clusters have zero frequency and could thus anchor the regression line), the data are restricted to singletons and attested clusters. There were 12 unique singletons and 28 unique attested CC onsets for a total of 40 data points. The scatter plot reveals a negative relationship: the more frequent an insert is in word-initial position, the less likely its initial (or, in the case of singletons, its only) consonant is to syllabify as a medial coda.

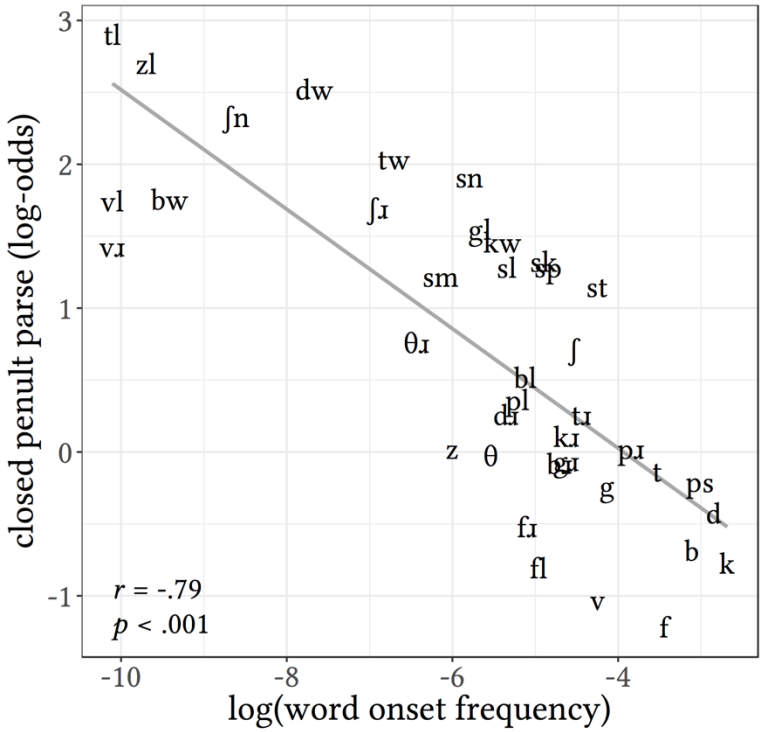

Figure 4.2. Log-odds of closed penults by initial frequency of each embedded insert, (singletons and attested CC onsets).

As shown in the lower-left corner of the panel, the correlation is statistically significant and relatively strong, with initial frequency capturing over 62% of the variance in the aggregated response data.

In order to test the influence of initial frequency on the parsing of pseudowords, a maximal, mixed-effects logistic regression was fit to the raw data. Again, since unattested onsets all shared a type frequency of zero, I conducted a more stringent test of the gradient hypothesis by excluding these items from the analysis and fitting the model to singletons and attested clusters only. Word onset frequency was found to significantly predict hyphenation behavior (β = -.69 *S.E.* = .07, p < .001), and the effect was in the direction seen in Figure 4.2: with each unit increase in initial frequency, the odds of splitting the cluster decreased by a factor of .50. In order to ensure that the effect was not driven by marginal onsets, I fit a second model to a subset of the data

with /ʃn, tl, vl, vɹ, zl/ excluded. The result was qualitatively unchanged, with initial frequency significantly predicting parsing behavior (β = -.78, *S.E.* = .08, p < .001)

The second gradient predictor under investigation was word-final frequency of the initial consonant of each insert. Figure 4.3 plots the correlation between this predictor and the log-odds of closing the penult. The correlation is statistically significant, with word offset frequency capturing 41% of the variance in the aggregate responses. The effect is in the expected direction, with consonants frequent in coda position more likely to be parsed as such by the participants. Note that, consistent with prior hyphenation studies, there appears to be a sonority effect, with sonorants more likely than obstruents to syllabify as codas. This effect is strongly correlated with offset frequency: with the exception of /m/, all sonorants are more frequent than all obstruents in word-final position.

Unlike word-initial probability, which is partly confounded with phonotactic legality, word-final frequency is in principle independent of insert status. For this reason, its influence on the parse was evaluated on the full set of inserts (as opposed to attested inserts only) with a maximal, mixed-effects logistic regression model. The effect of word offset frequency was significant (β = 2.49, *S.E.* = .49, p < .001): with each unit increase in offset frequency, the odds of closing the penult increased by a factor of 12.05. The effect persisted even after /ŋ/ was removed from the data (since this sound cannot begin a word, removing it represents a more rigorous test of gradience). Word offset frequency remained a significant predictor on the reduced data (β = 2.47, *S.E.* = .50, p < .001).

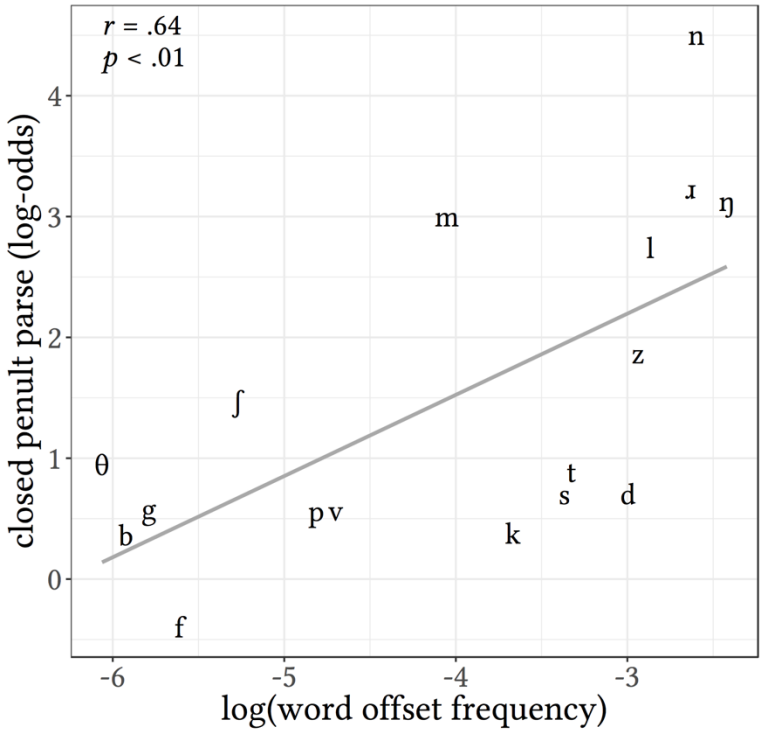

Figure 4.3. Log-odds of closed penults by word-final frequency of the initial consonant of each embedded insert.

The third gradient predictor investigated in this study is sonority slope. The correlation between this predictor and hyphenation behavior is plotted in Figure 4.1. Because this measure is correlated with insert status (no attested clusters feature negative sonority profiles), the dataset is limited to the 35 unique, initially unattested clusters. Recall from the discussion above that approximately 94% of these clusters were split — compared with the other insert types, there was relatively little variability in the responses. Nevertheless, the correlation is statistically significant, with sonority slope accounting for 28% of the variance in the aggregated responses. The effect is consistent with the SSP, with negative sonority profiles leading to a higher likelihood of a heterosyllabic parse.

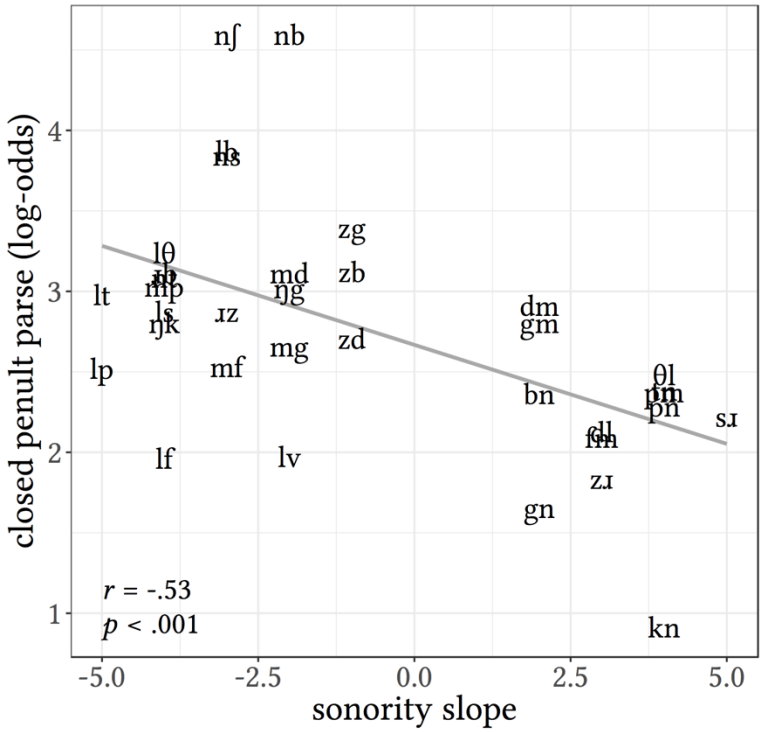

Figure 4.4. Log-odds of closed penults by sonority slope of each embedded insert (unattested clusters only).

To test whether sonority slope significantly predicted the parsing behavior, a mixed-effects logistic model was fit to the data. Again, due to the correlation between sonority slope and insert status, the gradient hypothesis was assessed by restricting the model to unattested clusters only. As with the word onset and offset frequency measures, sonority slope was centered and scaled, and the model included maximal random effects. The results revealed a significant effect of sonority on hyphenation behavior (β = -.36, *S.E.* = .08, p < .001). The effect was in the expected direction: with each unit increase in sonority slope, the odds of closing the penult decreased by a factor of .70.

Considered in isolation, each gradient predictor thus had a significant effect on hyphenation. In order to examine the joint performance of the measures, a multiple logistic regression model containing onset frequency, offset frequency and sonority

slope (residualized; recall section 3.4.1 for justification) was fit to the full data set. Each predictor was scaled and centered, and the model contained maximal random effects consisting of by-participant and by-frame slopes for every predictor as well as random intercepts for participant and frame. The model significantly outperformed a null version according to the likelihood ratio test ($\chi^2(3) = 88.32$, $p < .001$). The output is presented in Table 4.2, while the odds ratio estimates and marginal effects are plotted in Figure 4.5.

Table 4.2. Gradient model output (hyphenation task).

	Estimate (Std. Error)
Intercept	2.217 (0.244)***
Word Onset Frequency	-1.944 (0.143)***
Word Offset Frequency	0.389 (0.160)*
Sonority Slope	-0.334 (0.097)***
Observations	7,793
Log Likelihood	-2,712.643
Bayesian Inf. Crit.	5,640.349
Note:	*p<0.05; **p<0.01; ***p<0.001

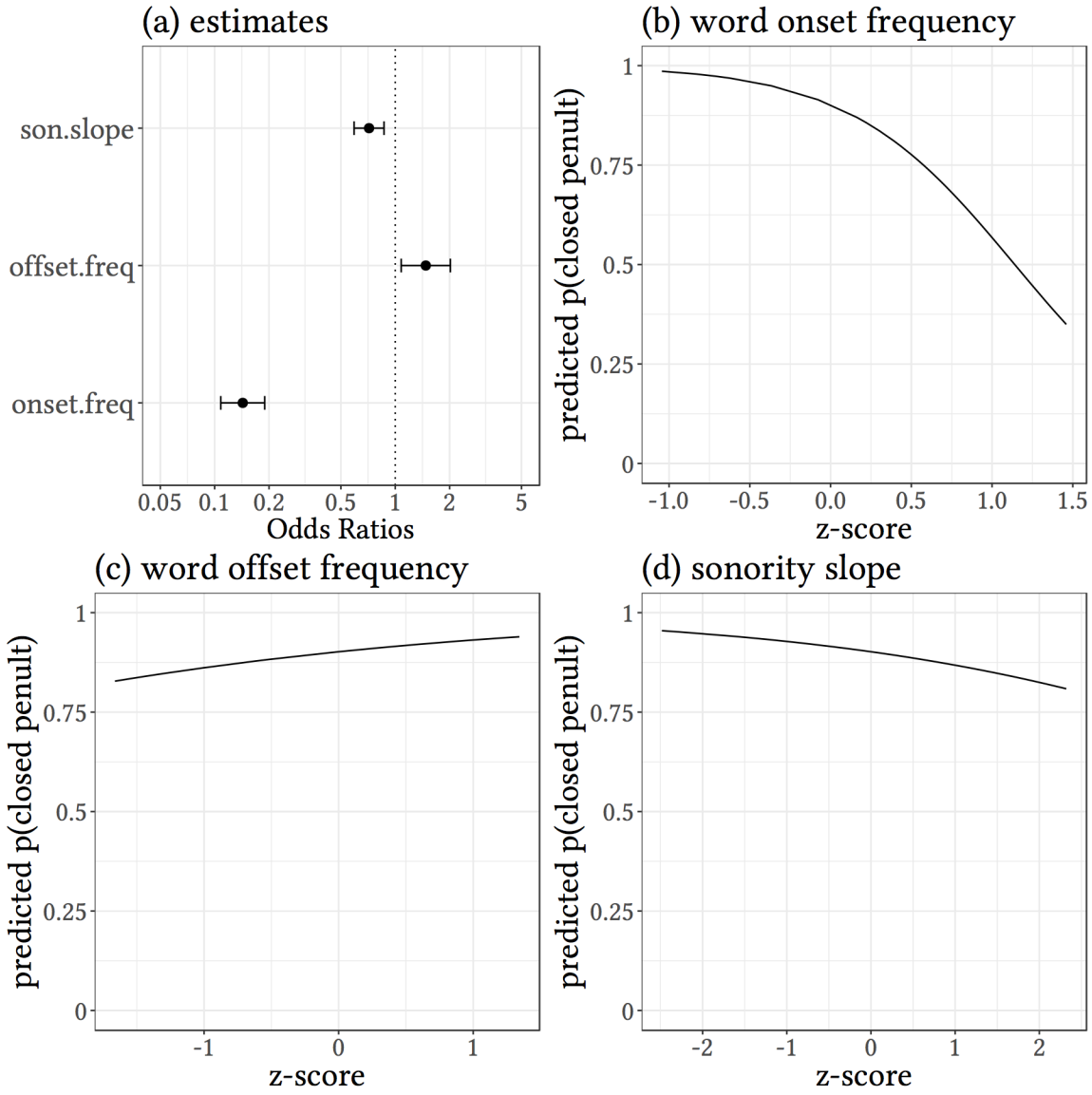

Figure 4.5. Gradient model estimates (panel [a]; dotted vertical line represents the null hypothesis) and marginal effects (panels [b]-[c]).

Each gradient predictor had a significant effect on hyphenation in the presence of the others. Relative to the grand mean, each unit increase in word onset frequency decreased the odds of closing the penult by a factor of .14. By contrast, increasing the offset frequency by one unit increased those odds by 1.48. Finally, for each unit increase in sonority slope (residualized), the odds of closed penults decreased by a factor of .72.

4.2.3.3 Model Comparison

The finding that lexical support and sonority each made significant contributions to predicting the hyphenation results suggests that syllable boundaries are computed in accordance to fine-grained phonotactic generalizations. In this section, I continue pursuing the question by evaluating the performance of the categorical parsing model (Table 4.1) relative to that of the gradient parsing model (Table 4.2). Both models were fit to the same data, but because they were non-nested, they could not be compared with a likelihood ratio test. Instead, two strategies for assessing relative fit were adopted. The first measured predictive accuracy on aggregate responses. First, predictions were generated from each model by conditioning on the fixed effects only. These were then averaged by insert and correlated with the actual responses. The scatterplots of predicted against observed values are displayed in Figure 4.6.

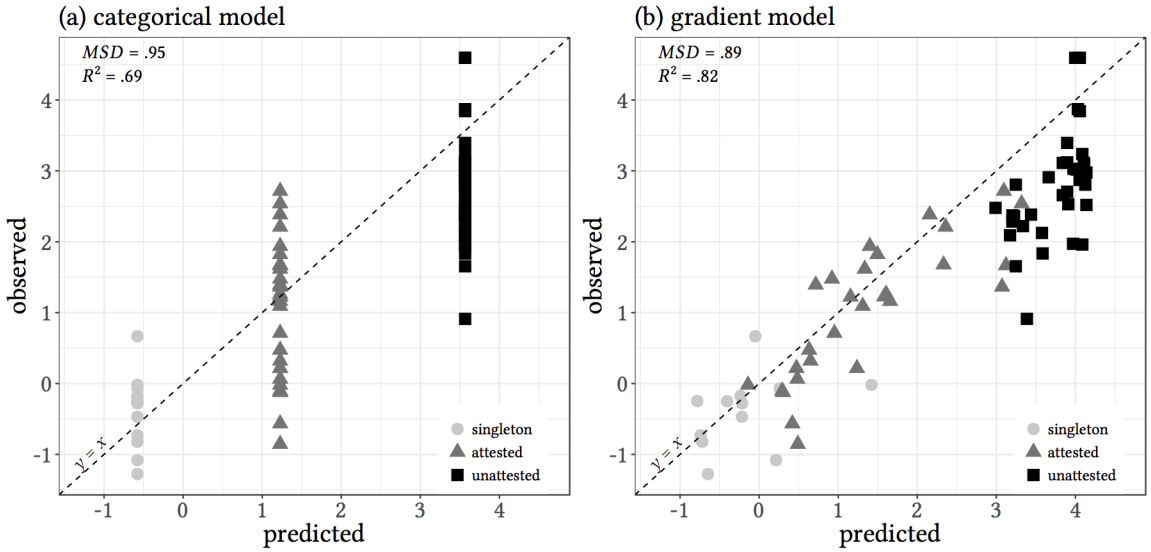

Figure 4.6. Comparison of model predictions (hyphenation task). Values are in log-odds.

There are 75 data points in each panel of the figure, each representing one insert. As expected, the categorical model generated predictions at the three distinct levels of insert status, while the predictions of the gradient model were more evenly distributed across the range of values. Both models appear to have somewhat over-predicted the probability of splitting in the unattested items (the dark squares are mostly below the dotted diagonal). However, the variability in the aggregated responses appears to have been better captured by the gradient model. The impression is supported by two statistical measures. First, the Mean Squared Deviation (MSD) between the observed and predicted values is higher for the categorical model, indicating higher prediction error. Second, the coefficient of determination (R^2) indicates that the gradient model accounted for 13% more variance in the aggregated responses.

The second model comparison strategy aimed to balance prediction with generalization. Including the random effects, the gradient model contained 24 free parameters whereas the categorical model contained only 9. It was therefore important to establish that the performance improvement was not due to overfitting. A common method of evaluating non-nested models is by comparing their scores on the Akaike Information Criterion (*AIC*) or the Bayesian Information Criterion (*BIC*). Both fit statistics penalize a model's maximum likelihood as a function of its complexity; here I use the BIC because it imposes a stricter penalty and thus puts the gradient model at a larger disadvantage. The BIC for model M_i is defined as

$$BIC(M_i) = -2\log L_i + k_i \log n,$$

where L_i is the model's maximum likelihood, k_i is the number of free parameters, and n is the number of observations. A lower *BIC* score indicates a better fit. Given its 24 free parameters, the gradient model received a complexity penalty of over 93 points, 62 more than the categorical model. Nevertheless, its *BIC* score was lower by 170 points than that of the categorical model (cf. Tables 4.1 and 4.2. To interpret the magnitude of this difference, I followed Wagenmakers (2007) in calculating the *BIC* approximation of the Bayes Factor and then calculating the posterior probability of the gradient model given the data. The Bayes Factor is simply a ratio of the posterior probabilities of the two models, under the assumption that the two models have equal prior probabilities, and was approximated for the gradient model by using the equation from Wagenmakers (2007:790):

$$BF_G \approx \exp(\Delta BIC_{CG} / 2),$$

where G and C stand for gradient and categorical, respectively, and $\Delta BIC_{CG} = BIC_C - BIC_G$. The Bayes Factor for the gradient model can be easily converted to its posterior probability:

$$p(M_G \mid D) = \frac{BF_G}{(1 + BF_G)}$$

Since BF_G was found to be approximately 7.9×10^{36}, the probability that a rational learner would choose the gradient over the categorical model was essentially equal to 1. In other words, the data strongly support the additional complexity contained in the gradient model.

4.2.4 Discussion

Overall, the results of Study 1 suggest that, to the extent that overt hyphenation recruits phonotactic knowledge, it is fine-grained rather than categorical generalizations that guide parsing behavior. To summarize, there are several pieces of evidence for this conclusion. First, when the data were stratified by word-initial legality of the inserts, initial insert frequency and sonority made additional contributions to predicting the hyphenation of legal and illegal onsets, respectively. In other words, there was gradience within each coarse-grained phonotactic category which was unaccounted for by traditional metrical theories. Second, singletons tended to be parsed as codas to the extent that they are frequent in word offset position. Apparently, it is not just knowledge of word onsets that transferred to the syllabification task; treatment of medial clusters appears to have been influenced by generalizations over both word edges. Third, when entered into a multiple regression model fit to the entire data set, all three gradient predictors made significant, independent contributions to parsing behavior: word onset statistics featured the largest estimated effect size, followed by approximately equal contributions from offset frequency and sonority slope. Finally, when the categorical and gradient parsing models were directly compared, the latter was shown to provide more accurate predictions. Importantly, comparison of the *BIC* scores revealed that this performance advantage was genuine and not due to overfitting.

The success of word edge statistics and sonority in predicting hyphenation behavior is consistent with the body of work on phonotactic well-formedness reviewed

in section 2.3. That is, the same sources of gradience which inform wordlikeness judgments and processing asymmetries in monosyllables appear to be implicated in judgments of syllable boundaries. Before the argument can be given much weight, however, there are a number of concerns about the generalizability of the Study 1 results. A number of these reference the shortcomings of metalinguistic tasks in general, which have been argued to reference sources of knowledge unrelated to the grammar (Goslin & Floccia, 2007; Smith & Pitt, 1999; Titone & Connine, 1997; Treiman et al., 2002). These objections will be addressed in chapter 5, which will present two experiments which utilize implicit tasks in perception and production, respectively. Here, I focus on two issues specific to the stimuli used in Study 1.

The first potential objection is that the pseudowords were presented orthographically. Generally speaking, orthographic knowledge has been argued to influence syllabification tasks independently of phonological knowledge. For example Treiman & Danis (1988) found that intervocalic singletons spelled with a double letter (*collar*) were more likely to elicit ambisyllabic responses than those spelled with a single letter (*color*). This was true not only in a written task where the participants were provided with alternate syllabifications, but also in an oral task involving syllable reversals. Somewhat more recently, Treiman et al. (2002) confirmed these findings with a partial repetition task. Their study investigated both children and adults, and found that the orthographic effects were present in 6th graders but not 2nd graders, suggesting that by the time learners reach moderate levels of literacy, knowledge of spelling begins to interact with grammatical knowledge in metalinguistic tasks.

While the pseudoword stimuli in Study 1 did not contain any double graphemes word-internally, the general concern about orthography remains valid. Specifically, the

issue lies with the uncertainty about how the vowel graphemes were interpreted by the participants. For example, given the orthographic nonce form *sibistoss*, hyphenation does not provide insight about whether the second vowel was interpreted as lax [ɪ] or tense [i]. As noted in section 4.1, vowel quality matters: syllabification studies employing real words have found that, all else being equal, lax vowels tend to attract codas and tense vowels may be more likely to attract onsets (Eddington et a.l., 2013a; Treiman & Danis, 1988; Treiman et al., 1994; Treiman et al., 2002, see also section 4.3.1). Although the vowels in the nonword stimuli were held constant across the coarse phonotactic categories of the clusters (recall section 3.3), it is entirely possible that variability in the interpretation of vowels contributed noise to hyphenation and potentially confounded the results.

The second major objection to the idea that the English metrical parse is gradient is that the pseudowords were not very similar to real English words (recall that, on average, the stimuli were about 5 edits away from the 10 closest lexical neighbors). As described in section 3.3, this similarity was intentionally avoided for the benefit of Studies 3 and 4 (see chapter 5), where it was important to discourage stress assignment by analogy to close neighbors. However, the low degree of similarity invites the criticism that the participants treated the stimuli as somehow deviant or exceptional. If the items were seen as very foreign, then participant behavior was potentially less constrained by the native grammar, and thus provides little insight into the phonological parser.

In order to address both of these concerns, it is important to show that the results obtained in Study 1 generalize to items that do not suffer from the potential shortcomings of our stimuli. Real English words of course fit this description: the

mapping between orthography and phonology is known to all literate speakers, so that — dialect differences notwithstanding — there is little ambiguity about the interpretation of vowel graphemes. Furthermore, real lexical items must by definition obey the native grammar, so there is no question of exceptional treatment. In the next section, I examine the gradient parsing hypothesis by reanalyzing the results of Eddington et al. (2013a,b), a large-scale hyphenation study of real English disyllables.

4.3 Study 2: Hyphenation of Real Words

4.3.1 Summary of Eddington et al. (2013a,b)

The megastudy by Eddington and colleagues constitutes the largest metalinguistic syllabification experiment conducted with English words to date. The test items consisted of 4,990 disyllabic words collected from the Hoosier Mental Lexicon (Pisoni et al., 1985). The participants were 841 native English speakers, most of whom were students at Brigham Young University. Each person syllabified a randomized list of 125 items, resulting in an average of 22 responses per word for a total of over 100,000 data points.

The format was an online survey where each trial provided a written word and asked the participants to choose from among quasi-phonemic, alternative parses. For example, given the word *victim*, the participants were provided with the following response choices:

- VI/KTUHM

- VIK/TUHM
- VIKT/UHM
- not sure

The results of the megastudy were released in two companion articles (Eddington et al., 2013a,b): one analyzed words with intervocalic singletons, while the other dealt with medial clusters of up to four segments. As in the three studies reported in the next chapter, Eddington et al. treated glides as consonants and rhoticized vowels as [Vɹ] sequences. In addition to the published analyses, the authors have made their data available to the public, preprocessed so that the responses were aggregated within words and across participants.[7]

In their investigation, Eddington and colleagues mainly focused on evaluating prior theoretical and empirical proposals about syllabification. As such, the potential factors considered in the analyses included previously hypothesized phonological, morphological and orthographic properties of the words. Categorical phonotactics and orthotactics were captured by coding the word-initial and word final legality of medial consonants and preceding vowels. Other predictors included consonant sonority, quality of the second vowel (tense vs. lax), stress placement, and the presence of morphological boundaries.

Similar to the approach taken here, Eddington et al. analyzed their responses using mixed-effects regression models. However, rather than entering all of the responses and predictors into one multinomial model, the authors first split the data

[7] Available to download at http://linguistics.byu.edu/faculty/deddingt/research%20data.html

into singleton- and cluster- containing words, and then fit a number of logistic regression models to each subset. None of the models included random slopes; for the most part, random intercepts for word and participant were included.

The singleton items were analyzed with three separate models. The first model contained morphological boundaries, the second featured categorical phonotactics, and the third categorical orthotactics. The authors' rationale for the split was that these predictors were too strongly correlated to be included in the same analysis. Each model also contained consonant sonority: the morphology model featured a four-level sonority scale (*rhotic* > *lateral* > *nasal* > *obstruent*) while the other two models made a two-way distinction between sonorants and obstruents. The remaining predictors were identical across the models: V1 legality in word-final position, initial vs. final stress, and V2 quality (tense vs. lax).

The analysis of words containing medial CC clusters[8] was divided along different lines. Rather than one multinomial regression with a three-level response variable, the authors fit separate logistic regressions to .CC, C.C and CC. responses. The set of predictors was identical across the models and included all of the variables present in the singleton models (the sonority predictor was binary, with obstruents opposed to sonorants). The .CC and CC. models featured random intercepts for both words and participants, while the C.C analysis could only converge with by-participant intercepts.

Across the singleton models, every predictor was found to significantly affect the syllabification choices. The participants preferred for syllable boundaries to

[8] Eddington et al also analyzed words with longer clusters; these results are not germane to the present study and will not be summarized here.

coincide with morpheme and word boundaries, stressed syllables were found to attract consonants to their margins, and tense second vowels attracted onsets. As for sonority, obstruents were generally more likely to be placed in onset position than were sonorants. In the .CC model, every predictor except orthographic onset legality was significant, with the effects being in the same direction as in the singleton analyses. The C.C parses were similarly affected, with the further exception of V2 quality. As for the CC. syllabifications, every predictor except word-final legality of V1 was found to have a significant effect.

Altogether, the Eddington et al. findings largely supported the prior accounts of syllabification the authors set out to evaluate, leading them to conclude that syllables are very 'word-like'. A close examination of their results further reveals that, among the factors examined, none turned out to be a categorical predictor of metalinguistic knowledge. Although singleton obstruents were more likely than sonorants to parse into onsets, they only did so with .80 probability. Vowels unattested in word-final position nonetheless closed medial syllables 64% of the time. Although tense second vowels were more likely than lax ones to attract onsets, both attracted medial singletons at rates of over 70%. A second conclusion could thus be reached: metalinguistic judgments of syllable boundaries reflect stochastic competition among generalizations over various phonetic and lexical properties. This conclusion was consistent with prior work (Fallows, 1981; Redford, 2008; Redford & Randall, 2005; Treiman & Danis, 1988; Treiman et al., 1992, 1994, 2002; Treiman & Zukowski, 1990,).

Crucially for the present purposes, however, there was unexplored variability within cluster sets defined by coarse-grained phonotactics. Specifically, only about half of initially legal CC were parsed as legal onsets, while unattested word onsets parsed as

such 94% of the time. Qualitatively, these findings are comparable to the pseudoword responses presented in section 4.2.3.1, and thus motivate a closer look at the phonotactic generalizations at play in the Eddington et al. data. Could it be that frequency explains some of the variance within legal clusters? The authors do not pursue this matter directly. However, in a follow-up model they find that some of the variability in the treatment of legal onsets can be attributed to sC clusters, which were more likely than the others to be split. While this is consistent with theoretical treatments of the initial /s/ as an affix (Kaye et al., 1990), sC clusters are also not among the most common word onsets (cf. Figure 4.2), suggesting a role of frequency.

The authors also under-explore the effect of sonority. As noted above, most of the singleton models featured a binary split between obstruents and sonorants, with only the morphology model coding sonority into a four-level scale. Unfortunately, the authors did not report follow-up, simple comparisons to that model, so it is unclear which sonority levels differed from each other (it is possible that the effect was entirely driven by obstruents vs. sonorants). The CC models considered only the obstruent/sonorant distinction in the initial consonant of the cluster. Ignoring the second consonant made it impossible to assess the scalar effect of the SSP on the well-formedness of complex onsets.

In the remainder of this chapter, I conduct a partial reanalysis of the Eddington et al. megastudy. Rather than challenge all of their conclusions, the aim is to simply take a closer look at the phonotactics involved. In line with the logic of this dissertation, the aim is to investigate whether gradient phonotactic knowledge can account for some of the variance in the results, just as they did in the nonword hyphenation experiment. As in Study 1, I compare the performance of a categorical

phonotactic model to that of a gradient phonotactic model in accounting for the data. To these ends, the results of Eddington et al. (2013a,b) were subjected to an analysis that largely parallels that presented in Study 1. To anticipate the results, the Eddington et al. data pattern in remarkably similar ways to the findings of Study 1. In spite of the differences in stimulus properties between the two studies (size, lexical status), the gradient model captures both data sets better than the categorical model, strengthening the conclusion that fine-grained, lexicon-derived phonotactic knowledge is the appropriate source of stochastic generalizations responsible for the emergence of syllable-like representational units.

4.3.2 Method

In order to facilitate the comparison to Study 1, the analysis was restricted to a subset of the data collected by Eddington et al. (2013a,b). Specifically, words with medial clusters longer than two consonants were excluded, so that the remaining inserts matched the C and CC inserts from Study 1 in length. As in Study 1, CC. responses were also excluded. Thus reduced, the data set consisted of 83,131 responses to 3,868 unique, disyllabic words. Of these, 2,297 items contained medial singletons, 441 featured CC inserts attested as word onsets, and the remaining 1,148 had initially illegal CC clusters embedded between the two vowels.

The total number of insert types was 232. These consisted of 23 singletons, 46 attested CC word onsets, and 163 unattested CC word onsets. Of the 75 inserts analyzed in Study 1, 67 were also present in the Eddington et al. data. All 12 singletons from Study 1 were represented, as were 26 of 28 attested word onsets and 29 out of 35

unattested word onsets. For inserts unique to the Eddington experiment, sonority slope was calculated as in Study 1 and the word-edge statistics were based on the same lexicon.

As in the pseudoword hyphenation task, the data were analyzed with mixed-effects logistic regression models. The response variable was binary, representing the two .(C)C and C.(C) parsing options. Because the Eddington et al. results were pre-aggregated across participants, it was not possible to match Study 1 and include participant-based random effects in the models. Furthermore, unlike the frames used in the pseudoword study, each word only contained one insert. For these reasons, the random effects structure of the models reported below contained only by-word intercepts.

4.3.3 Results

4.3.3.1 Coarse-Grained Phonotactics

This section examines the effect of the categorical predictor (insert status) on the parsing results of the Eddington et al. (2013ab) participants. Approximately 26% of singleton items were parsed with the singleton belonging to the penult coda. For words with inserts consisting of attested CC onsets, about 44% were split. For unattested CC onsets, the number rose to about 96%. These results are displayed in Figure 4.7.

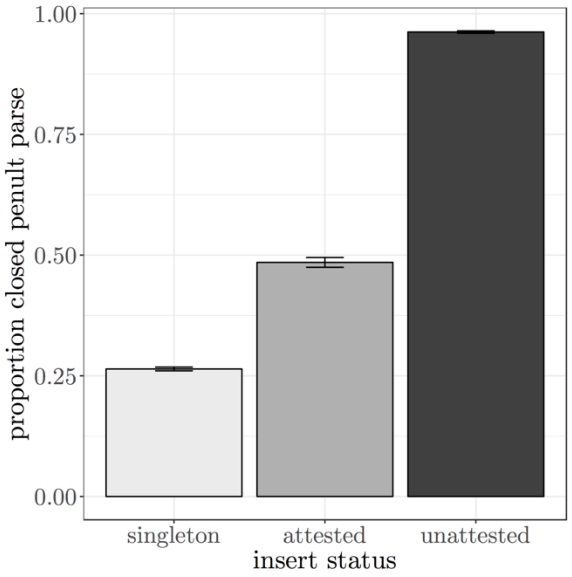

Figure 4.7. Closed penults by insert status (Eddington et al., 2013a,b study). Error bars are 95% confidence intervals based on the proportion test.

To test for the effect of insert status on penult rime structure, a mixed-effects logistic regression with maximal random effects was fit to the data. The model significantly outperformed an intercept-only version according to the likelihood ratio test ($\chi^2(2) = 3{,}762.6$, $p < .001$). The output is shown in Table 4.3.

Table 4.3. Categorical model output (Eddington et al., 2013ab data).

	Estimate (Std. Error)
Intercept (Status = singleton)	-1.264 (0.035)***
Status = attested	1.140 (0.083)***
Status = unattested	5.233 (0.074)***
Observations	3,868
Log Likelihood	-9,196.717
Bayesian Inf. Crit.	18,426.480
Note:	*p<0.05; **p<0.01; ***p<0.001

With singleton items set as the reference category, words with both initially attested and unattested clusters featured significantly higher rates of closed penults. In the case of initially attested CC clusters, the odds of closing the penult increased by a

factor of 3.13 over singletons. The odds ratio of unattested clusters to singletons was 187.34.

A follow-up model comparing the two cluster types found that unattested clusters had significantly higher odds of being split than attested onsets by a factor of 93.66 (β = 4.54, *S.E.* = .14, p < .001).

4.3.3.2 Fine-Grained Phonotactics

As in the analysis of the hyphenation study presented in Section 4.2.3.2, I begin by plotting each gradient predictor against aggregated responses. Figure 4.8 shows the effect of word onset frequency on inserts attested in word-initial position.

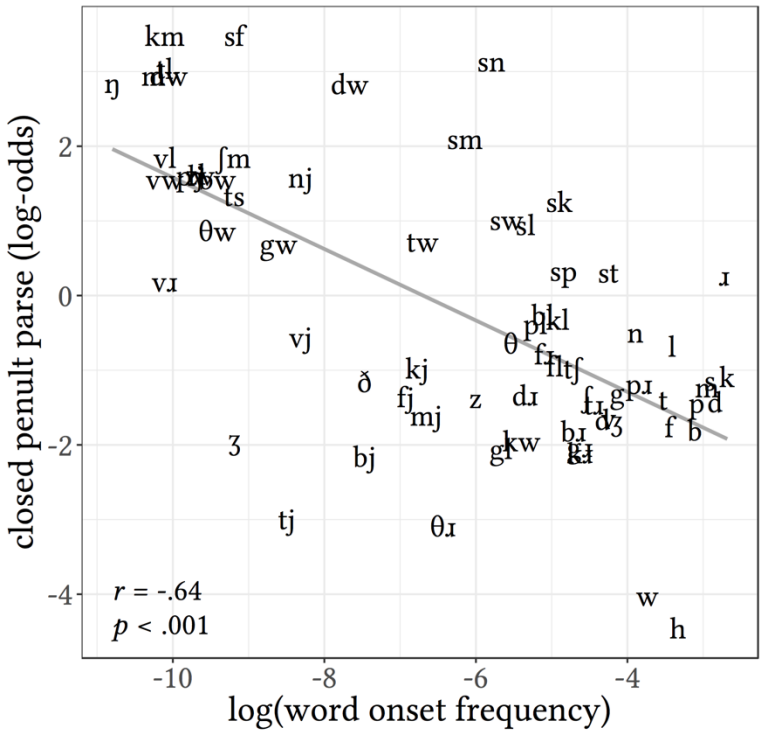

Figure 4.8. Log-odds of closed penults by word-initial frequency of each embedded insert in the Eddington et al. (2013a,b) data (singletons and attested CC onsets).

In total, there were of 23 singletons and 46 attested CC clusters embedded between the first and second vowel in the words used by Eddington et al. (2013ab), for a total of 68 data points. The correlation was statistically significant, with word onset frequency accounting for about 41% of the variance in the aggregate responses. The effect was in the expected direction, with frequent word onsets resisting the penult parse.

To test the effect of word onset frequency on the parsing judgments of the Eddington et al. (2013) participants, a maximal, mixed-effects logistic regression was fit to the responses to singleton and attested clusters. The model revealed a significant effect of onset frequency (β = -.26, *S.E.* = .02, p < .001): with each unit increase in word onset frequency, the odds of closing the penult decreased by a factor of .77.

As a more stringent test of gradience, a second regression was run on a subset of the data which excluded /ŋ/ (since it cannot be a word onset), /w/ and /h/ (which cannot end a word), and seven marginal CC onsets. The effect of onset frequency remained significant and in the same direction (β = -.19, *S.E.* = .02, p < .001).

Figure 4.9 plots the correlation of responses with word offset frequency. As in section 4.2.3.2, the responses are collapsed across items and participants and averaged by the initial consonant of each embedded insert. There were 23 unique consonants occupying this position. Of these, word-finally illegal /w, h/ were excluded from the plot, leaving 21 unique data points. The correlation was significant, with word offset frequency accounting for about 36% of the variance in the aggregate responses. The effect was in the expected direction, with consonants frequent in coda position more likely to be parsed as such by the participants in the Eddington et al. (2013a,b) study.

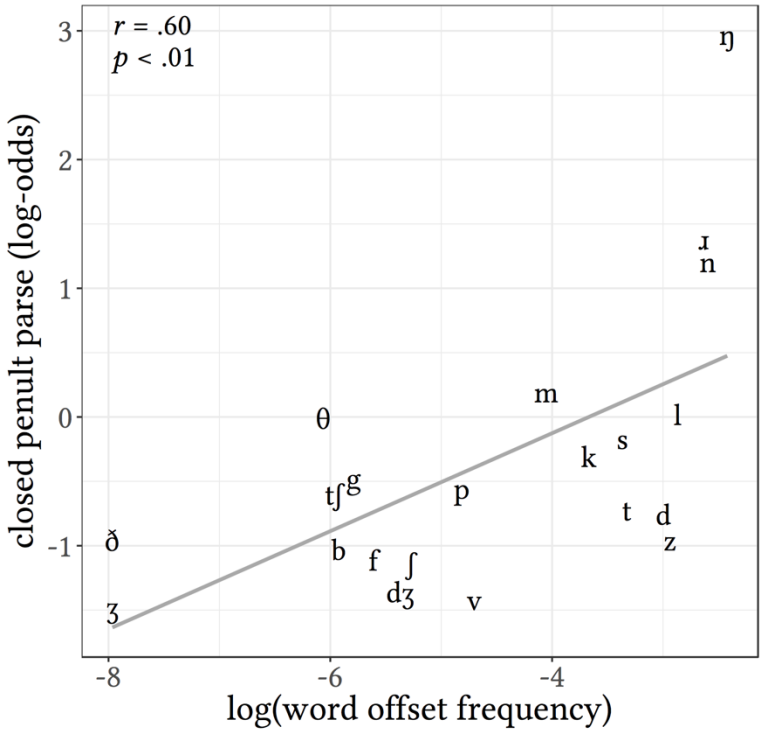

Figure 4.9. Log-odds of closed penults by word-final frequency of the initial consonant of each embedded insert (Eddington et al., 2013a,b data).

The results of a maximal, mixed-effects logistic regression fit to the entire data set (minus /w, h/) confirmed the pattern seen in the figure. Word offset frequency was found to significantly predict hyphenation judgments (β = .84, *S.E.* = .04, p < .001). With each unit increase in offset frequency, the odds of closing the penult increased by a factor of 2.33. The effect of word offset frequency persisted in a reduced data set which excluded inserts beginning with /ŋ/ (β = .81, *S.E.* = .04, p < .001), indicating that it was not driven by categorical preferences.

The correlation between the responses to items with initially unattested inserts and sonority slope are plotted in Figure 4.10. The participants in the Eddington et al. (2013a,b) study overwhelmingly preferred to split these clusters. Even so, the correlation is significant, with sonority slope accounting for about 17% of the

aggregated responses. The effect is consistent with the SSP, with rising sonority profiles somewhat more resistant to the heterosyllabic parse.

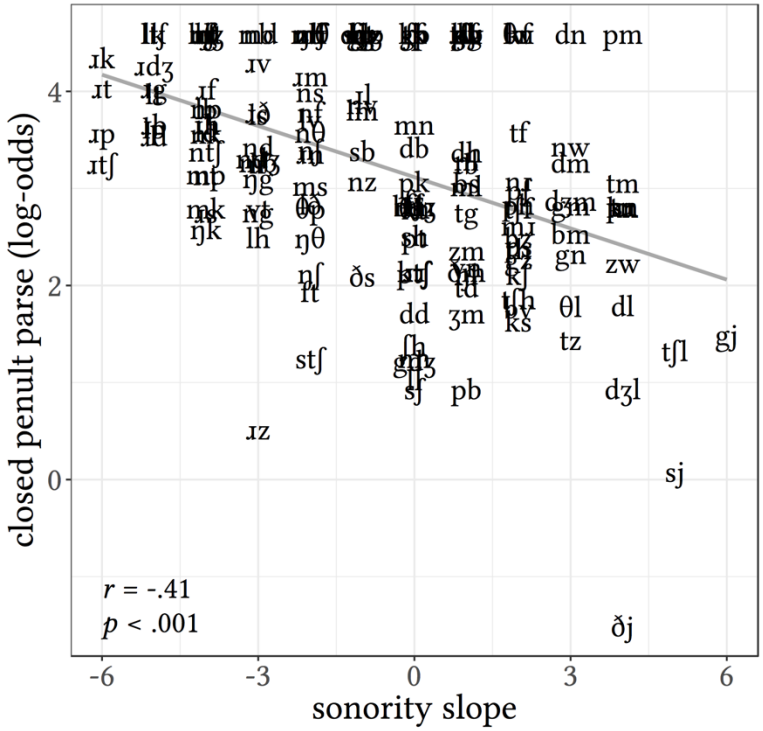

Figure 4.10. Log-odds of closed penults by sonority slope of each embedded insert in the Eddington et al. (2013a,b) data (unattested clusters only).

A mixed-effects, logistic regression model fit to the unattested items revealed a significant effect of sonority (β = -.25, *S.E.* = .02, p < .001). With each unit increase in sonority slope, the odds of closing the penult were reduced by a factor of .78.

On their own, each gradient measure had a significant effect on the Eddington et al. (2013a,b) results. In order to examine their performance in the presence of each other, a multiple logistic regression model containing onset frequency, offset frequency and sonority slope was fit to the full data set. As in the Experiment 1 analysis, sonority slope was residualized on the two lexical support measures, all predictors were centered and scaled, and the model featured maximal random effects. A likelihood ratio test

revealed that the model was a significant improvement over an intercept-only version ($\chi^2(3) = 4{,}496.9$, $p < .001$). Table 4.4 shows the output and Figure 4.11 plots the odds ratio estimates and marginal effects.

Table 4.4. Gradient model output (Eddington et al., 2013ab data).

	Estimate (Std. Error)
Intercept	0.411 (0.027)***
Word Onset Frequency	-2.183 (0.030)***
Word Offset Frequency	0.718 (0.027)***
Sonority Slope	-0.559 (0.026)***
Observations	3,868
Log Likelihood	-8,829.550
Bayesian Inf. Crit.	17,700.400
Note:	*p<0.05; **p<0.01; ***p<0.001

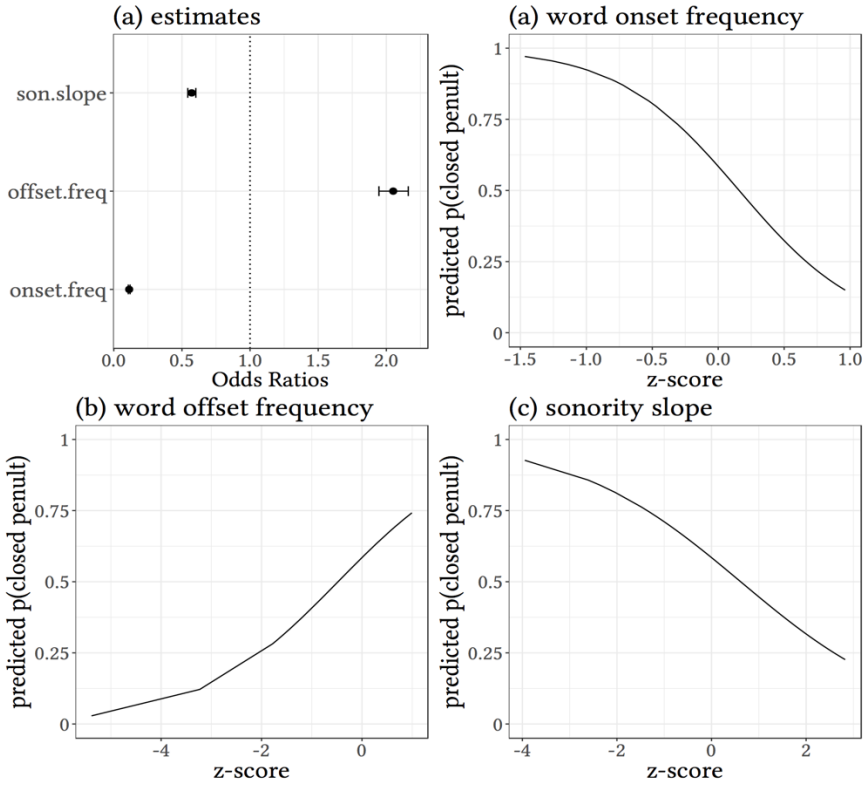

Figure 4.11. Marginal effects of gradient model predictors.

All three predictors had a significant effect on the responses, with word onset frequency returning the largest effect size. Each unit increase in word onset frequency decreased the odds of closing the penult by .11 while the same change in sonority slope decreased the odds by .57. Increasing word offset frequency by one standard deviation increased the odds of closed penults by 2.05.

4.3.3.3 Model Comparison

This section compares the performance of the categorical versus gradient parsing models on the Eddington et al. (2013a,b) data. As with Experiment 1, two comparisons were made: the first checked the predictive accuracy of each model on aggregate responses, while the second computed posterior probabilities based on the *BIC* approximation of the Bayes Factor (see section 4.2.3.3 for a description of this procedure).

Beginning with the aggregate responses, Figure 4.12 plots the by-insert predicted versus observed values for each model. As in the hyphenation task, the predictions were generated by conditioning on the fixed effects only.

There are 232 data points in each panel, each representing a unique insert in the Eddington et al. (2013a,b) word list. The categorical model predicted that 22% of words with embedded singletons should be parsed with a closed penult. For words with attested and unattested CC inserts, the predicted rates were 47% and 98% , respectively. The predictions of the gradient model were mode evenly distributed, with most of the values falling between 7% and 98%.

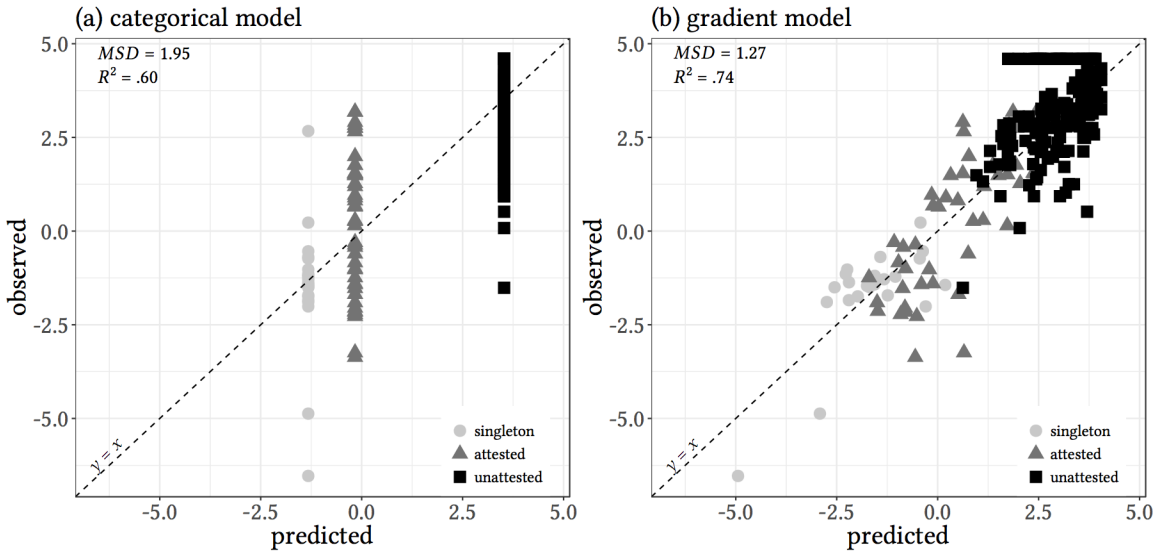

Figure 4.12. Comparison of model predictions (Eddington et al., 2013ab data). Values are in log-odds.

The additional level of detail available to the gradient model conferred a predictive advantage. The mean squared deviation was considerably lower than that of the categorical model (1.27 vs. 1.95), indicating a closer correspondence between the aggregate predictions and observations. Comparison of the R^2 values indicated that the gradient model accounted for approximately 14% more variance in the aggregate responses.

That said, there was a number of inserts for which the categorical model yielded slightly better predictions. The property these had in common was that the words which contained them invariantly led to a closed penult parse. The categorical model predicted this parse with a probability of .98 for all of these items, but the gradient predictions ranged between 86% and 99% (note the horizontal "bar" in the upper-right corner of Figure 4.12b). A closer inspection of these inserts revealed that the vast majority of them were instantiated in a single English word (unique for each insert), making it impossible to tease apart the influence of word-level from insert-level

properties on the aggregate responses to these items. In a follow-up comparison, all inserts with type frequency of 1 (67 out of 232) were removed from the analysis. The resultant R^2 values were .72 for the categorical model and .83 for the gradient model. Thus, both models yielded better predictions on the reduced data set, but the gradient model maintained its advantage.

The second type of model comparison was based on the *BIC* scores. Having only 1 predictor, the categorical model was simpler and thus incurred a smaller likelihood penalty than the gradient model. Nevertheless, the categorical model had a higher *BIC* score (18,426) than the gradient model (17,700), indicating worse fit to the data. This 363-point difference resulted in a Bayes Factor in excess of 4.6×10^{157} for the gradient model, yielding a posterior probability essentially equal to one. In other words, given the availability of both models, the gradient model is virtually always better justified by the data. In chapter 7 I will show that this *BIC* advantage holds for unbiased learners of the lexicon regardless of vocabulary size.

4.3.4 Discussion

This reanalysis of Eddington et al. (2013a,b) was motivated by the need to test the generalizability of the Study 1 findings and to address potential objections to the pseudowords used in that task. Overall, the results of the two studies were remarkably consistent. As in Study 1, the hyphenation preferences in the Eddington et al. task were influenced by fine-grained phonotactic generalizations. The similarities persisted in every analysis. First, word onset frequency and sonority slope made significant contributions to predicting the hyphenation of real words with medially-embedded

legal and illegal word onsets, respectively. Second, word offset frequency affected the probabilities with which the C1 of each insert was placed into the penult coda. Third, each gradient predictor was found to make a significant, independent contribution in a multiple regression model fit to the entire dataset, with word onset frequency having the largest estimated effect (see standardized coefficients inTable 4.4 and odds-ration plots in Figure 4.11a). Finally, comparison of the categorical and gradient parsing models revealed a significant advantage for the latter: its predictions were closer to the observed human behavior, and its *BIC* score indicated that this advantage was not due to the inclusion of more parameters.

The fact that the hyphenation of two very different types of stimuli — trisyllabic nonwords on one hand and real, English disyllables on the other — appears to have been guided by similar phonotactic generalizations lends important converging evidence for the gradient parser hypothesis. However, it also raises an interesting issue about the original analyses presented in Eddington et al. (2013a,b). As detailed in section 4.3.1 above, those models featured categorical phonotactics in addition to a number of other predictors, including morphological boundaries, vowel quality and stress. Would the substitution of gradient phonotactics improve fit above and beyond those variables? Would the gradient parsing model presented here outperform the original analysis? In other words, did word-edge statistics and sonority slope serve as actual cues to hyphenation, or did their effects arise as epiphenomena through correlations with the factors investigated by Eddington and colleagues? Unfortunately, the authors have not made their predictor coding publicly available, making it difficult if not impossible to perform a fair model comparison between the original analyses and the gradient parsing model advocated here. That said, the fact that the same

phonotactic generalizations captured behavior in nonwords does lend support to the validity of the gradient parsing model, since the nonwords did not have morphological boundaries.

Although the consistency between the two hyphenation studies is encouraging, both experiments suffer from additional interpretation problems common to all hyphenation tasks. First, several researchers have warned that the metalinguistic knowledge tapped in such tasks applies at relatively late stages of stimulus processing, where school-taught rules of written word division and other orthographic conventions may obscure the nature of the underlying phonological representations (Goslin & Floccia, 2007; Smith & Pitt, 1999; Titone & Connine, 1997; Treiman et al., 2002). Second, it is not clear that the participants in these tasks are parsing out syllables rather than possible words. The distinction is important: it is well known that across languages, word edges are not always coextensive with internal syllable edges (e.g. Broselow, 2003; Hammond, 1999; Pierrehumbert, 1994), and several authors have explicitly cautioned against interpreting all word-edge phenomena as indicative of syllable properties in general (Côté & Kharlamov, 2011; Davis, 1989; Frisch, 2000; Harris, 1994; Kaye et al., 1990; Kessler & Treiman, 1997; Pierrehumbert, 1994; Pierrehumbert & Nair, 1995). It is thus possible that fine-grained, word-edge phonotactic knowledge is relevant to word segmentation but not necessarily to syllabification.

The argument for a fine-grained metrical parse would thus be bolstered by converging evidence from a task that specifically and unambiguously targets internal syllable boundaries. Furthermore, an implicit task would offer higher ecological validity than one where behavior is potentially mediated by metalinguistic introspection. A number of such tasks have been applied to the question of syllables over the years. For

example, Treiman et al. (1994) examined blending errors in a short-term memory task that required participants to memorize nonsense CVCVC strings under cognitive load (as mentioned above, the error patterns supported the influence of stress, vowel quality and sonority on the syllabic affiliation of the intervocalic singletons). Taking a different approach, Titone & Conine (1997) employed a phonological priming task to pit the influence of the Maximal Onset Principle against that of stress (the former was found to be more influential). Smith & Pitt (1999) investigated the interaction between phonotactic legality, onset maximization and morphology using a variant of the phoneme monitoring paradigm. The authors found that legality trumped maximizing onsets, and that phonology influenced earlier stages of processing than did morphology.

While speech error analysis, structural priming and phoneme monitoring represent well-established psycholinguistic methods, one could argue that they share one property with metalinguistic tasks like hyphenation: no person performs anything like them in daily life. Since a general goal of this dissertation was to investigate the natural deployment of phonotactic knowledge, a different approach was sought. Fortunately, there is a well-documented relationship between English stress and syllable structure: as a quantity-sensitive language, English has been argued to preferentially assign stress to heavy over light syllables (Hayes, 1982; Kager, 1989). As noted in section 2.1.1, most accounts of weight-sensitive stress assume that syllabification precedes stress assignment because syllable structure must be available for weight computation. This means that reversing the directionality provides a window on syllabification: stress assignment can be used to infer syllable structure. In

the next chapter, I exploit this relationship to probe the nature of the metrical parse from another angle.

CHAPTER V

STRESS ASSIGNMENT STUDIES

Portions of the work presented in this chapter will be published as a coauthored article: Olejarczuk, P. & Kapatsinski, V. The metrical parse is guided by gradient phonotactics. To appear in *Phonology*.

5.1 Background

The role of syllable weight is widely acknowledged in formal accounts of English stress (Halle, 1998; Halle & Vergnaud, 1987; Hayes, 1982, 1995; Kager, 1989; Liberman & Prince, 1977; Prince, 1991). The traditional view holds that, in non-final syllables, stress assignment is sensitive to a binary weight distinction carried by the rime: light rimes consist of a lax vowel (V̆) and so carry a single mora, whereas heavy rimes contain at minimum a tense vowel, a diphthong, or a coda (VX), making them bimoraic. In weight-sensitive systems, heavy syllables attract stress, and in the case of English this is perhaps most clearly exemplified by the well-known Latin Stress Rule, which captures much of the Latinate vocabulary that entered the English lexicon following the Norman conquest (Halle & Keyser, 1971). According to this rule, main stress in trisyllabic and longer words tends to fall on the penult if it is heavy, else it falls on the antepenult.

Under the Hayesian version of metrical theory, Latin Stress follows from the interaction of foot type, edge alignment and extrametricality: bimoraic trochees are

constructed right-to-left, skipping the final syllable unless its rime is 'superheavy' (VVX); main stress is then assigned to the head of the rightmost foot. To illustrate this phenomenon, consider the words *stamina* and *cicada*, which feature CV̆ and CVV penults, respectively. Their metrical parses are shown below (by convention, syllable boundaries are indicated by periods, feet enclosed by parentheses and extrametrical syllables contained within angle brackets).

(5.1) a. (ˈstæ.mɪ.)<nə> b. sɪ.(ˈkeɪ.)<də>

As seen in example 5.1, the light penult in *stamina* foots with the preceding syllable, whereas the bimoraic penult of *cicada* parses into its own trochee. The difference in stress follows from the fact that trochees are left-headed.

In classical Optimality Theory, Latin Stress is captured with a strict ranking of a number of metrical constraints, which are used to evaluate competing outputs and select the winning candidate eventually produced by the speaker. One representative constraint set is presented in example 5.2 below, followed by ranking tableaux (Table 5.1) that account for the stress patterns in *cicada* and *stamina*, respectively. Constraint rankings are represented left-to-right in the tableaux: the further left a constraint appears, the higher its ranking. A candidate wins if (a) the highest constraint it violates is ranked below at least one constraint violated by some competitor, or (b) the competitor incurs more violations of the same constraint than the winner.

(5.2) A typical set of metrical constraints (see Tesar & Smolensky, 2000)

TROCHEE: Feet are left-headed

IAMB: Feet are right-headed

NONFINAL: Final syllable is extrametrical

FOOT BINARITY (FTBIN): Feet contain either two syllables or two moras

ALIGN FOOT-R (AFR): The right edge feet and words are aligned

WEIGHT-TO-STRESS PRINCIPLE (WSP): Heavy syllables are stressed

PARSE: All syllables parse into feet

Table 5.1. Constraint rankings that produce the correct outputs for *cicada* and *stamina*.

/sɪ.keɪ.də/	TROCHEE	NONFINAL	FTBIN	AFR	WSP	IAMB	PARSE
→ sɪ.(ˈkeɪ.)də				*	*		**
sɪ.(ˈkeɪ.də)		*!			*	*	*
(sɪ.ˈkeɪ.)də	*!			*	*		*
(ˈsɪ.keɪ.)də				*	**!	*	*
(ˈsɪ.)keɪ.də			*!	**	**		**
sɪ.(keɪ.ˈdə)	*!	*			*		*
sɪ.keɪ.(ˈdə)		*!			*		**
/stæ.mɪ.nə/							
→ (ˈstæ.mɪ.)nə				*	*	*	*
(ˈstæ.)mɪ.nə			*!	**	*		**
(stæ.ˈmɪ.)nə	*!			*	*		*
stæ.(ˈmɪ.)nə			*!	*	*		**
stæ.(ˈmɪ.nə)		*!			*	*	*
stæ.(mɪ.ˈnə)	*!	*					*
stæ.mɪ.(ˈnə)		*!					**

Although weight is associated with English stress, it is by no means completely predictive. The lexicon contains a multitude of other surface patterns, many of which compete with weight sensitivity and with each other (see Kager, 1989 for a summary). For example, it is well-known that disyllabic nouns tend to behave differently from

verbs, with the former more likely to feature initial stress regardless of syllable weight (*récord* vs. *recórd*). In addition, the Latinate vocabulary often patterns separately from words of Germanic origin, which are weight-insensitive and often stressed on the root-initial syllable (Lahiri, Riad & Jacobs, 1999). Morphology also plays a role: lexical compounds tend to be stressed on the leftmost constituent (*gréenhouse*, not **greenhóuse*; Liberman & Prince, 1977; Hayes, 1995) and suffixes vary in whether they attract, shift or ignore stress (e.g. Alcántara, 1998). Collapsing across morphology, etymology and word class thus yields many surface exceptions to the Latin Stress Rule, with stress often landing on light penults (*narcótics*, *specífic*, *tobácco*, *unpléasant*) or skipping heavy ones (*ánarchy*, *députy*, *hándicap*, *pólygon*).

Nevertheless, experimental evidence suggests that English speakers acquire the weight generalization and appear to do so at an early age. For example, 9-month-old infants show a preference for initial, stressed syllables to be heavy (Turk et al., 1995), 22-month-old toddlers often incorrectly shift stress onto final syllables when these are heavy (Kehoe, 1998), and 5-year-old children are able to productively extend weight sensitivity to nonwords, stressing CVV.CVC probes on the initial syllable at higher rates than CV.CVVC items (Redford & Oh, 2015).

By the time they reach adulthood, English speakers exhibit quite sophisticated knowledge of the various stress patterns in the lexicon. For instance, Guion et al. (2003) demonstrated that adults rely on a number of strategies when assigning stress to nonce forms, including sensitivity to syllable weight, lexical class, and analogy to known words (see also Baker & Smith, 1976). Similarly, Ernestus & Neijt (2008) reported that English speakers are sensitive to the interaction of weight and word length present in the CELEX lexical database (Baayen, Piepenbrock & Gulikers, 1995), stressing heavy

penults in quadrisyllabic pseudowords more often than in trisyllables. In a similar study conducted by Domahs, Plag & Carroll (2014), productive extension of the Latin Stress pattern was modulated by the weight of the final syllable in a way that qualitatively resembled the stress distribution in CELEX. The identity of the final vowel also appears to be extended from the lexicon: Moore-Cantwell (2016) showed that nonsense trisyllables ending in [-i] had a stronger preference for being stressed on the antepenult than nonwords ending in [-ə]. Finally, the probability of stressing a syllable in a nonword seems to rise monotonically as rime complexity increases along a V̆ < V̆C < VV < VVC continuum, and onset structure may have a secondary, cumulative effect (Kelly, 2004; Ryan, 2011a, 2014; see section 5.4.4.1.1 for more discussion). Such findings challenge the traditional notions that English weight is binary and exclusive to the rime.

Taken together, these results suggest that, much like phonotactics, weight sensitivity forms part of English speakers' internal model of the language. Moreover, like phonotactic grammars, the stress grammars acquired by learners are considerably more complex than predicted by classical phonological theory. Studies that utilize pseudowords provide especially compelling evidence that stress assignment is *multiply determined*, with several lexical patterns serving as bases of stochastic generalizations extended to these forms. The interplay between these factors likely forms a crucial part of the puzzle of English stress, and future models of the system must take them into account (see chapter VIII). Rather than attempting to account for the entire stress system, the experiments described in this chapter control for most of the factors while focusing on one component: the link between stress and syllable structure. Presented with the same trisyllabic stimuli from Study 1, the participants were either asked to

choose their preferred stress pattern (Study 3), or else to stress the pseudoword themselves (Study 4). The metrical parse is inferred indirectly from stress: the crucial assumption is that, as long as the second vowel is realized as lax, antepenultimate stress implies that the cluster has been assigned to the onset of the final syllable, whereas stress on the second syllable is evidence of a closed penult. To illustrate, consider the minimal stress pair for the orthographic nonce form *vatablick*:

(5.3) a. (ˈvæ.tə.)<blɪk> b. və.(ˈtæb.)<lɪk>

Under the assumptions of metrical theory, the two stress patterns imply two different parses as shown above. Specifically, the initial [b] of the cluster is assigned to the onset of the final syllable in (a) and to the penult coda in (b). As example 5.3 makes clear, the two syllabifications have consequences for penult weight.

Following this logic, stress assignment can thus be used to indirectly probe the phonotactic generalizations involved in the syllabification of intervocalic clusters. As mentioned above, this technique has two advantages over the hyphenation task employed in Study 1. First, it is implicit rather than metalinguistic, making it more difficult for participants to arrive at rules through introspection. Second, whereas hyphenation can be argued to involve searching for word edges, Latin Stress unquestionably relies on word-internal syllable boundaries. Indeed, this pattern has been employed by historical linguists to reconstruct the syllable structure of Classical Latin: given well-founded assumptions about vowel length, the syllabification of intervocalic consonants and clusters was established on the basis of the regular rhythmic properties of Latin verse (see Cser, 2012 for discussion).

The two experimental approaches — hyphenation and stress assignment — thus complement each other. The former is explicit and direct, but potentially confounded by overt knowledge or by word-edge phenomena. The latter avoids these pitfalls but is indirect and thus requires a leap of faith (however reasonable) during interpretation. Confidence in the correct parsing model thus requires converging evidence from both methods: to the extent that they yield similar results, we can be sure that we are onto something.

5.2 Latin Stress in the Lexicon

Before introducing the experiments in this chapter, it is necessary to establish some baseline expectations about online productivity of Latin Stress. In other words, given nonsense trisyllables with either light or heavy penults, how inclined would an unbiased (i.e. probability-matching) learner be to stress each structure on the penultimate syllable? Establishing this baseline requires investigating the strength with which this pattern is instantiated in the lexicon. Are there many exceptions, or is Latin Stress relatively robust? Does it interact with other factors that might generalize to the pseudowords? In this section, I explore the strength of the lexical basis of the pattern by investigating the same lexicon used to calculate the word-edge statistics (see section 3.2 for a description).

5.2.1 Methodological Preliminaries

Quantifying the relationship between heavy penults and stress involved several non-trivial decisions. First, in order to determine weight, the lexicon had to be syllabified using some categorical algorithm. The nature of the parse is of course the empirical question driving this dissertation — how can the gradient parser hypothesis be reconciled with this categorical treatment? The answer is that, for the present purposes, the goal is to arrive at an approximate estimate of lexical support. Since we merely need to know the rough extent to which heavy penults attract stress over light penults, any reasonable definition of weight will suffice as long as it is consistently applied throughout the lexicon. In the descriptions that follow, the lexicon was thus syllabified according to the Maximal Onset Principle, ignoring morpheme boundaries (see also Moore-Cantwell, 2016 for the same treatment).

The second choice concerned the weight criteria: how is weight to be parameterized? Should weight assignment follow the classical binary criteria, or should it be scaled in proportion to rime (and perhaps onset) complexity? While the recent arguments for gradient weight are certainly compelling, this type of treatment is not necessary for the present purposes. Again, the idea is to get a rough estimate of Latin Stress. For the sake of simplicity, I therefore employed the traditional, binary weight distinction based on rime structure: rimes consisting of short (lax) vowels were coded as light, and all others counted as heavy.

That said, the weight criteria warrant a brief description, since they differed somewhat from recent work in this area. For instance, whereas both Carpenter (2016) and Moore-Cantwell (2016) classified all monophthong rimes as light, here I retained

the distinction between tense and lax vowels. Thus, while these researchers treated [Ci] and [Cu] syllables as light, here they counted as heavy (see also Ryan, 2011a for similar treatment). In addition, Moore-Cantwell (2016) coded unstressed syllables closed by sonorants as open (light), because these sonorants are often analyzed as syllabic (e.g. /Cəl/ → [Cl̩]). This coding scheme is inappropriate to the present purposes: since the goal is to predict stress from syllable structure, allowing the former to determine the latter is undesirably circular. That is, if some syllables count as light because they are unstressed, then one is sure to overestimate the relationship between stress and heavy syllables (because there will be fewer unstressed, heavy syllables). For this reason, all [V + sonorant] rimes were treated as closed, regardless of stress.

The third decision was more complex and concerned delimiting the set of English words which should constitute the lexical basis of the generalizations extended to the pseudoword stimuli. As noted above, the lexicon contains a number of stress patterns unrelated to syllable weight, with different lexical strata exhibiting different behavior. Which subset of the lexicon should be taken as the appropriate search space explored by the speakers?

The first relevant issue is word length: since the test probes are trisyllables, a prosodic template-based view might call for the search space to be constrained to trisyllabic words only. On the other hand, a more traditional approach based on a right-aligned stress window would extend the search space to also include words longer than three syllables. Both approaches have been employed in previous work. For example, while Domahs et al. (2014) restricted the lexical items to match the length of their pseudoword stimuli, Ernestus and Neijt (2008) explicitly investigated the interaction of

Latin Stress and word length, and Moore-Cantwell (2016) compared more and less-restricted search spaces.

A second issue concerns the role of morphological complexity. Many traditional stress rules explicitly reference morphology: affixes fall into different classes depending on whether they attract stress, cause retraction, or behave in a neutral way, and compounds follow their own set of rules. Should complex words be included in the search if the test probes contain no transparent morphology? Classical, compositional theories of morphology argue that complex words do not contribute to productivity because they are assembled from their subparts during online speech production rather than stored (e.g. Stockall & Marantz, 2006), and only stored items can contribute to the search space. Most lexicon-based studies of stress have tacitly adopted this approach. For example, in their comparisons of Dutch, English and German, both Ernestus & Neijt (2008) and Domahs et al. (2014) limited the investigation to monomorphemes found in CELEX. Similarly, in her analysis of light-syllable, English words, Moore-Cantwell (2016) filtered out some of the more productive affixes from the search space. On the opposite extreme, exemplar theories assume the storage of all auditory experiences (e.g. Bybee, 2001). On this approach, the search space should indeed be comprised of all word forms since everything has some bearing on the process of output selection. This approach is computationally implemented in network models where activation is some function of aggregate similarity.

The third issue is syntactic. It is well known that English stress is affected by lexical class (e.g. Kager, 1989). For instance, among disyllables, nouns are more likely than verbs to feature initial stress. As mentioned above, this particular fact is exploited by English speakers when assigning stress to novel disyllables (Guion et al., 2003). Does

word class play an appreciable role in trisyllables as well? While the design of Studies 3 and 4 encouraged the participants to interpret the stimuli as novel nouns (see below), one cannot be certain that the manipulation was successful. In other words, it is possible that the lexical search space extended beyond nouns to include other word classes.

Rather than taking theoretical positions on word length, morphological complexity and lexical class, I examined the interaction between each of these factors and Latin Stress. Specifically, I subdivided the lexicon into different parts and measured weight sensitivity within each lexical stratum. If Latin Stress is found to be stable across these different subsets of the lexicon, then the issue of correctly defining the search space becomes less crucial (since all spaces would support weight sensitivity in roughly equal measure). On the other hand, if Latin Stress is found to vary widely among the subsets, then the choice of search space might influence the interpretation of the participants' behavior.

5.2.2 Results

I begin with the stress window approach, which accepts any words longer than two syllables into the search space. Figure 5.1 plots the proportion penult stress in these words as a function of penult weight and morphology. For simplicity, the figure excludes the small number of words with primary stress on syllables other than the penult or the antepenult. The morphological coding was based on the CELEX database. The data are collapsed across word class.

Panel (a) represents the least restricted search space, where all long word forms are included. Thereafter, the search space grows increasingly constrained in terms of morphology. Panel (b) excludes inflected forms but allows all derived words, panel (c) further eliminates productive derivations but allows words with synchronically opaque morphology, and panel (d) includes only monomorphemic items. Compound words are shown separately in panel (e).

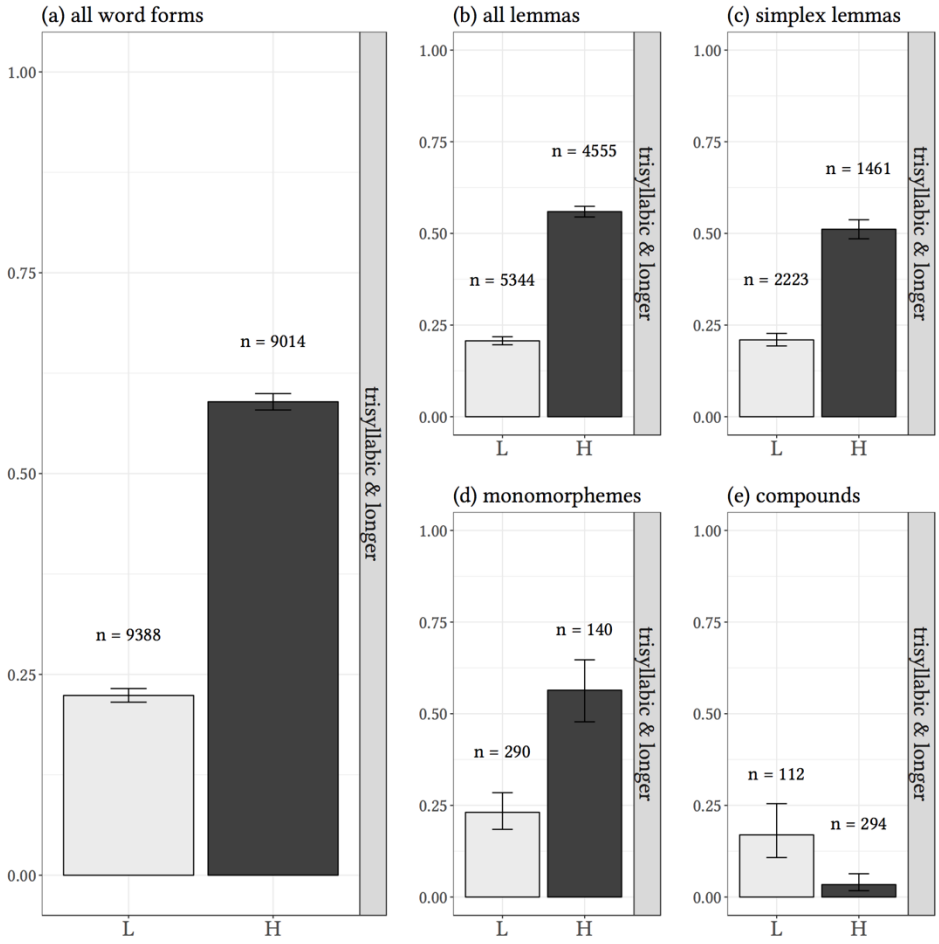

Figure 5.1. Latin Stress in English words of 3+ syllables, in different morphological subsets. The *y*-axis marks the proportion of penults receiving primary stress. The *x*-axis differentiates between L(ight) and H(eavy) penults. Error bars are 95% confidence intervals based on the proportion test. The number above each column represents the total cell size.

The figure shows that, as the search space becomes morphologically constrained from word forms to monomorphemes, Latin Stress remains remarkably robust. Across the lexicon subsets, heavy penults are more than twice as likely to be stressed as light penults: the former attract stress at rates between 51% and 59%, while the latter fall in the 21% to 25% range. Only compounds exhibit different behavior. First, they are much more likely overall to be stressed on the antepenult than the other items. Second, the weight generalization appears to be reversed, with light penults attracting more stress (however, this is a numerically small effect).

The prosodic frame approach is illustrated in Figure 5.2. Here, the search space is constrained to trisyllables only. To facilitate the comparison, the organization of the panels parallels that of Figure 5.1.

It is clear from the comparison that removing longer words had little effect on the lexical productivity of the pattern. Among the trisyllables, heavy penults attracted stress between 55% and 58% of the time, while light penults were stressed between 22% and 24% of the time.

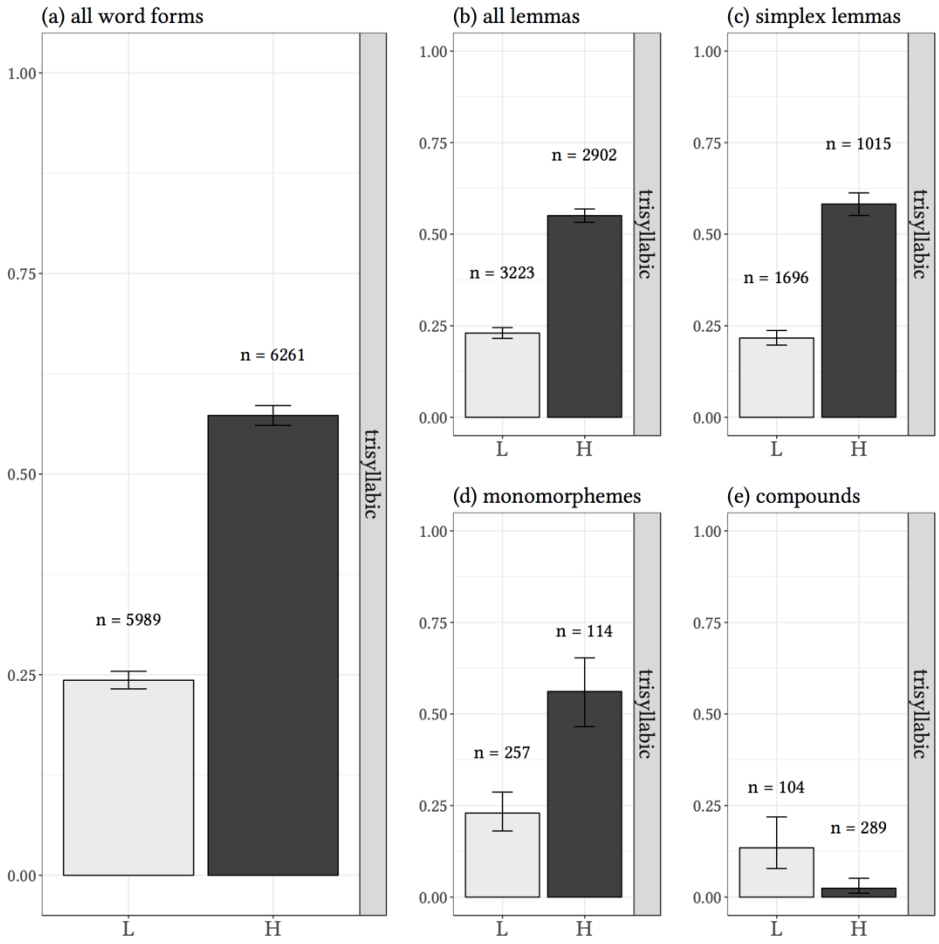

Figure 5.2. Latin Stress in English trisyllables, in different morphological subsets. The *y*-axis marks the proportion of penults receiving primary stress. The *x*-axis differentiates between L(ight) and H(eavy) penults. Error bars are 95% confidence intervals based on the proportion test. The number above each column represents the total cell size.

In order to explore the interaction between penult weight and part of speech, all of the lemmas of at least 3 syllables in length (see Figure 5.1b above) were grouped according to their CELEX-assigned word class. Figure 5.3 illustrates the productivity of Latin Stress within the four major classes of Nouns, Verb, Adjective and Adverb (closed classes were ignored).

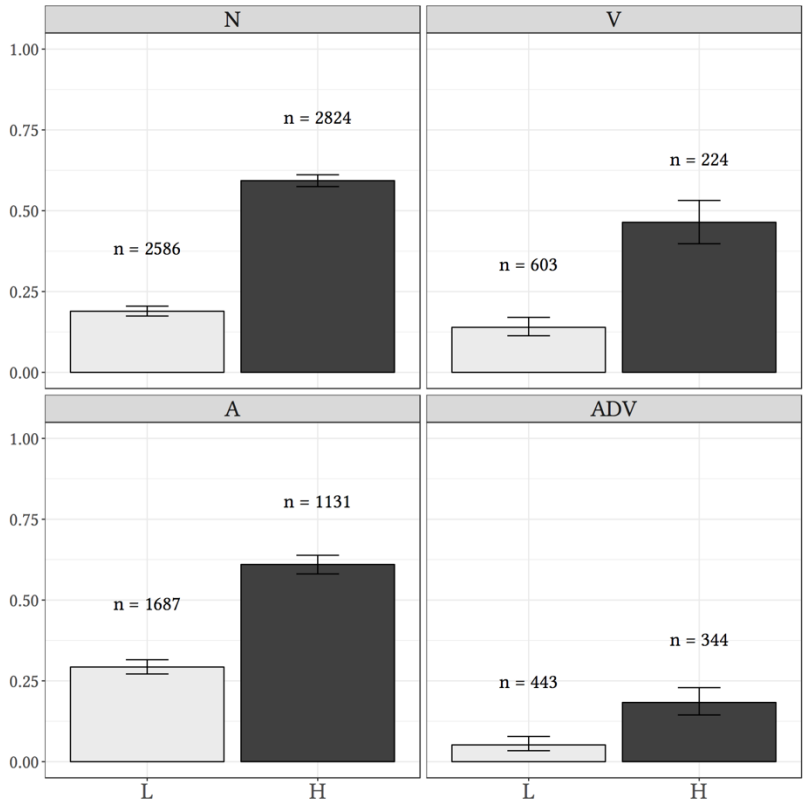

Figure 5.3. Latin Stress in English words of 3+ syllables, by major lexical class. The *y*-axis marks the proportion of penults receiving primary stress. The *x*-axis differentiates between L(ight) and H(eavy) penults. Error bars are 95% confidence intervals based on the proportion test. The number above each column represents the total cell size.

The comparison revealed that, at least in long words, adjectives are somewhat more likely than nouns to favor penult stress across the board, which in turn feature higher penult stress rates than do verbs. That said, the differences among these three classes are rather small. It is the adverbs that stand out, being by far more likely than the other classes to be stressed on the antepenult. Crucially however, Latin Stress appears to be stable across the parts of speech, with stress preferring heavy over light penults by a factor of at least 2.

5.2.3 Implications for Productivity

The results of the lexicon study point to a remarkable stability of Latin Stress across nearly all words capable of carrying the pattern. Under the assumption that learners probability match, this makes the expected baseline productivity relatively easy to establish: no matter which subset of the lexicon is used as the search space by the participants, we can expect LHL pseudowords to be about twice as likely as LLL pseudowords to attract penult stress. Roughly, an unbiased learner should stress H penults between 45% and 60% of the time, and L penults between 20% and 30% of the time. The only exceptions to this would be if participants somehow interpreted the pseudowords as adverbs (in which case the overall penult rates would be lower) or as compounds (in which case the L penults would attract more stress than H penults). Neither scenario is very likely: the stimuli do not resemble adverbs (for example, none end in *-ly*), and none can be exhaustively decomposed into recognizable free morphemes.[9]

An important implication of these findings is that the results of the stress experiments in this chapter are not expected to align perfectly with the hyphenation results from Study 1. Specifically, recall that in the hyphenation task, unattested word

[9] An alternative to the probability matching assumption is to follow Yang (2005; see also Legate & Yang, 2012) in modeling productivity as a sub-linear function of type frequency. According to Yang, a grammatical rule is only productive if the number of exceptions to it is sufficiently small. Specifically, the number of exceptions m must be less than $N/\ln(N)$, where N is the total number of words that meet the structural description of the rule in question. The threshold function is sub-linear in the sense that, as the relevant search space N grows in size, productivity permits a smaller proportion of exceptions. Under the assumption that both stressed L penults and unstressed H penult constitute exceptions to Latin Stress, Yang's model predicts that Latin Stress would not be productive in any of the sublexicons in Figures 5.1 - 5.3 (in all cases, the number of exceptions exceeded the productivity threshold). The only way to ensure productivity is to include disyllabic words (and perhaps even monosyllables) in the search space and allow these to support weight sensitivity (and thus reduce m). In this dissertation, I take the view that vacuous rule application does not constitute lexical support: in order to be relevant for the productivity of Latin Stress, words must have an antepenult.

onsets were split about 94% of the time, yielding heavy penults. It is highly unlikely that stress will mirror these rates. Only about half of the heavy penults in the lexicon receive stress; under the assumption that, like other lexicon-based generalizations, the weight-to-stress relationship is extended probabilistically, one can expect a lower penult stress ceiling on the pseudowords with medially-embedded unattested onsets. As a consequence, the stress judgment and assignment tasks are likely to show diminished sensitivity to phonotactics simply because the range of possible responses is compressed by an overall reluctance to stress the penult. That said, to the extent that stress is found to be modulated by the same phonotactic factors as hyphenation, it is reasonable to hypothesize that both processes are subserved by the same parsing mechanism. A gradient metrical parse is expected to produce variable syllable boundaries, the location of which should be reflected in the probabilistic treatment of penult stress in both perception and production.

5.3 Study 3: Stress Preferences

5.3.1 Overview

Study 3 examined the interaction of phonotactics and stress during the processing of spoken pseudowords using a well-formedness judgment task. As reviewed in section 2.3, well-formedness judgments constitute one of the main sources of data in phonotactic models (e.g. Albright, 2009; Daland et al., 2011; Hayes & Wilson, 2008). These tasks generally come in three variants. They can be categorical, forcing the participant to make a binary choice when evaluating a nonword (e.g. *'Is this form*

possible as a new English word?'; Scholes, 1966). They can also employ a Likert-style ratings scale, which can capture gradient preferences within each subject (Coetzee, 2009; Frisch & Zawaydeh, 2001). Finally, tasks can force participants to directly compare two or more alternatives (Berent & Shimron, 1997); the task employed here is of this latter variety.

Participants in Study 3 performed a *two-alternative, forced-choice (2AFC)* task. They were aurally presented with minimal stress pairs (e.g. *vátablick ~ vatáblick*) and asked to choose the more natural-sounding pronunciation. To minimize perception errors, orthographic support was also provided (see also Hayes & White, 2013). The hypotheses were once again based on those outlined in section 2.4. To the extent that a medial cluster was interpreted as a bad complex onset, the item containing it should be favored when stressed on the penult, since this stress pattern would reflect a C.C parse.

The 2AFC task was similar to that employed in Guion et al. (2003) and Daland et al. (2011). The decision to use it in lieu of a Likert task was motivated by two factors. First, I reasoned that presenting the stimuli individually (as in the Likert task) would cause the effects of cluster phonotactics to be masked by the shape of the context frames, since the latter constituted about 75% of the phonological makeup of each item (including the perceptually salient beginning and end). Second, Daland et al. (2011) compared the two methods and found the 2AFC preference task to be more sensitive to gradient phonotactics of word onsets because the Likert scale was subject to floor effects, where all unattested clusters were treated as equally deviant (see also Coetzee, 2009 for similar results).

5.3.2 Method

5.3.2.1 Participants

One hundred and thirty participants took part in the experiment as part of a larger study on the learning of stress patterns. Of these, 40 were excluded because they either (a) had significant exposure to a language other than English and self-reported as fluent speakers of that language, or (b) were found to be uncooperative during the subsequent learning task and thus judged to be sources of random noise. Data from the remaining 90 individuals were retained for analysis.

5.3.2.2 Materials

The stimuli used in Study 3 consisted of half of the pseudowords used in the hyphenation experiment. Specifically, while all of the inserts from Study 1 were represented, only 22 of the 44 CVCV__VC frames were retained.[10] As a result, Study 3 featured 85 unique pseudowords; these are listed in Appendix A.

Test trials involved both visual and auditory presentation of the items. On the visual side, the pseudowords were represented in lower-case, Courier font. For each spelling, two auditory versions were prepared: one with stress on the penult and the other with stress on the antepenult. These 'minimal stress pairs' were recorded by a an adult, male native American English speaker with training in phonetics. The speaker was instructed to pronounce each pseudoword in the most natural, native-like way,

[10] Withholding half the items was necessary in order to measure generalization to unseen items in a learning study which took place immediately following this task. The results of the learning study are reported elsewhere (Olejarczuk, 2014; Olejarczuk & Kapatsinski, in revision).

imagining that it was a novel word entering the English language. He was further instructed to maintain a constant mapping between orthography and pronunciation. In stressed syllables, vowels remained lax, so that *e* was always pronounced as [ɛ], *i* as [ɪ], and *a* as [æ]. This was done in order to eliminate the influence of phonological length on syllable weight, which would have confounded the interpretation that stress indicated closed syllables. Vowels in unstressed syllables were reduced to either [ə] or [ɪ] as appropriate. Prior to recording, the speaker practiced the item list several times in order to get acquainted with the spelling.

The speaker provided several productions of each minimal stress pair. These were recorded in a quiet, sound-treated room using a USB condenser microphone connected directly to a laptop computer. Each set of productions was saved to a .wav file at 16-bit, 44.1kHz resolution. From each series, the production judged to be most natural and representative of the desired stress pattern was excised and saved to a separate file. The resultant files were then batch normalized to the same peak amplitude in Praat (Boersma, 2001). Peak rather than average amplitude was used in order to prevent amplitude compression.

In order to ensure that each production contained phonetic cues to the intended stress patterns, the recordings were segmented and measured in Praat. Segmentation followed criteria standard in the field (e.g. Klatt, 1976). Specifically, vowel offsets were identified by abrupt lowering of energy in the upper formants, nasals by the presence of anti-resonances, and liquids by upper formant movements and changes in amplitude relative to neighboring vowels. Stress was verified by reference to two acoustic correlates known to correspond to perceptual cues: duration and intensity (see Cutler, 2005 for a review of perceptual cues to stress). To illustrate, Figures 5.4 and 5.5 show

the segmentations, spectrograms and intensity contours for the minimal stress pair of *tabasmub* as pronounced by the speaker. Note that, in the antepenult-stressed variant (Figure 5.4), the antepenultimate vowel is longer and higher in intensity than the penultimate vowel. This relationship is reversed in the penult-stressed version (Figure 5.5).

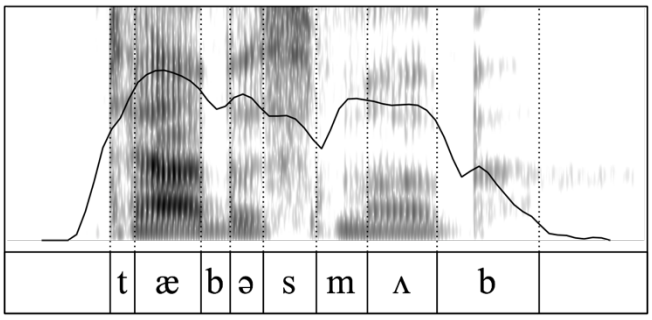

Figure 5.4. Spectrogram and segmentation of the pseudoword *tabasmub* with stress on the antepenult. Intensity curve is superimposed on the spectrogram. Time is on the x-axis. Frequency (spectrogram) and intensity (curve) are on the y-axis.

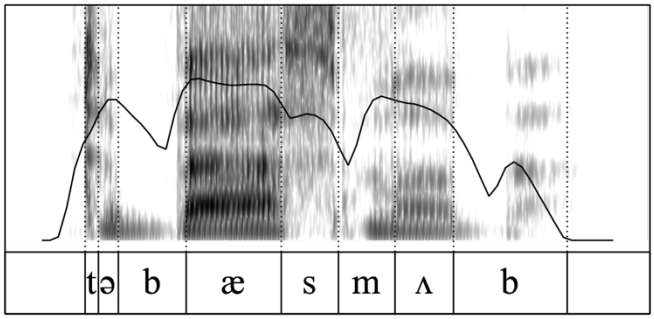

Figure 5.5. Spectrogram and segmentation of the pseudoword *tabasmub* with stress on the penult. Intensity curve is superimposed on the spectrogram. Time is on the x-axis. Frequency (spectrogram) and intensity (curve) are on the y-axis.

To verify the intended stress patterns across all items, the antepenultimate and penultimate vowels were compared on two acoustic measures: V2:V1 duration ratio and V2-V1 intensity difference. The average values of these measures are plotted in Figure 5.6.

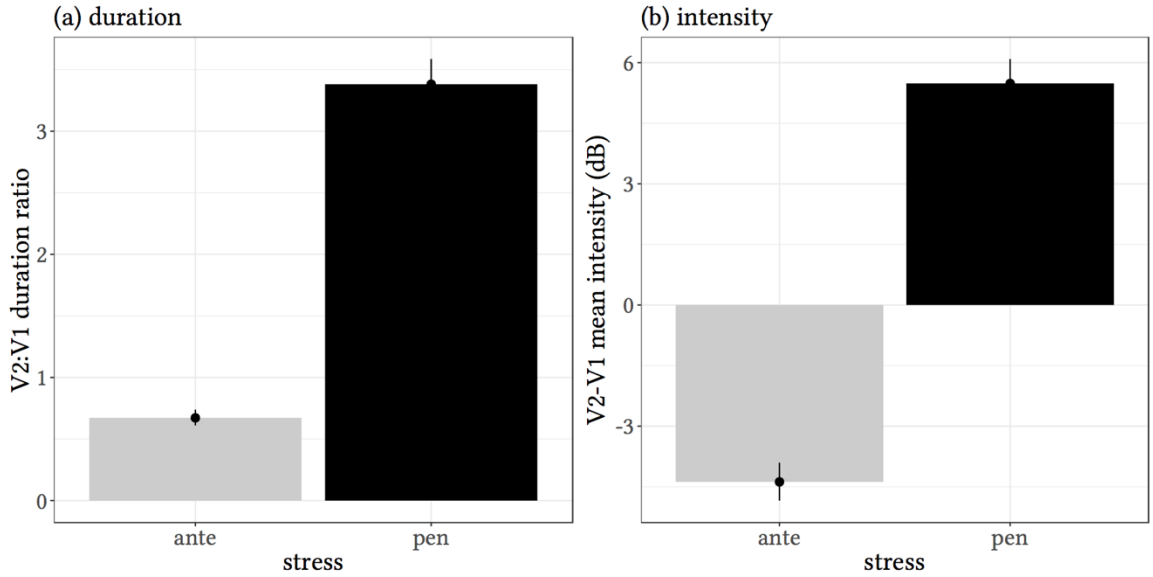

Figure 5.6. Mean acoustic correlates of stress in the auditory stimuli. Error bars are confidence intervals obtained via nonparametric bootstrap.

Panel (a) shows the duration ratios. For words stressed on the antepenult, the mean V2:V1 duration ratio was less than .67, indicating longer antepenultimate vowels. Conversely, stressed penults were longer than unstressed antepenults by a factor of 3.38. A mixed-effects linear regression predicting log-transformed duration ratios from stress and containing random by-word intercepts confirmed that the stress effect was significant (β = 1.62, *S.E.* = .05, p < .001).

Panel (b) displays the intensity difference between the two vowels in question. For items stressed on the antepenult, the stressed vowel was higher in intensity than the unstressed vowel by approximately 4.38dB. For items stressed on the penult, the difference was about 5.48dB. A mixed-effects linear regression predicting the intensity differences from stress and containing random by-word intercepts found a significant stress effect (β = 9.86, *S.E.* = .35, p < .001).

Acoustic analysis thus confirmed the reliable presence of duration and intensity differences, two important cues to the perception of English stress. Furthermore, both of the measured effects were well beyond the just noticeable difference (JND) thresholds established in the psychoacoustic literature (see Moore, 2013).

5.3.2.3 Procedure

The experiment was administered individually via the E-Prime 2.0 software environment (Schneider, Eschman & Zuccolotto, 2002). Participants were seated alone in a small, quiet room in front of a monitor screen. Each trial began with the orthographic presentation of a pseudoword in black font, centered against a white background. After an interval of 500ms, the minimal stress pair was presented over headphones at a comfortable listening level. Pair members were separated by 500ms, and within-pair stress order (penult/antepenult) was counterbalanced across participants. Each trial was presented only once, in random order.

Participants were instructed to listen to each pair, consider the written form, and decide which pronunciation would be a better fit if the word were to be introduced into the English language as a new noun. The participants entered their choice by pressing a button on a serial response box. Trials advanced 500ms after a response was recorded.

The preference task lasted approximately 15 minutes. Immediately following its completion, the participants took part in miniature artificial language learning experiments reported elsewhere (Olejarczuk, 2014; Olejarczuk & Kapatsinski, in revision).

5.3.3 Results

5.3.3.1 Nuisance Covariates

Before evaluating the phonotactic models, I begin by examining the effects of the two nuisance covariates on the responses (see section 3.4.2 for the description). Figure 5.7 plots each covariate against the proportion of penult- over antepenult-stressed pseudowords chosen by the participants. With the data aggregated by word, there are 85 data points in each plot.

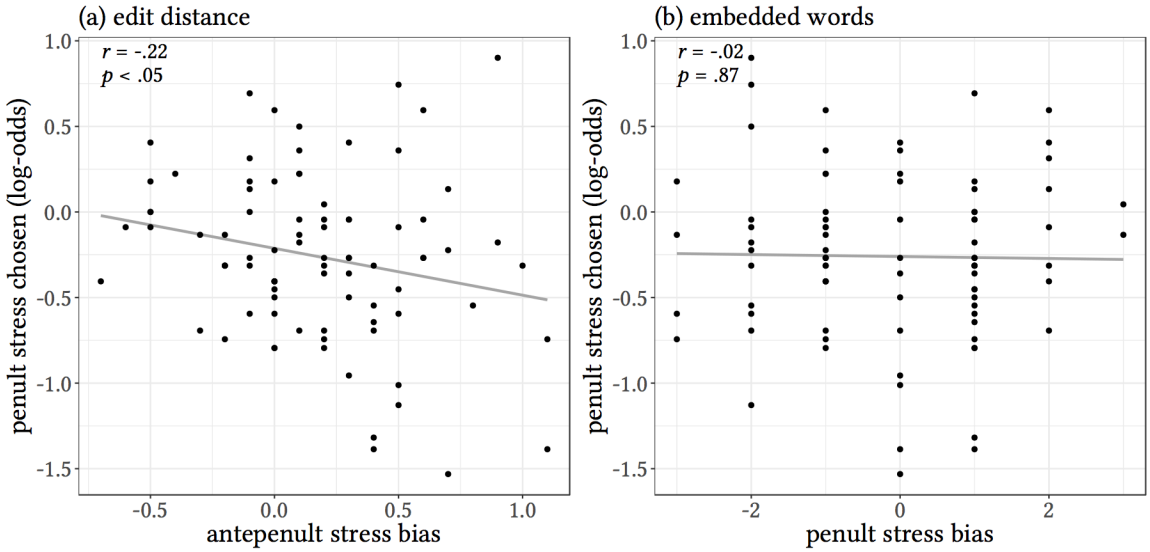

Figure 5.7. Effects of nuisance covariates on stress preferences (all test items).

Panel (a) shows the effect of the edit distance-based covariate. For each pseudoword, the x-axis represents the difference in mean edit distance to the nearest 10 antepenult- versus penult-stressed lexical items. There is a significant trend in the aggregate data wherein items closer to penult-stressed neighbors tend to be preferred

when stressed on the penult. The trend was confirmed in a univariate, mixed-effects model fit to the raw responses, which revealed a significant effect of edit distance (β = -.50 *S.E.* = .21, $p < .05$). With each edit closer to antepenult-stressed neighbors, the odds of selecting the penult-stressed variant decreased by a factor of .61.

Panel (b) shows the effect of the embedded words on aggregate responses. Here, the positive values on the x-axis indicate that the short words embedded in the test items favored penult stress. As seen in the figure, there is virtually no correlation with the preferences indicated by the participants in the study. The univariate, mixed-effects logistic regression predicting the raw response data did not return a significant effect of edit distance (β = -.003 *S.E.* = .05, p = .94).

5.3.3.2 Coarse-Grained Phonotactics

This section examines the effects of insert status on the stress preferences exhibited by the participants in Study 3. Figure 5.8 plots the proportion of penult-stressed versions chosen over their antepenult-stressed counterparts. Approximately 38% of singleton-bearing items were preferred with penult stress. For words with attested CC inserts, this number was 43%, while for items with unattested CC inserts it was 48%.

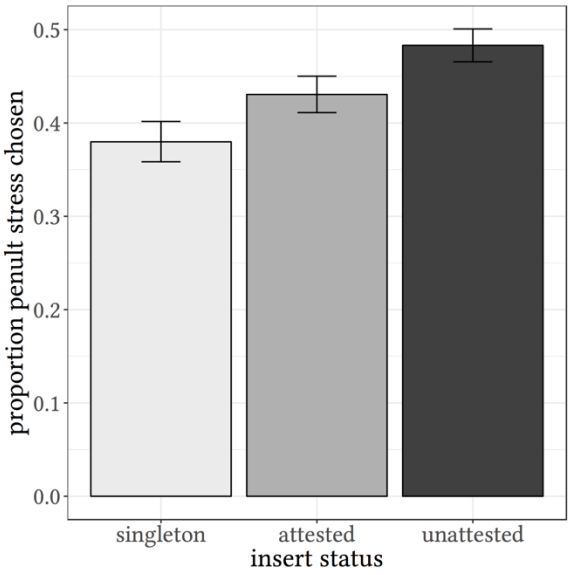

Figure 5.8. Penult preferences by insert status. Error bars are 95% confidence intervals based on the proportion test.

To test for the effect of insert status in the presence of the nuisance covariates, a mixed-effects logistic regression was fit to the data. Both edit bias and the embedded word bias were centered and scaled. Because the maximal model failed to converge, the number of parameters was reduced by removing the random correlation estimates (see Bates et al., 2015). This reduced model converged successfully, and it significantly outperformed an intercept-only version according to the likelihood ratio test ($\chi^2(4) = 14.43$, $p < .01$). The output is shown in Table 5.2.

With the reference level set to singletons at the mean covariate values, only the effect of unattested CC inserts emerged as statistically significant. Specifically, the odds of preferring the penult-stressed versions of these items were higher than those of singletons by a factor of 1.61. In contrast, likelihood ratio tests revealed that, with coarse phonotactics in the model, neither covariate significantly contributed to fit (edit distance: $\chi^2(1) = 1.30$, $p = .26$; embedded words: $\chi^2(1) = 1.66$, $p = .20$).

Table 5.2. Categorical model output (stress preference task).

	Estimate (Std. Error)
Intercept (Status = singleton)	-0.548 (0.132)***
Status = attested	0.186 (0.140)
Status = unattested	0.475 (0.131)***
Edit distance bias	-0.058 (0.051)
Embedded word bias	-0.066 (0.052)
Observations	7,650
Log Likelihood	-5,055.275
Bayesian Inf. Crit.	10,316.230
Note:	*p<0.05; **p<0.01; ***p<0.001

To test the difference between attested and unattested CC inserts, a second model was fit to a subset of the data. The model revealed a significant effect of insert type (β = .29, *S.E.* = .12, p < .05). Relative to attested items, the odds of preferring penult-stressed versions of unattested items increased modestly by a factor of 1.43.

To sum up, singleton and attested items behaved similarly, whereas unattested items were more likely than both to elicit preferences for their penult-stressed versions. I now turn to examining the effects of fine-grained phonotactics.

5.3.3.3 Fine-Grained Phonotactics

This section presents the gradient analysis of the 2AFC results. As in the preceding chapter, I begin with a look at the relationship between individual predictors and aggregate responses. The correlation between word onset frequency and stress preferences in words with embedded singletons and attested clusters in shown in Figure 5.9. As in Study 1, there were 12 unique singleton and 28 unique attested inserts for a total of 40 data points.

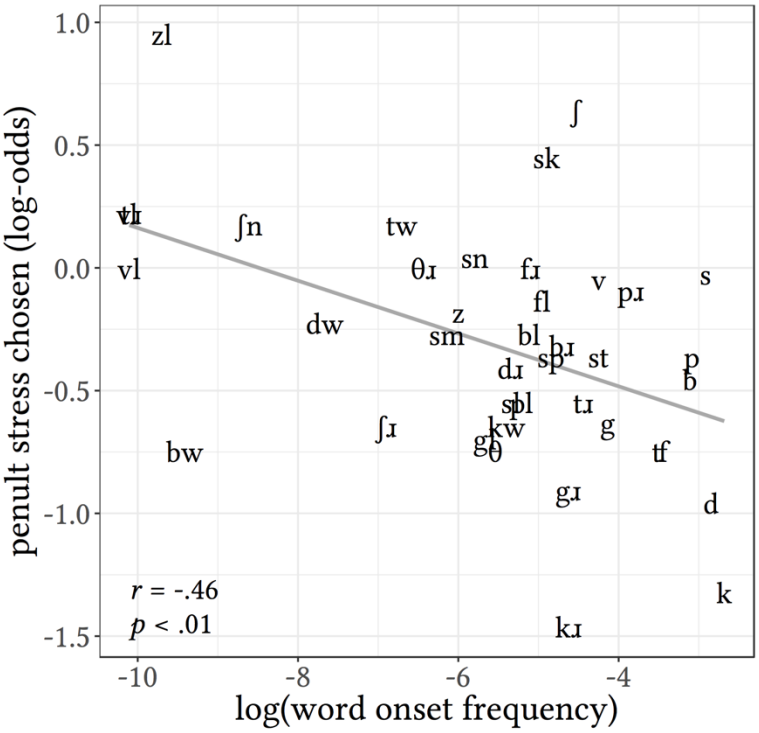

Figure 5.9. Log-odds of penult-stressed variants chosen, by word-initial frequency of each embedded insert (singletons, attested CC onsets).

The negative correlation is consistent with the gradient hypothesis, with penult-stressed variants preferred more often in items with rare word onsets embedded between V2 and V3. On its own, word onset frequency captured about 21% of the variance in the aggregated responses.

To test the onset frequency effect on the raw responses, a maximal, mixed-effects model was fit to the items with singletons and attested clusters. Word onset frequency was found to significantly predict stress preferences (β = -.12 *S.E.* = .05, p < .05). With each unit increase in log frequency, the odds of preferring the penult-stressed variant dropped by a factor of .89. The effect persisted after the exclusion of marginal onset clusters from the data (β = -.08 *S.E.* = .02, p < .01).

The second gradient predictor, word offset frequency of the insert C1 (17 data points), is plotted against aggregate responses in Figure 5.10.

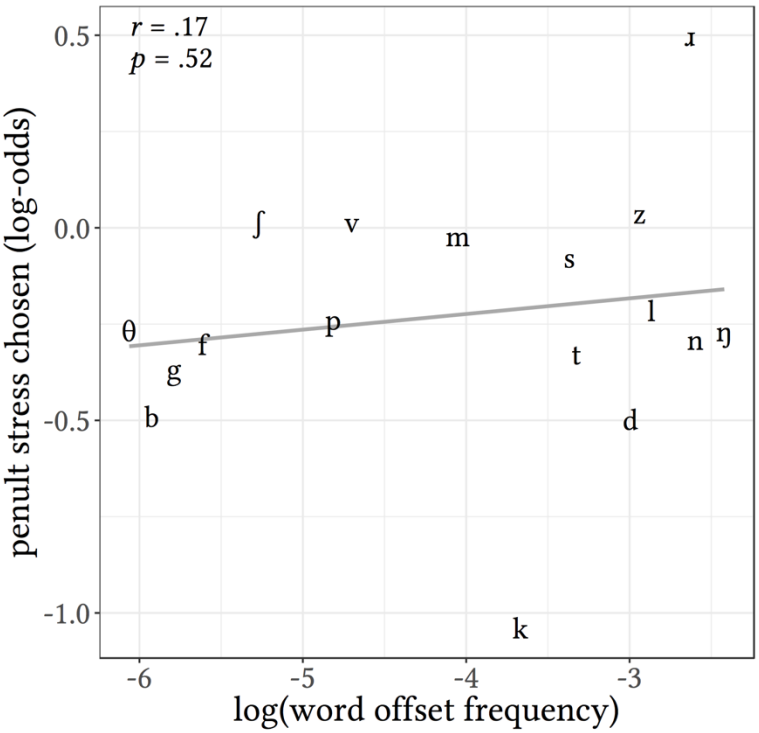

Figure 5.10. Log-odds of penult-stressed variants chosen by word-final frequency of the C1 of each embedded insert.

The correlation was weak and failed to reach significance, with offset frequency accounting for about 3% of the variance in the averaged responses. However, a mixed-effects logistic regression fit to the raw data revealed that offset frequency did significantly predict preferences (β = .14, *S.E.* = .07, p < .05). The effect was modest, with each unit of offset frequency increasing the odds of choosing the penult-stressed variant by a factor of 1.15. Furthermore, the effect disappeared after the removal of /ŋ/ from the data (β = .08, *S.E.* = .05, p = .08).

The final gradient predictor under investigation was sonority. The correlation between sonority slope and the averaged responses to items with unattested clusters is plotted in Figure 5.11.

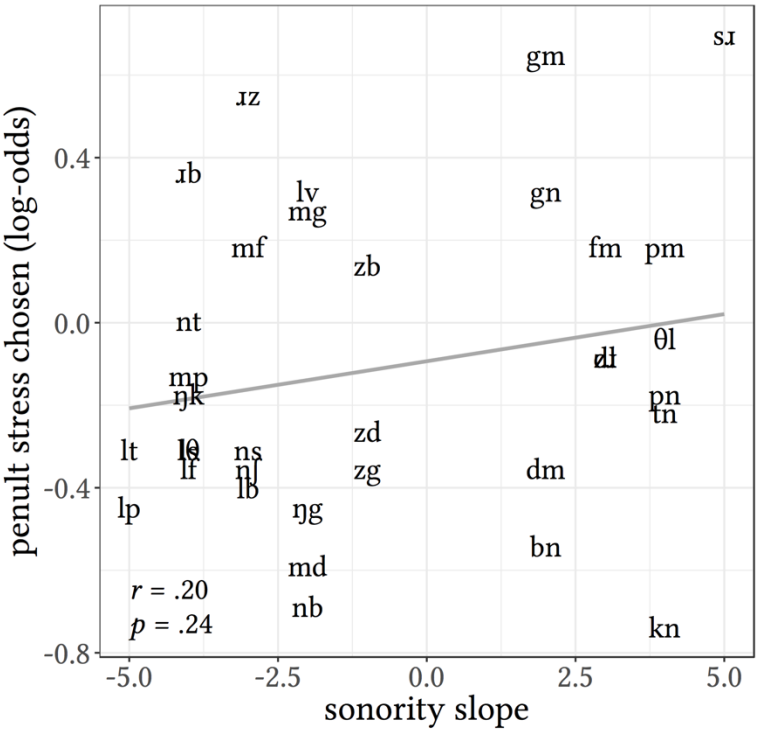

Figure 5.11. Log-odds of penult-stressed variants chosen by sonority slope of each embedded insert (unattested clusters only).

Unlike in Study 1 and Study 2, the correlation was not significant. Furthermore, the trend was *positive*, with penult-stressed variants more likely to be preferred in items with rising-sonority inserts. A mixed-effects model fit to the raw data likewise found no significant effect of sonority on the preferences (β = .02, *S.E.* = .02, p = .20).

In order to examine how each gradient measure predicted stress preferences in the presence of the others, a multiple, mixed-effects logistic regression was fit to the entire data set. All predictors were centered and scaled, and sonority slope was residualized against the onset and offset frequency measures to reduce collinearity. As in the categorical model (Table 5.2), the random correlation parameters were omitted in order to facilitate convergence. The model significantly outperformed a null version according to the likelihood ratio test ($\chi^2(5) = 10.61$, $p < .01$). Table 5.3 lists the model output. Figure 5.12 plots the estimates and marginal effects.

Table 5.3. Gradient model output (stress preference task).

	Estimate (Std. Error)
Intercept	-0.305 (0.067)***
Word Onset Frequency	-0.211 (0.059)***
Word Offset Frequency	-0.026 (0.043)
Sonority Slope	0.034 (0.060)
Edit Distance Bias	-0.140 (0.072)
Embedded Words Bias	-0.102 (0.057)
Observations	7,650
Log Likelihood	-5,054.529
Bayesian Inf. Crit.	10,270.020
Note:	*p<0.05; **p<0.01; ***p<0.001

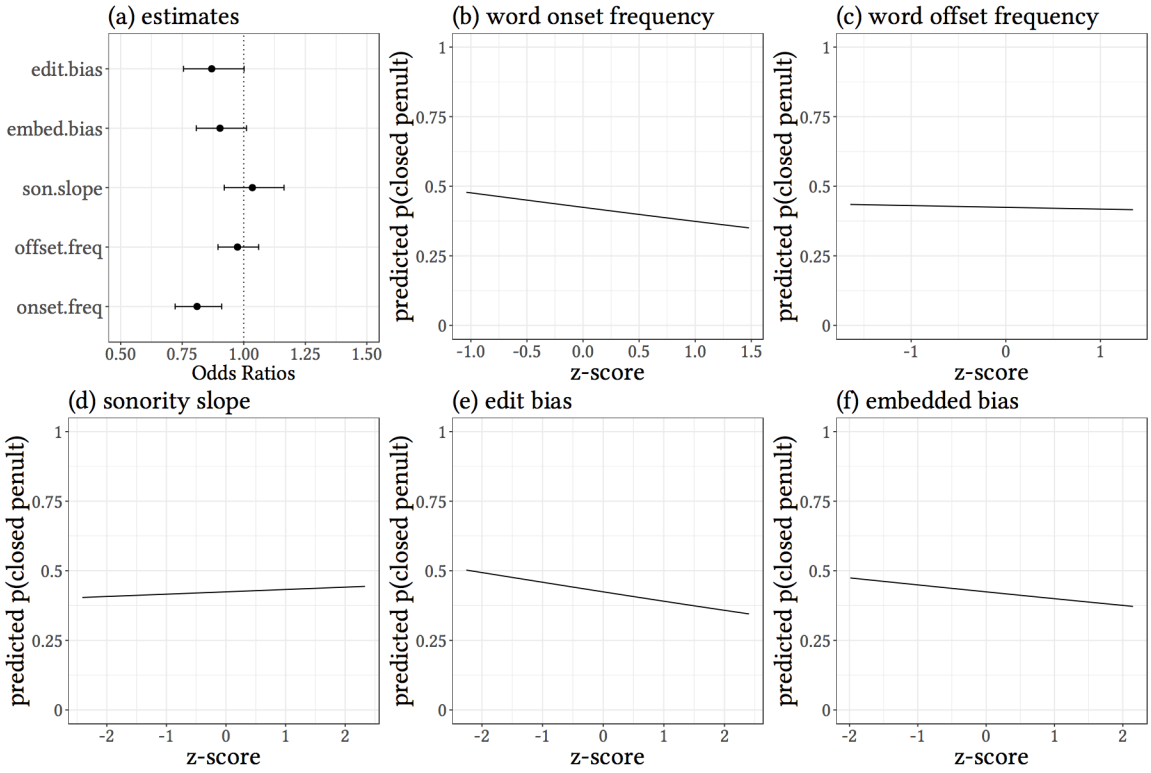

Figure 5.12. Gradient model estimates (panel [a]; dotted vertical line represents the null hypothesis) and marginal effects (panels [b]-[f]).

The output revealed that, of all five predictors, only word onset frequency had a significant effect on the preferences. With each *z*-score increase in onset frequency, the odds of choosing the penult-stressed variant decreased by a factor of .81. There were

also numerical trends in the two nuisance variables. Curiously, the trend for embedded bias was in the opposite direction than expected: pseudowords with embedded items cuing penult stress were somewhat less likely to be preferred with penult stress. However, neither trend was statistically significant. I now turn to the question of whether the gradient model fit the data better than the categorical model.

5.3.3.4 Model Comparison

This section compares the fit of the categorical and gradient phonotactic models to the stress preference data. Continuing the strategy in the previous chapter, the comparison consists of predictive accuracy on aggregate responses and posterior probabilities derived via the *BIC* approximation.

To begin, Figure 5.13 plots the correlation between predicted and observed values. There are 75 data points in the plots, each of which represents the average value for a unique insert. The predicted values were conditioned on fixed effects only.

Unlike in Study 1 and Study 2, the predictions of the categorical model are somewhat distributed rather than confined to three discrete values (cf. Figures 4.6a and 4.12a). This is of course because the categorical model in Study 3 contained two nuisance predictors which were in fact continuous (only the phonotactic predictor was categorical). Visual examination of the two scatter plots suggests marginally better performance for the gradient model, where the predictions are slightly more distributed. The fit statistics confirm this pattern: relative to the categorical model, the gradient model had a marginally lower mean squared deviation and accounted for about 5% more variance in the aggregated data.

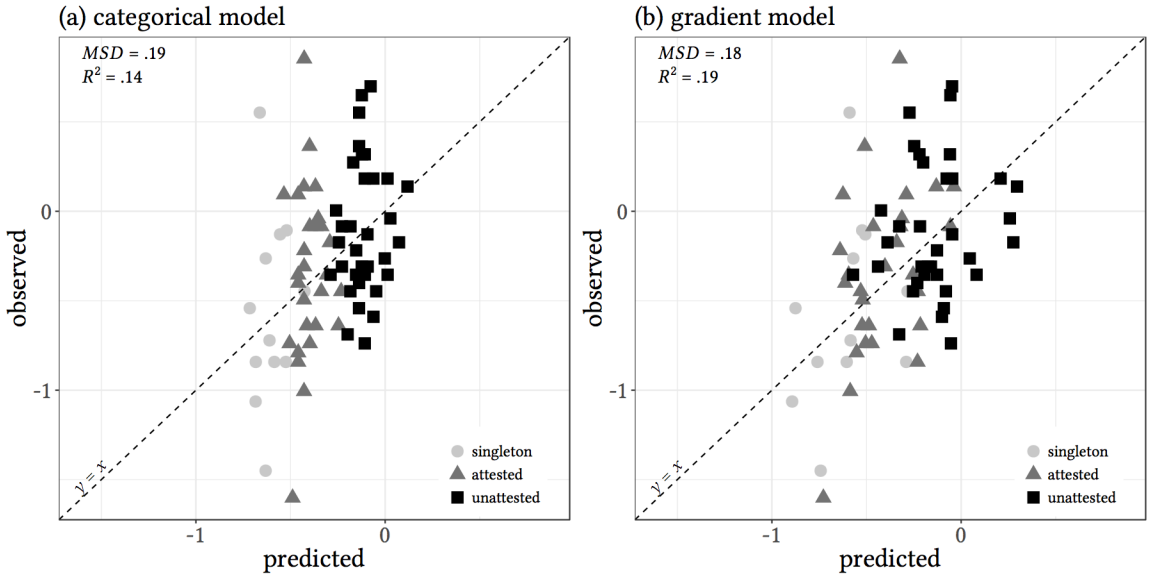

Figure 5.13. Comparison of model predictions (stress preference data). Values are in log-odds.

This performance advantage on the aggregated predictions was marginal. However, a comparison of the posterior probabilities tells a different story. On unaggregated data, the *BIC* scores for the two models were 10,316 (categorical) and 10,270 (gradient). This difference of 46 points corresponded to a Bayes Factor of nearly 1.1×10^{10} for the gradient model, which in turn translated to a posterior probability of nearly 1. Provided with the learning data and a choice between both models, a rational, unbiased learner would almost always choose the gradient model.

5.3.4 Discussion

To summarize, the results of Study 3 are consistent with the gradient parser hypothesis, but the effects were somewhat weaker than in the hyphenation tasks used in Studies 1 and 2. Specifically, of the three gradient predictors, only word onset

frequency consistently and significantly contributed to stress preferences, with rare onsets more likely than frequent onsets to elicit preferences for penult-stressed variants. This effect was significant both within the legal onsets and in the multiple regression model fit to the full data set. Neither sonority slope nor offset frequency emerged as significant predictors in the full model. Nevertheless, the gradient parsing model held a slight R^2 advantage over the categorical alternative and emerged as the clear winner in the comparison of *BIC* scores.

There are a number of possible reasons why the fit of the gradient parsing model was weaker in this task relative to word division. One that can be ruled out immediately is perceptual noise — the idea that the participants had difficulty perceiving the difference between the penult- and antepenult-stressed productions they were asked to compare. This was not the case because immediately following the judgment task, the subjects participated in a learning study (Olejarczuk & Kapatsinski, in revision), wherein training consisted of repeating the same items. The training productions were recorded and checked, revealing that the participants were nearly perfect in reproducing the stress patterns. The source of the difference likely cannot be attributed to misperception.

That said, a number of other factors could have been implicated. First, it is possible that stress placement is cued by more phonological factors than is hyphenation (see section 5.1). Since there is more competition for stress, each predictor may have accounted for a lower unique share of the total variance. Second, although the 2AFC task has been shown to be more sensitive to gradient phonotactics than the Likert scale (Coetzee, 2009; Daland et al., 2011), a binary choice task may result in more guessing than an open-ended task like hyphenation. Comparing the coarse-grained results in

Figures 4.1 and 5.8 reveals that, whereas hyphenation exhibited a wide range of responses across insert type, 2AFC results hovered closer to chance. Third, closed-set tasks like the 2AFC have been argued to reduce listener sensitivity to phonetic variability and lexical neighborhood effects during spoken word recognition (Sommers, Kirk, & Pisoni, 1997). It is thus possible that providing the illicit forms essentially primed them, boosting their acceptability (see also Harmon & Kapatsinski, 2017; Luce & Pisoni, 1998; Luka & Barsalou, 2005; Snyder, 2000).

In addition, some of the difference may have been due to conflicting parses between stress cues and phonetic juncture cues. For instance, illegal inserts that began with liquids may have featured the 'dark', velarized variant of /l/, regardless of stress pattern. This phonetic realization may have cued coda assignment, which came into conflict with the parse assigned by antepenultimate stress. To check for this possibility, I compared stress preferences between items with liquid-initial and nasal-initial clusters. No significant difference emerged (β = -.06 S.E. = .18 z = -.32, p = .75). A more likely possibility is that the relative durations of the two members of the CC inserts (as pronounced by the trained speaker) may have served as a perceptual cues syllable boundaries. Redford & Randall (2005) investigated the interaction of various phonetic juncture cues and phonological knowledge in the hyphenation of disyllabic nonce words. They found that, for items with embedded legal CC onsets and second syllable stress, longer C2 durations yielded fewer .CC syllabifications of the clusters. To check for this possibility, I calculated the C1:C2 ratios for both antepenult- and penult-stress variants of the Study 3 stimuli, subtracted the former from the latter, and predicted stress preferences from the resulting ratio differences. Figure 5.14 plots the results

separately for items with initially attested and unattested medial clusters. Because singleton items by definition lack C2, they are excluded from the plot.

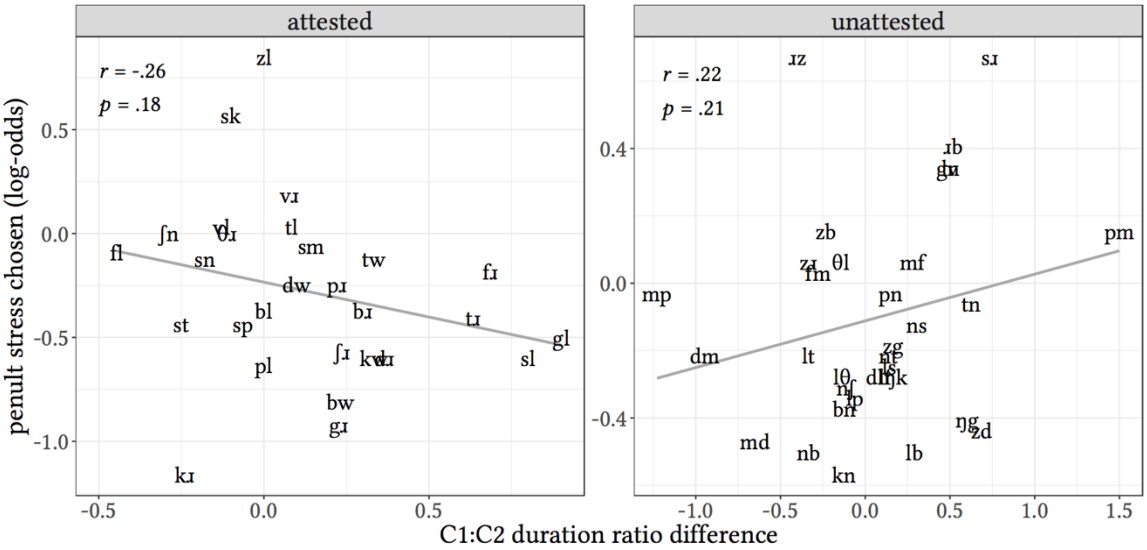

Figure 5.14. Log-odds of penult-stressed variants chosen by difference in C1:C2 duration between antepenult- and penult-stressed variants. Larger values on the x-axis indicate larger ratios for items with penult stress. Pseudowords with embedded singletons are excluded.

In the figure, positive values along the x-axis indicate larger C1:C2 ratios in penult- relative to antepenult-stressed items. For example, in the rightmost cluster in the left panel ([gl]), the initial [g] was longer relative to the following [l] when the item was stressed on the penult than when it was stressed on the antepenult. The relationship holds to a much lesser extent for [fl], the leftmost cluster in the panel. Although the correlations shown in the panel failed to reach significance (see inset *r* and *p* values), there appears to be a numerical interaction between attested and unattested items. Among the former, relatively long C1 in penult-stressed (relative to antepenult-stressed) variants leads to numerically greater preferences for antepenult stress (i.e. .CC parse). The direction of this relationship is consistent with the

hyphenation results reported in Redford & Randall (2005). Among the latter, the relationship is numerically reversed.

To recapitulate, although the stress preferences appeared to be guided by a gradient metrical parsing model, it is possible that task effects and phonetic juncture cues captured in C1:C2 duration ratios interacted with the phonotactics, or at least contributed some noise to the results. In addition, the syllable's role in the processing of spoken words seems to be controversial for stress-timed languages like English (recall section 2.1.3). Taken together, these potential complications suggest that a 2AFC perceptual task may not be optimally sensitive in uncovering the relationship between stress and syllabification. In the remainder of this chapter, I present the results of an online production task that overcomes many of these issues.

5.4 Study 4: Stress Assignment

5.4.1 Overview

Study 4 was an online production task where participants were presented with orthographic prompts of the same pseudowords used in Study 1 and simply asked to produce each form as naturally as possible. In producing each form, the participants assigned stress to one of the three syllables. The location of stress was coded, spot-checked against a second rater naive to the purpose of the study, and verified with acoustic measurements. As in Study 3, stress placement was treated as an indirect window on the metrical parse. To the extent that a medial cluster was interpreted as a

bad complex onset, the item containing it should be more likely to receive penultimate stress (see section 5.1).

Online production tasks have been used to probe various aspects of metrical knowledge in a number of studies dating back to at least the 1970s. Baker & Smith (1976) employed orthographic nonsense prompts to study the effectiveness of *Sound Patterns of English (SPE)* rules (Chomsky & Halle, 1968), analogy and word class in predicting stress assignment. Walch (1972) likewise investigated the role of stress rules using written nonwords. More recently, the method has been adopted by a number of studies examining factors beyond the scope of traditional metrical theory. Both Kelly (2004) and Ryan (2011a) used orthographic prompts to explore gradient weight (see section 5.4.4.1.1). Ernestus & Neijt (2008) likewise employed written stimuli (transcribed in IPA) to investigate the effect of word length on stress placement in German, Dutch and English. Shelton, Gerfen & Gutiérez Palma (2012) used a naming task to examine stress-attracting properties of falling and rising diphthongs in Spanish. Domahs et al. (2014) investigated differences in the sensitivity to syllable structure in German, Dutch and English. Hirsch (2014) employed orthographic prompts to argue that the weight-bearing unit is the V-to-V interval rather than the syllable (see section 5.4.4.1.2 for details).

Taken together, these studies establish the link between the metrical grammar and productivity. In Study 4, this link is exploited to examine the nature of the phonotactic generalizations relevant to weight-sensitive stress assignment.

5.4.2 Method

5.4.2.1 Participants

Thirty-six undergraduates were recruited from the same pool as in Exp. 1. All participants self-reported to be monolingual, native speakers of American English with corrected-to-normal vision and no hearing impairments. Data from six participants were excluded: two due to self-reported dyslexia, and an additional four due to failure to meet the accuracy criterion of 60% useable productions (see below for fluency criteria). The data from the remaining 30 participants were analyzed.

5.4.2.2 Materials

The target items consisted of the same 170 nonce words used in Study 1. In addition to these, 506 nonword fillers were randomly generated with the Wuggy software program, which is designed to produce phonotactically legal pseudowords (Keuleers & Brysbaert, 2010). The fillers were 1-5 syllables in length and were created by concatenating legal English syllables of various structures. The rationale for using nonwords rather than real words for fillers was that the former have been argued to encourage grammatical processing (e.g. by referencing phonotactic probabilities) while the latter may be processed by reference to lexical neighborhoods (Shademan, 2006; Vitevitch & Luce, 1998).

5.4.2.3 Procedure

The experiment was administered in a laboratory setting. The participants were seated alone in a quiet room in front of a computer screen. Test items were presented in black, lower-case font on a white background, randomly paired with images representing unique alien creatures. The participants were told that the words represented the creature names, a manipulation intended to contextualize the pseudowords as nouns. Trial order was pseudo-random, with each target item separated by three fillers of varying length in order to minimize potential sequence effects between trisyllabic metrical frames. The trials advanced automatically after a time interval of 5 seconds for the targets and 3-5 seconds for the fillers, depending on length. The participants were instructed to consider each word silently, decide how to pronounce it so that it would sound as natural and English-like as possible, and finally to read it out loud. No mention of stress or syllables was made. A headset microphone was used to record responses for offline coding of stress placement and acoustic analysis.

5.4.2.4 Data Pre-Processing

Stress was coded offline with reference to loudness, duration, pitch movement and vowel centralization (see Cutler, 2005). In the event of multiple productions within the 5 second response window, only the final attempt was considered. Responses were coded into five categories: antepenult stress, penult stress, final stress, ambiguous stress, and production error. A total of 5,100 response trials were recorded (30

participants x 170 items). Of these, 956 (18.7%) were coded as errors and excluded from the analysis (these are analyzed separately in Study 5 below).

Of the 4,144 error-free responses, 174 (4.2%) featured tense or diphthong realizations of stressed vowels. These responses confounded the inference of syllable boundaries because codas were not required to make the syllables heavy; they were therefore excluded from the analysis. Finally, 191 items (4.6%) received final stress and 364 productions (8.8%) elicited 'ambiguous' judgments. These items were included in the reliability check (see below); however, the main analysis was restricted to those productions where stress was clearly placed on either the antepenult or the penult. These amounted to 3,415 tokens, about 82% of the error-free productions.

5.4.2.5 Reliability

To assess the reliability of the coding, 878 randomly selected tokens (~25% of total, evenly distributed across the cluster types and speakers) were judged by a second listener who was a native American English speaker trained in phonetics. Agreement was near perfect (97.5% of cases, Cohen's $\kappa = .933$, $z = 27.7$). The 22 tokens which resulted in coding disagreement were reviewed before making the final decision.

In addition to being subjected to inter-rater reliability, the coding was checked against the same two acoustic correlates used to verify the stimuli in Study 3: duration, and intensity. To calculate the relevant measures, all 3,527 error-free productions (including final and ambiguous stress, but excluding stressed long vowels and diphthongs) were hand-segmented and phonetically transcribed in Praat. For the vast majority of the items, the visual information provided in the spectrogram and

waveform views was sufficient to clearly identify segment transitions. The only exceptions occurred in a small subset of illegal fall items that featured heavily coarticulated vowel+liquid sequences. Two strategies were simultaneously adopted to deal with these tokens. The first was to simply place the boundary at roughly the midpoint of the sequence, assigning half of the duration to each segment (see also Redford, 2008). The second was to treat the entire unit as vocalic as in Morrill (2012). For example, a heavily coarticulated production of *thanarbiss* (stressed on the antepenult) would be transcribed in two ways: as [θænəɹbɪs] and [θænəˈbɪs]. The two segmentations are illustrated in Figures 5.15 and 5.16, respectively.

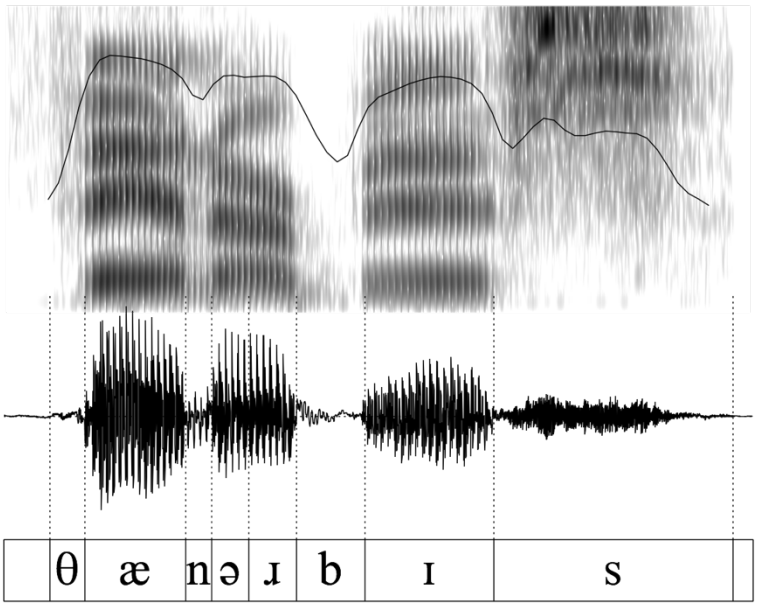

Figure 5.15. Spectrogram with superimposed intensity contour (top), segmented wave form (middle) and transcription (bottom) of the pseudoword *thanarbiss* (antepenult stress), with the rhotic separated from the penultimate vowel. Time is on the x-axis. Frequency (spectrogram), intensity (curve) or pressure (waveform) on the y-axis.

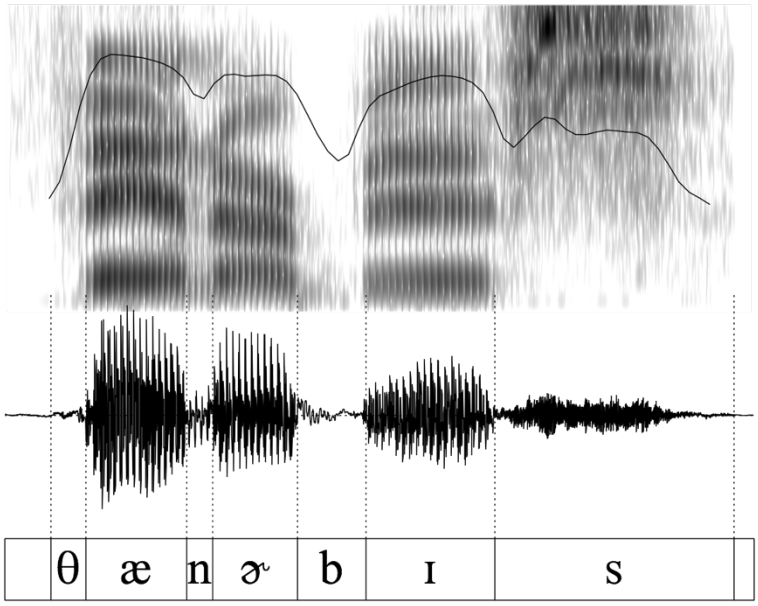

Figure 5.16. Spectrogram with superimposed intensity contour (top), segmented wave form (middle) and transcription (bottom) of the pseudoword *thanarbiss* (antepenult stress), with the rhotic included in the penultimate vowel. Time is on the x-axis. Frequency (spectrogram), intensity (curve) or pressure (waveform) on the y-axis.

Since the acoustic correlate measures relied on vocalic intervals, I took the conservative approach of keeping both segmentation versions and deriving measures for each one; these were subsequently entered into separate statistical models. Because the results were qualitatively unaffected by the segmentation strategy, I arbitrarily report the measures derived from the segmentations that split coarticulated vowels and liquids at the midpoint.

Figure 5.17 presents the two acoustic correlates plotted as a function of coded stress. The left panel shows the duration-based correlate. In order to derive this measure, I calculated the durations of the first and second vocalic intervals, and divided the latter by the former in order to normalize for speech rate differences. As the panel shows, items coded as having penultimate stress featured longer penultimate vowels (ratio = 4.12), whereas in words coded with initial stress, the vowels were

approximately equal in duration (ratio = .97). Note also that the ambiguous cases were intermediate on the measure.

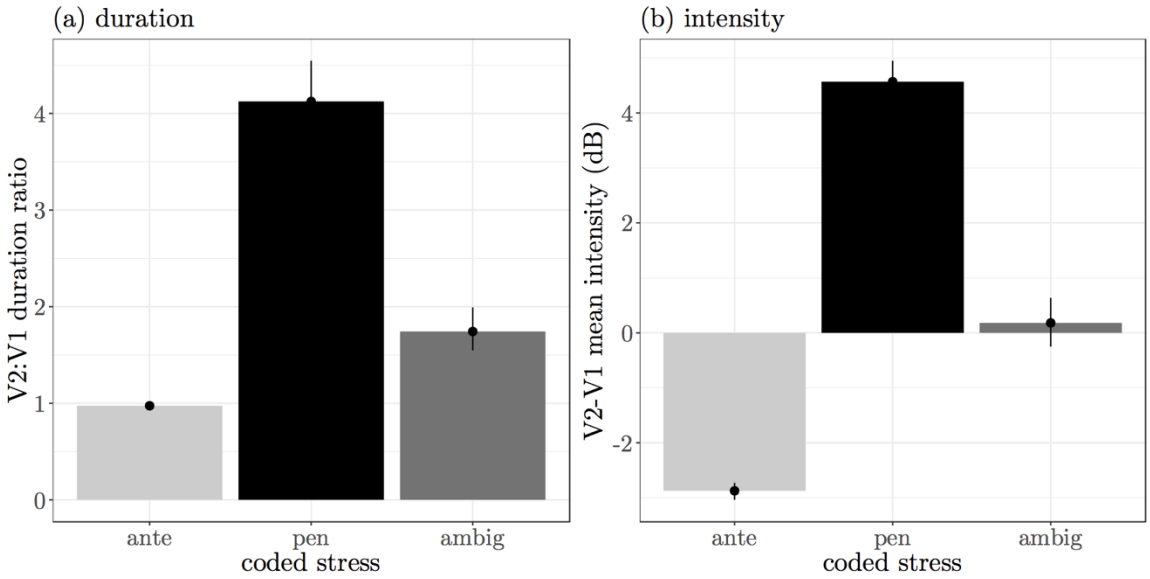

Figure 5.17. Acoustic correlates by coded stress. Error bars are 95% confidence intervals obtained via non-parametric bootstrap.

To test for the significance of the pattern seen in the figure, a linear model was fit to the data, predicting the log-transformed duration ratios from the stress coding. The model significantly improved fit over a null model that featured only the random effects ($\chi^2(2) = 81.33$, $p < .001$). The results of planned comparisons revealed items coded with penult stress featured significantly higher V2:V1 duration ratios than items perceived as antepenultimate-stressed ($\beta = 1.25$, S.E. = .07, $t(52.73) = 16.84$, $p < .0001$) and items perceived as ambiguous ($\beta = .64$, S.E. = .06, $t(22.08) = 9.96$, $p < .0001$). Words coded as ambiguous also featured significantly higher V2:V1 duration ratios than words placed in the antepenult category ($\beta = .51$, S.E. = .05, $t(29.80) = 11.03$, $p < .0001$).

The right panel in Figure 5.17 shows the intensity correlate. This measure was calculated by subtracting the mean intensity of the first vocalic interval from that of the

second (the values for each interval were calculated by averaging the intensity contour over the interval's duration). The plot reveals a similar pattern to that of the duration ratios. Stressed vowels (especially penults) were higher in mean intensity than unstressed vowels, whereas words where both vowels were approximately equal in intensity elicited ambiguous judgments. A linear model testing this relationship significantly improved fit over a null model ($\chi^2(2) = 57.16$, $p < .0001$). Results of the simple comparisons revealed that the intensity measure was distributed across the stress judgments as depicted in the figure (penult vs. antepenult: $\beta = 7.00$, S.E. = .54, $t(36.97) = 13.04$, $p < .0001$; penult vs. ambiguous: $\beta = 4.02$, S.E. = .44, $t(53.93) = 9.17$, $p < .0001$; ambiguous vs. antepenult: $\beta = 2.57$, S.E. = .33, $t(22.80) = 7.77$, $p < .0001$).

Taken together, the results of the reliability analysis indicate that the coders were consistent with each other in relying on duration and intensity, two of the acoustic correlates implicated in the realization and perception of English lexical stress. I now turn to the main results of the experiment.

5.4.3 Results

5.4.3.1 Nuisance Covariates

This section examines the effects of the two nuisance covariates on the stress assignment responses. Figure 5.18 shows the scatterplots of each nuisance measure against the log-odds of penult stress assigned by the participants. The data were aggregated by test item, yielding 170 unique data points for each panel.

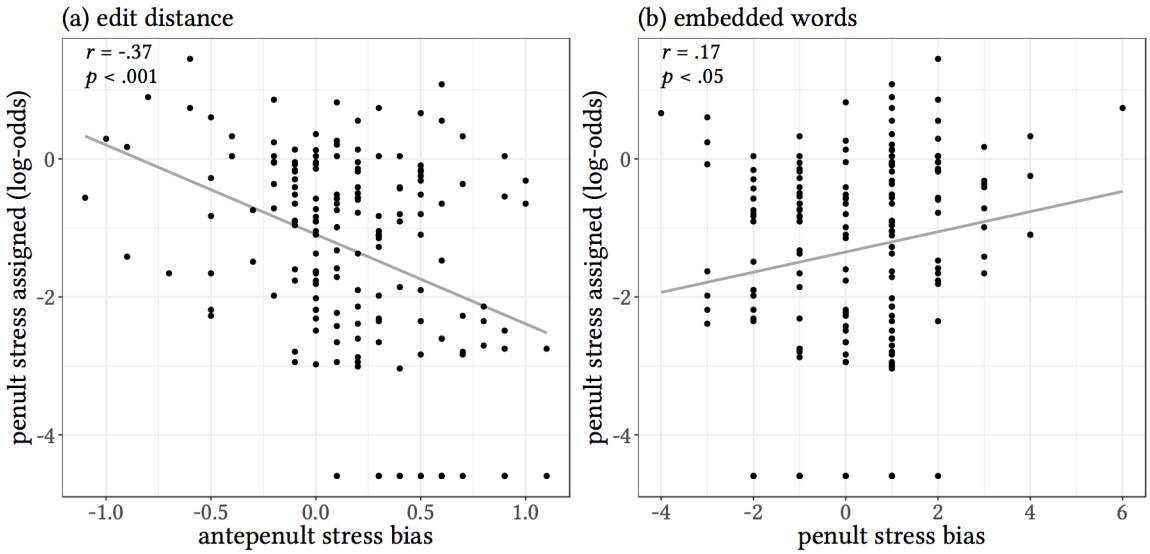

Figure 5.18. Effects of nuisance covariates on stress assignment (all test items).

Panel (a) displays the effect of the covariate based on edit distance. Positive values on the x-axis indicate test items that, on average, were closer to antepenult- than penult-stressed lexical neighbors. The relationship to the responses was in the expected direction, with pseudowords closer to penult-stressed neighbors more likely to receive penult stress than pseudowords closer to antepenult-stressed neighbors. The correlation was significant, with the edit distance measure capturing some 14% of the word-level variance in responses. A univariate, mixed-effects logistic regression fit to the raw data indicated that the effect of edit distance was significant (β = -1.32 *S.E.* = .51, $p < .05$). With each edit closer to antepenult-stressed neighbors, the odds of stressing the penultimate syllable decreased by a factor of .27.

Panel (b) displays the covariate based on embedded words. Positive values on the x-axis signify test items for which the number of embedded words cuing penult stress outnumbered those favoring antepenult stress. Unlike in Study 3, the relationship was clearly positive, indicating that embedded words had an effect on stress assignment. The correlation was statistically significant, though the effect was rather

small with embedded words accounting for about 3% of the variance in the word-level responses. A univariate, mixed-effects logistic regression fit to the raw data failed to find a significant effect of embedded words (β = -.07 *S.E.* = .16, p = .65).

5.4.3.2 Coarse-Grained Phonotactics

This section investigates the effect of coarse-grained phonotactics on stress assignment. Figure 5.19 plots the proportion of penult stress at each level of insert status. For items containing singleton inserts, approximately 11% were stressed on the penult. This rate rose to 23% in pseudowords with attested clusters and 42% in items with unattested CC inserts.

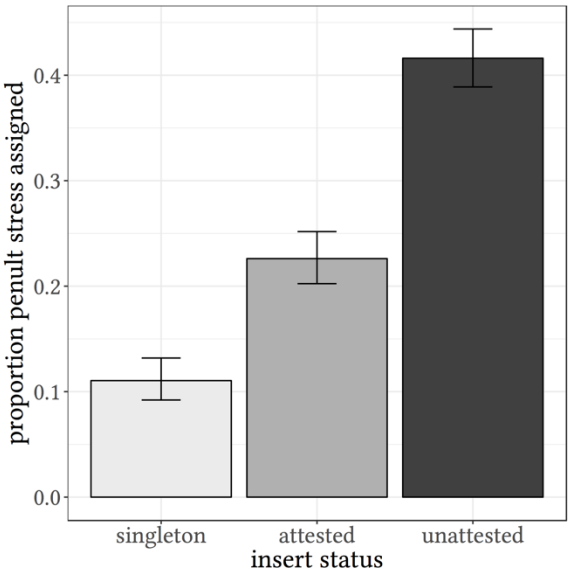

Figure 5.19. Penult stress by insert status. Error bars are 95% confidence intervals based on the proportion test.

To test for the significance of the differences seen in the figure, a maximal, mixed-effects logistics regression was fit to the data. In addition to insert status, the

model contained the two nuisance predictors (edit distance bias and embedded word bias), which were centered and scaled prior to their inclusion. The model significantly improved fit over an intercept-only version ($\chi^2(4) = 44.93$, $p < .001$). The model output is presented in Table 5.4.

Table 5.4. Categorical model output (stress assignment task).

	Estimate (Std. Error)
Intercept (Status = singleton)	-2.965 (0.370)***
Status = attested	0.896 (0.261)***
Status = unattested	2.392 (0.284)***
Edit distance bias	-0.295 (0.119)*
Embedded word bias	0.162 (0.124)
Observations	3,415
Log Likelihood	-1,414.026
Bayesian Inf. Crit.	3,112.809
Note:	*p<0.05; **p<0.01; ***p<0.001

With the intercept set to singleton items at mean covariate values, the effect of insert status emerged as statistically significant. Specifically, the odds of penult stress on items with attested CC onsets increased by a factor of 2.45 over singleton-containing items. For words with unattested inserts, the odds ratio over singleton items increased to 10.93. A likelihood ratio test indicated that edit distance bias also significantly improved fit ($\chi^2(1) = 5.27$, $p < .05$). However, this was not the case for the embedded word bias ($\chi^2(1) = 1.53$, $p = .22$).

To test whether pseudowords with the two cluster types differed from each other in stress placement, a simple comparison was conducted via a second logistic regression. The results indicated that the odds of penult stress in items with unattested

clusters were significantly higher than in items with attested items by a factor of 4.48 (β = 1.50, *S.E.* = .25, p < .001).

To sum up, the patterns seen in Figure 5.19 were confirmed. Each level of insert status elicited significantly different rates of penult stress, indicating that the participants were sensitive to coarse-grained phonotactics during online stress assignment. I now turn to the question whether this phonotactic awareness was more fine-grained than suggested by these results.

5.4.3.3 Fine-Grained Phonotactics

In this section, I examine the influence of fine-grained generalizations on stress assignment. I begin by investigating insert-level correlations between each phonotactic predictor and the responses. For singletons and attested cluster inserts, Figure 5.20 plots the relationship between stress assignment and word onset frequency. There are 40 data points representing the 12 unique singletons and 28 unique attested clusters.

The relationship seen in the plot is negative, with frequent word onsets resisting penult stress when placed between the penultimate and final vowels. The correlation was relatively strong and statistically significant. Onset frequency accounted for approximately 38% of the variance in insert-level responses.

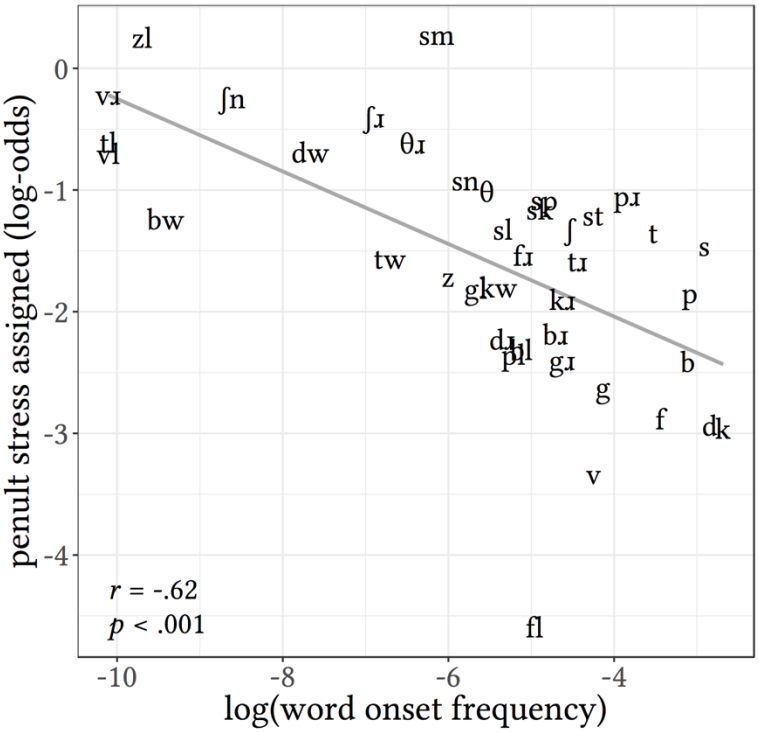

Figure 5.20. Log-odds of penult stress assigned by word-initial frequency of each embedded insert (singletons, attested CC onsets).

A mixed-effects logistic regression tested this relationship on the raw responses to singleton and attested items. The results were consistent with the gradient hypothesis, with word onset frequency significantly predicting the rate of penult stress (β = -.37 *S.E.* = .07, p < .001). With each log unit increase in onset frequency, the odds of stressing the penult decreased by a factor of .69. The effect remained significant even after the five marginal onsets (/bw, tl, vl, vɹ, zl/) were removed from the model (β = -.39 *S.E.* = .12, p < .01).

Figure 5.21 plots the correlation between the responses and the second gradient predictor, word offset frequency of the C1 of each insert.

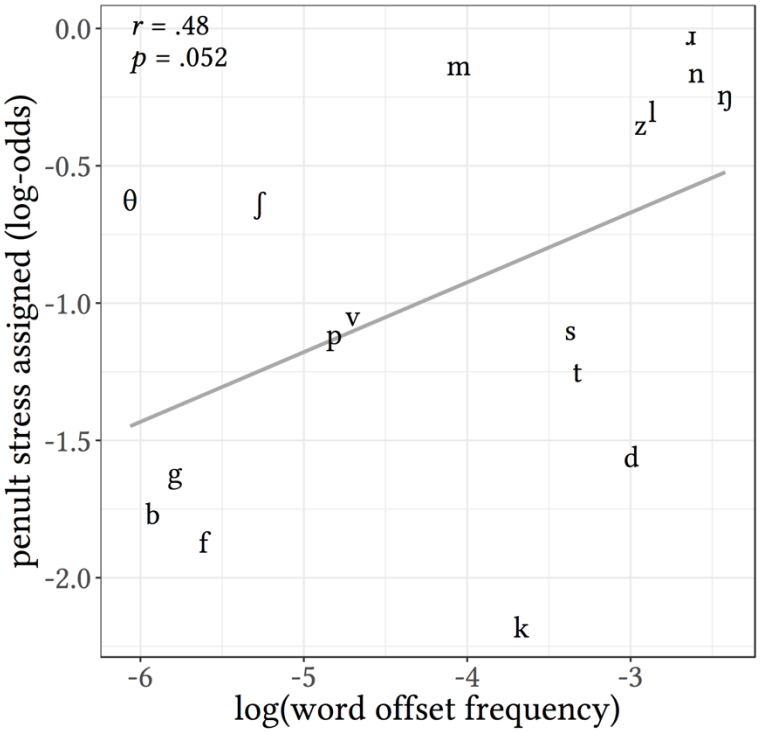

Figure 5.21. Log-odds of penult stress assigned by word-final frequency of the C1 of each embedded insert.

The relationship between the predictor and responses is positive, with frequent word offsets more likely to lead to penultimate stress when placed in medial position. In spite of there only being 17 data points, the correlation was nearly significant.

To test the significance of the effect on actual response data, a mixed-effect logistic regression was fit to the raw data. Word offset frequency was found to significantly affect stress placement (β = .92 *S.E.* = .18, p < .001). As offset frequency increased by one unit on the log scale, the odds of stressing the penult increased by a factor of 2.5. The effect persisted even after removing /ŋ/ from the data (β = .87, *S.E.* = .20, p < .001), indicating that this categorically illegal onset did not drive the relationship.

The correlation between penult stress and the final gradient predictor of interest, sonority slope, is shown in Figure 5.22. As in Studies 1-3, the data were limited to the 35 unattested inserts.

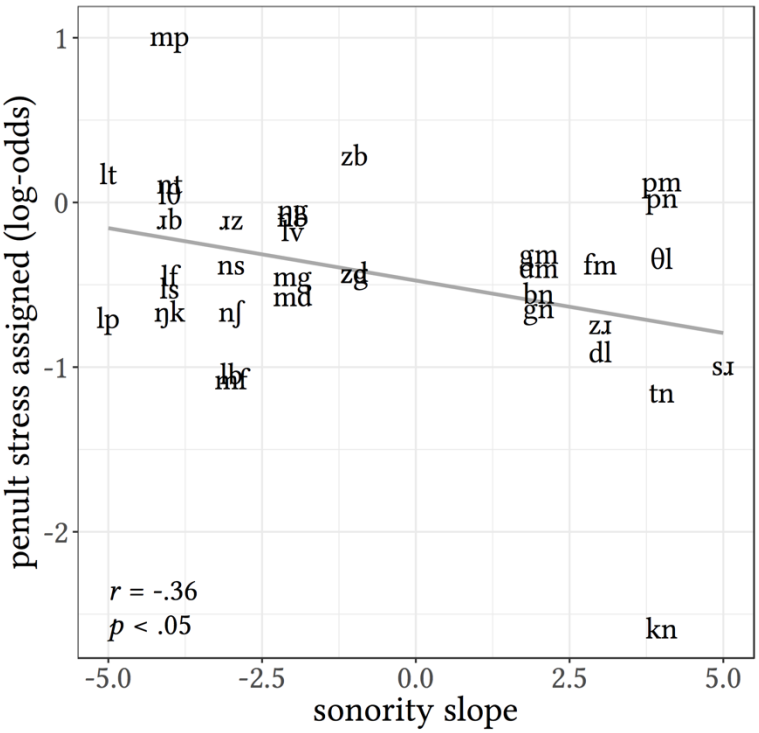

Figure 5.22. Log-odds of penult stress assigned by sonority slope of each embedded insert (unattested clusters only).

As seen in the figure, the correlation was statistically significant, and the direction of the relationship was consistent with the SSP: among the unattested onsets, those with rising sonority were slightly more likely to resist penultimate stress. Sonority slope captured approximately 12% of the variance in the insert-level responses.

To test the significance of the sonority effect on level-1 responses, a mixed-effects, logistic regression model fit to the unattested items data. Although the model suggested a trend in the expected direction, sonority slope failed to reach statistical significance (β = -.08, *S.E.* = .04, p = .08).

In order to examine the performance of each gradient predictor in the presence of the others, a multiple, mixed-effects model was fit on the full data. In addition to onset frequency, offset frequency and residualized sonority slope, the model contained the edit distance and embedded word-based nuisance covariates. After the maximal model failed to converge, the random-effects correlation parameters were removed from the estimating formula. This reduced model converged successfully and was a significant improvement over an intercept-only model according to the likelihood ratio test ($\chi^2(5) = 54.56$, $p < .001$). The model output is presented in Table 5.5 while the odds ratio estimates and marginal effects are plotted in Figure 5.23.

Table 5.5. Gradient model output (stress assignment task).

	Estimate (Std. Error)
Intercept	-1.702 (0.321)***
Word Onset Frequency	-1.013 (0.117)***
Word Offset Frequency	0.122 (0.128)
Sonority Slope	-0.040 (0.102)
Edit Distance Bias	-0.355 (0.101)***
Embedded Words Bias	0.151 (0.123)
Observations	3,415
Log Likelihood	-1,387.334
Bayesian Inf. Crit.	2,921.115
Note:	*p<0.05; **p<0.01; ***p<0.001

The output of the model revealed that onset frequency had a significant effect on stress assignment. With each standardized unit increase in onset frequency, the odds of stressing the penultimate syllable decreased by a factor of .36. The effect of the edit distance-based nuisance variable also emerged as significant: as the similarity to antepenult-stressed words increased by one *z*-score, the odds of stressing the penult

decreased by a factor of .70. The other predictors in the model all showed numerical trends in the expected direction, but none emerged as statistically significant effects.

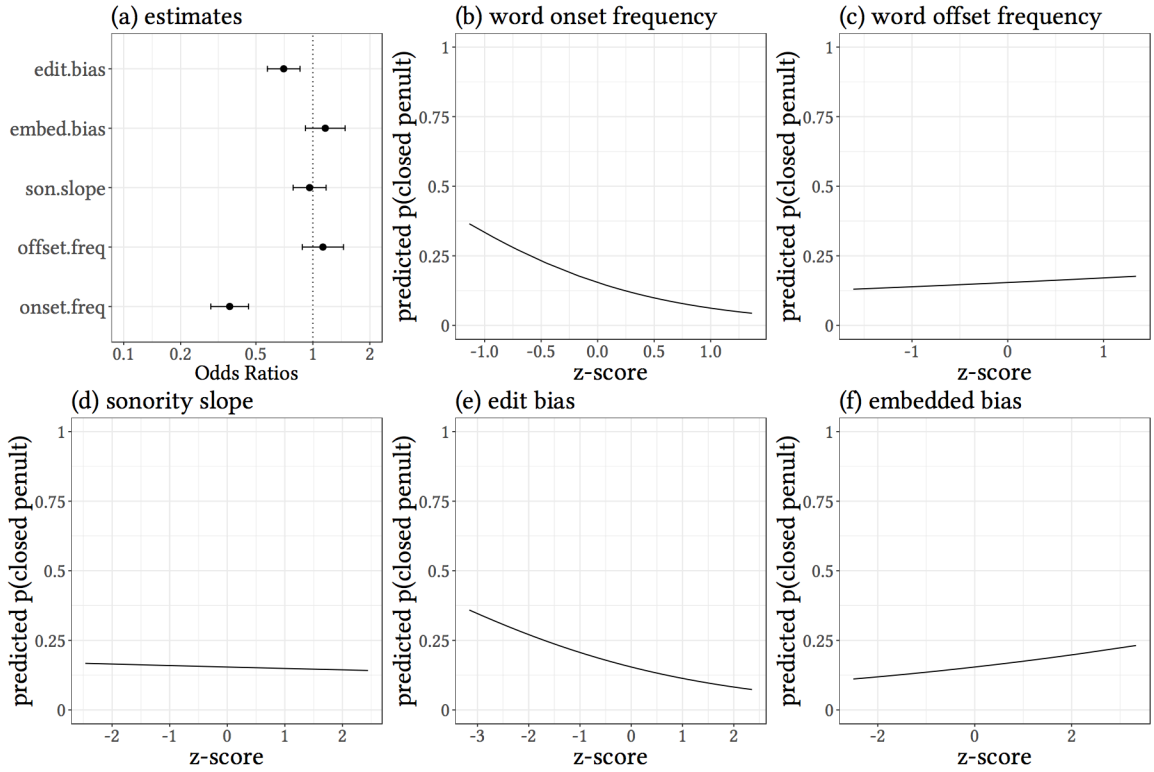

Figure 5.23. Gradient model estimates (panel [a]; dotted vertical line represents the null hypothesis) and marginal effects (panels [b]-[f]).

To sum up, the participants appeared to be sensitive to gradience in stress assignment, with the caveat that their sensitivity was restricted to the phonotactics of word onsets. In addition, they were influenced by analogy to known words, as captured by edit distance. In the next section I ask how the gradient model compares to the categorical model in fitting the stress assignment data.

5.4.3.4 Model Comparison

Prior to comparing the two models, some adjustments were necessary. Recall that the gradient model failed to converge in the maximal configuration, necessitating the removal of the random correlation parameters. Because of this, the gradient model's likelihood penalty (assigned by the *BIC* formula) was disproportionally lenient relative to the maximal categorical model. In order to facilitate the comparison on an equal footing, the categorical model was therefore refit with the random correlation parameters removed. With respect to the fixed effects, the results of this reduced model were nearly identical to the original model's findings and led to the same conclusions. In this section, the reduced models are compared.

Following the comparison strategy in Studies 1-3, I begin by comparing each model's insert-level predictions to the observed values. The predictions were generated by conditioning on the fixed effects. The correlation plots are shown in Figure 5.24, where each of the 75 point represents the average values for a unique insert.

As in Study 3, the categorical model contained two continuous covariates. Therefore, its aggregate predictions are distributed along the x-axis rather than restricted to 3 values (as in Studies 1 and 2). That said, closer inspection of the scatterplot in panel (a) reveals that the variation in predictions is largely within levels of insert type, indicating that the two covariates rarely pushed the model to predict against the categorical phonotactics. This is not the case for the gradient model plotted in panel (b), where the predicted values of attested clusters overlap greatly with singletons and to a lesser extent with unattested onsets. Of course, this difference is because the gradient model did not contain insert level as a predictor and was not

forced to bin its predicted values. As it turns out, the additional phonotactic flexibility resulted in a predictive advantage, as evidenced by the lower mean squared deviation and and additional 13% of captured variance in the aggregate responses. This improvement in variance reduction is larger than than that seen in the stress preference task (Figure 5.13), but very much in line with the hyphenation and Eddington et al. (2013ab) reanalysis results (Figures 4.6 and 4.12).

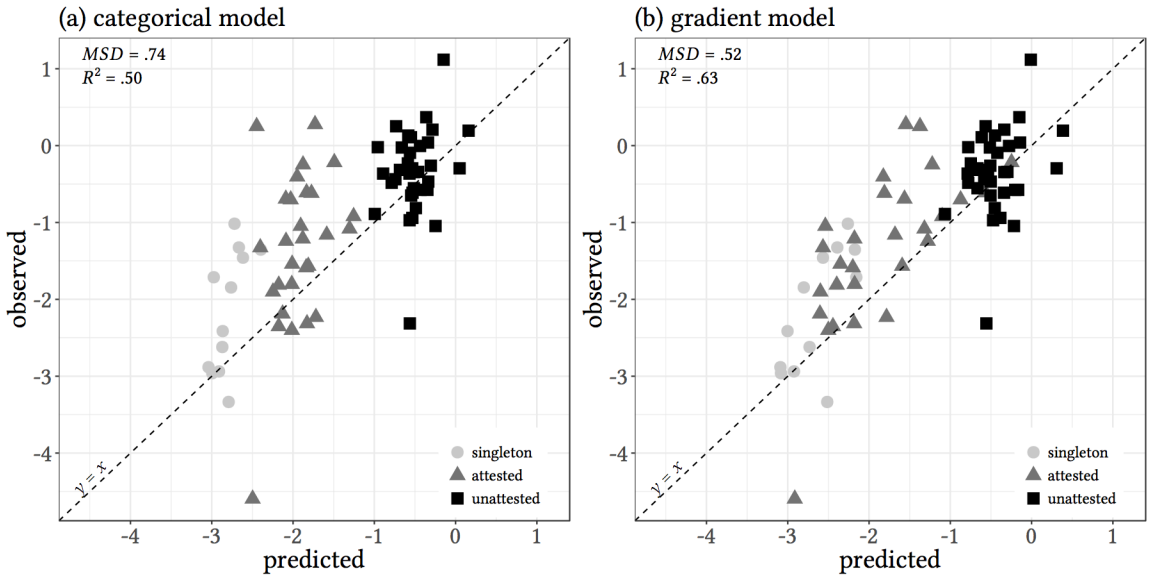

Figure 5.24. Comparison of model predictions (stress assignment data). Values are in log-odds.

In terms of the raw response data, the *BIC* scores were about 3,025 for the categorical model and about 2,921 for the gradient model (again, both models did not estimate random correlation parameters and were thus on an equal footing). This difference of approximately 104 points translated to a Bayes Factor in excess of 4.2×10^{22} for the gradient model, which in turn corresponded to a posterior probability essentially equal to 1. As in Studies 1, 2, and 3, a rational learner provided with a choice

between the two parsing models would virtually always infer the gradient model despite some penalty for its increased complexity.

5.4.4 Discussion

Like hyphenation (Studies 1 and 2) and stress preferences (Study 3), online stress assignment appears to have been driven by gradient rather than categorical phonotactics. More specifically, word onset frequency captured a significant portion of the variance, even after edit distance and embedded words were controlled for in the model. Although offset frequency and sonority slope showed trends in the expected direction, neither predictor reached significance in the full model. Nevertheless, the gradient parsing model outperformed the categorical alternative, both in predictive power and according to the *BIC* score comparison. Furthermore, the stress assignment task proved to be more sensitive to gradience than the 2AFC task in Study 3 — the effect size of word onset frequency was considerably larger in the production study. This was an expected result (see section 5.3.4).

If the relationship between the hyphenation and stress assignment results argues that both tasks were subserved by the same metrical parse, it also reveals some inconsistencies. Namely, the range of responses in Study 4 was narrower than in the hyphenation task. To illustrate, compare Figures 4.1 and 5.19: the difference in closed penult rates between singletons and unattested CC words onsets is about 53% in the former but only 31% in the latter. Furthermore, hyphenation yielded much higher rates of closed penults overall than did stress: 72% vs. 26%, respectively. Why the difference between the two tasks?

One possibility is that, relative to the underlying parse, the participants in Study 1 were too liberal in closing syllables. Recall that 41% of the singleton inserts were assigned to the preceding syllable, a finding at odds with well-established theoretical arguments for onset filling (e.g. Itô, 1989). It is plausible that this coda bias was a manifestation of the Possible Word Constraint (Norris et al., 1997) which emerged during the sequential processing of orthography. As the participants worked their way across the character string, they likely felt some pressure to produce heavy syllables in order to satisfy the word minimality requirements of English. If the vowel was interpreted as lax, this meant appending a coda. If the order of processing was indeed left-to-right, the minimality bias emerged prior to and thus was able to compete for the parse with the downstream phonotactic dependencies. One piece of evidence consistent with this argument is the structure of antepenults in the hyphenation study. Recall that every pseudoword contained a singleton between V1 and V2. Subsequent analysis revealed that 54% of these consonants were parsed as antepenult codas, suggesting that word minimality was indeed competing with onset filling.

Independently of the minimality bias, the second reason for the difference between the two experiments may originate in the lexical statistics of English stress. There are at least two possibilities. First, the lexicon could have affected productivity through shorter words, which are overwhelmingly stressed on the initial syllable regardless of weight (Cutler & Carter, 1987). Allowing any of these words to infiltrate the search space would result in competition between the Germanic (i.e. initial) and Latin stress patterns, resulting in lowered productivity of the latter (but see Yang, 2005; Legate & Yang, 2012 for a different model of productivity). Second, recall that, while weight sensitivity is robust across the lexicon, it is by no means categorical. For

instance, Figure 5.3a shows that, across all trisyllabic and longer word forms, only about 58% of heavy penults are stressed. Even if we restrict the definition of heavy syllables to those with long vowels (and thus circumvent the potential issues arising from the choice to syllabify the lexicon in section 5.2.2 according to the Maximal Onset Principle), the rate of heavy penults receiving stress barely crosses 61%. It is therefore possible that the lexical statistics of weight add another stochastic dimension to the results: having parsed the pseudowords according to the fine-grained model, the participants may have probability-matched the weight generalization in the lexicon. Indeed, it is not unlikely that multiple weight generalizations are involved in stress assignment, perhaps organized into a weight gradient based on vowel quality (Carpenter, 2010; Hitchcock & Greenberg, 2001, see next section for discussion). A comprehensive treatment of the interaction between the gradient parser and gradient weight phenomena remains an area for future work.

5.4.4.1 Alternative Explanations

Before accepting the idea that stress assignment was guided by gradient phonotactic generalizations over word edges, a number of alternative explanations must be addressed. These include the possibility that (a) the parser was categorical but the resultant syllables differed along a weight continuum, (b) insert phonotactics do not matter because the domain of weight computation is not the syllable but the V-to-V interval, and (c) rather than generalizing over word edges, the participants were tracking the relationship between medial clusters and stress in the lexicon. This section addresses each of these major objections in turn.

5.4.4.1.1 Categorical Parse, Gradient Weight

Recent empirical work on weight-sensitive stress systems has argued for a gradient treatment of weight in some languages previously assumed to have a binary L/H distinction. For example, Ryan (2011b) examined poetic corpora from Homeric Greek, Kalevala Finnish, Old Norse and Middle Tamil, and argued that each meter showed evidence of a four-level weight system based on rime complexity. After conducting a quantitative analysis of the Portuguese lexicon, Garcia (2017) demonstrated that stress assignment is stochastic and dependent on a complex interaction between onset size, nucleus size, coda size and the position of the syllable within the trisyllabic stress window. In Spanish, evidence for gradient weight was presented in Shelton, Gerfen & Gutiérez Palma (2012), who used a pseudoword naming task to investigate the stress attracting properties of diphthongs. Shelton and colleagues found that penults with falling diphthongs (*fa.tei.ga*) attracted more stress than penults with rising diphthongs (*do.bia.na*), leading to the novel conclusion that Spanish CVG syllables are heavier than CGV syllables.

A number of studies have demonstrated gradient weight effects in English. Kelly (2004) was among the first to note the influence of onset structure, finding that word onset length correlated positively with initial stress in English disyllables. Crucially, native speakers extended this generalization to disyllabic pseudowords. As for rime complexity, Ryan (2011a) showed that English monomorphemic disyllables follow a four-level weight hierarchy, which is also extended to nonce forms by adult speakers. In subsequent work, Ryan (2014) developed a gradient weight model that integrated

onset and rime effects. The proposal was based on the idea that the left edge of the weight domain is not at the onset-rime boundary but rather at the perceptual center (p-center), the moment at which a syllable is registered by the perceptual system (see Morton et al., 1976). Increasing onset complexity shifts the p-center leftward; Ryan (2014) calculated that adding a segment to the onset adds about a third of the weight to the syllable compared to adding a segment to the coda.

Whenever relevant, these gradient weight proposals have made the assumption that the metrical parse is phonotactically coarse-grained. This assumption made the assignment of weight relatively straightforward; one simply needed to correlate categorically-determined syllable structure with stress. The cross-linguistic success of the gradient weight hypothesis raises an important objection to the claim that the results of Study 4 reflected a stochastic parser. The suggested alternative is that syllabification was in fact categorical, but the penults which resulted from this parse varied along a weight continuum, resulting in gradient stress assignment.

Before addressing this possibility, it is of interest to examine the extent to which Latin Stress exhibits any weight-based gradience in the English lexicon. As a first step, it is important to determine whether the results reported in Kelly (2004) and Ryan (2011b) extend to the penultimate syllables in longer words. This is by no means a foregone conclusion, as there has been evidence of structure interacting with position elsewhere. Specifically, Garcia's (2007) study of Portuguese revealed that the relationship between rime complexity, onset complexity and stress was different for antepenultimate and penultimate syllables. In antepenults, longer onsets attracted stress while longer codas repelled it; in penults, these correlations were reversed.

To examine the sensitivity of English stress to fine-grained penult structure, I relied on the same lexicon examined throughout this dissertation. As in section 5.2, the words were syllabified in accordance with the Maximal Onset Principle, and only words longer than 2 syllables were examined. Figure 5.25 plots the proportion of penult stress as a function of rime complexity (short vowels and consonants were each assumed to contribute one mora). The panels vary in morphological restrictions on the words (cf. section 5.2).

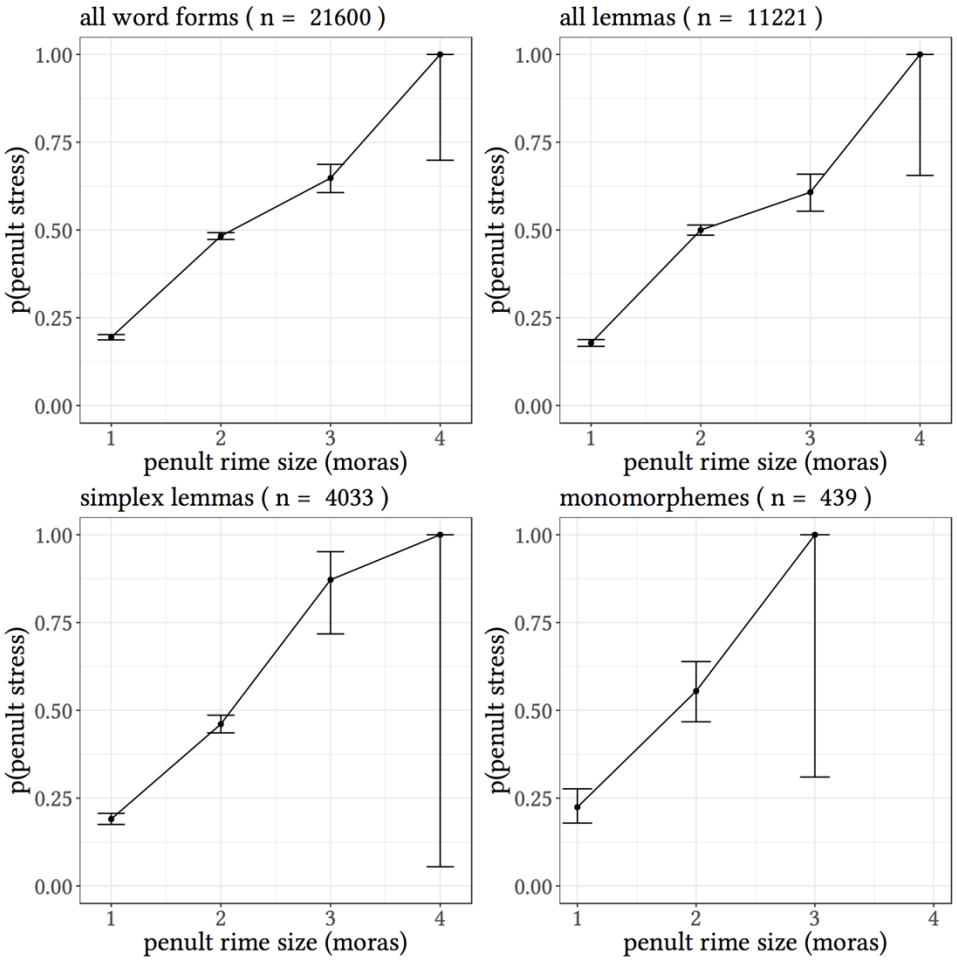

Figure 5.25. Penult stress as a function of penult rime complexity across different subsets of the lexicon (trisyllabic and longer words). Error bars are 95% confidence intervals based on the proportion test.

The panels show a clear trend whereby the probability of stress appears to rise monotonically with rime complexity. This trend was tested with a series of mixed-effects logistic regressions comparing each level of rime complexity to the adjacent level. The models featured random intercepts for words and the alpha levels were Bonferroni-adjusted to account for the number of comparisons. As suggested by the error bars, the difference between monomoraic and bimoraic rimes was significant for each subset of the lexicon (all *p*s < .001). Furthermore, the difference between bimoraic and trimoraic rimes was significant for all but the monomorphemes (all *p*s < .001). No lexicon subset featured a significant difference between trimoraic and longer rimes, likely due to the very low number of words instantiating the latter. At least for the less restricted lexicons, these results point to a gradient weight system with 3 distinct levels.

Figure 5.26 plots the effect of penult onset length on stress attraction. The two largest subsets of the lexicon appear to feature a positive correlation, but the pattern seems to break down in the smaller data sets.

Mixed-effects regressions revealed a significant four-level onset weight hierarchy in the word form lexicon (all *p*s < .001) and a binary distinction (CCC vs. others) in the lemma lexicon ($p < .001$). The smaller lexicons did not exhibit a significant effect of onset length on stress. These results suggest that, in order to show a gradient onset effect, learners must have access to a word form lexicon in which syllabification does not necessarily respect morpheme boundaries.

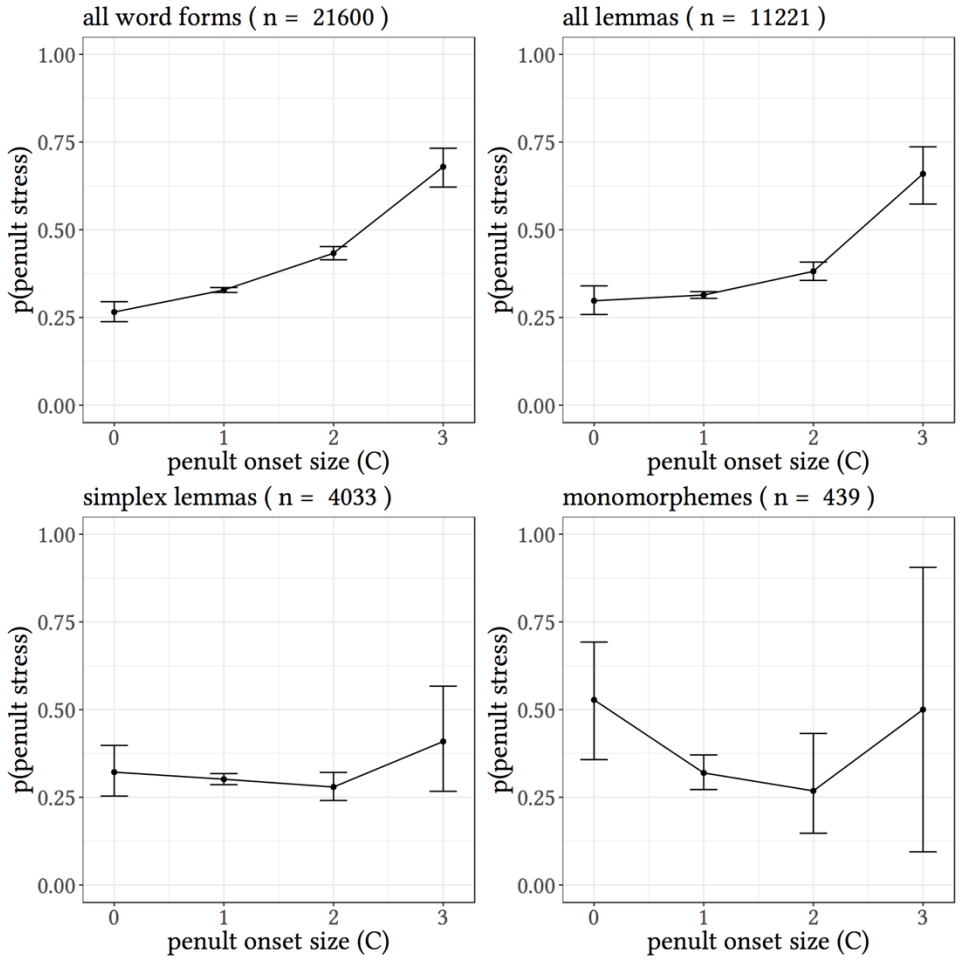

Figure 5.26. Penult stress as a function of penult onset length across different subsets of the lexicon (trisyllabic and longer words). Error bars are 95% confidence intervals based on the proportion test.

Taken together, the rime and onset findings support the idea that Latin Stress can be modeled with a gradient weight model, particularly under the assumption that learners generalize over a minimally restricted lexicon. With this caveat, the findings in Kelly (2014) and Ryan (2011a,b) can be extended beyond disyllables and initial stress. Nonetheless, the categorical parse/gradient weight account cannot explain the results of Study 4. The reason is that there was not enough variability in rime and onset size among the stimuli. Under the categorical parsing model, all penults featured obstruent C onsets, and the rimes were either VC (for items with initially unattested clusters) or V

(singleton, attested). Furthermore, recall that productions with tense penult vowels were excluded from the analysis, so vowel length did not contribute variability to rime weight. The gradient weight hypothesis is especially not equipped to explain the considerable variance in stress assignment in items with medial singletons and attested onsets (both featured CV penults under the maximal onset parse). The gradient parsing model, on the other hand, provided a good fit to the results.

As for the CVC penults in unattested items, the only way to retain the gradient weight hypothesis is to argue that English coda weight parallels sonority (recall from Figure 5.22 that sonority correlated with stress assignment in the pseudowords). There is some precedent for this idea in languages like Kwakwala and Lithuanian, where sonorants are more likely to attract stress than are obstruents (Zec, 1995), but it is not the standard view of English weight. In order to determine whether the lexicon supports this generalization in Latin Stress, I measured the proportion of stress on V̆C penult rimes in trisyllabic and longer words. Figure 5.27 plots the proportions across four subsets of the lexicon.

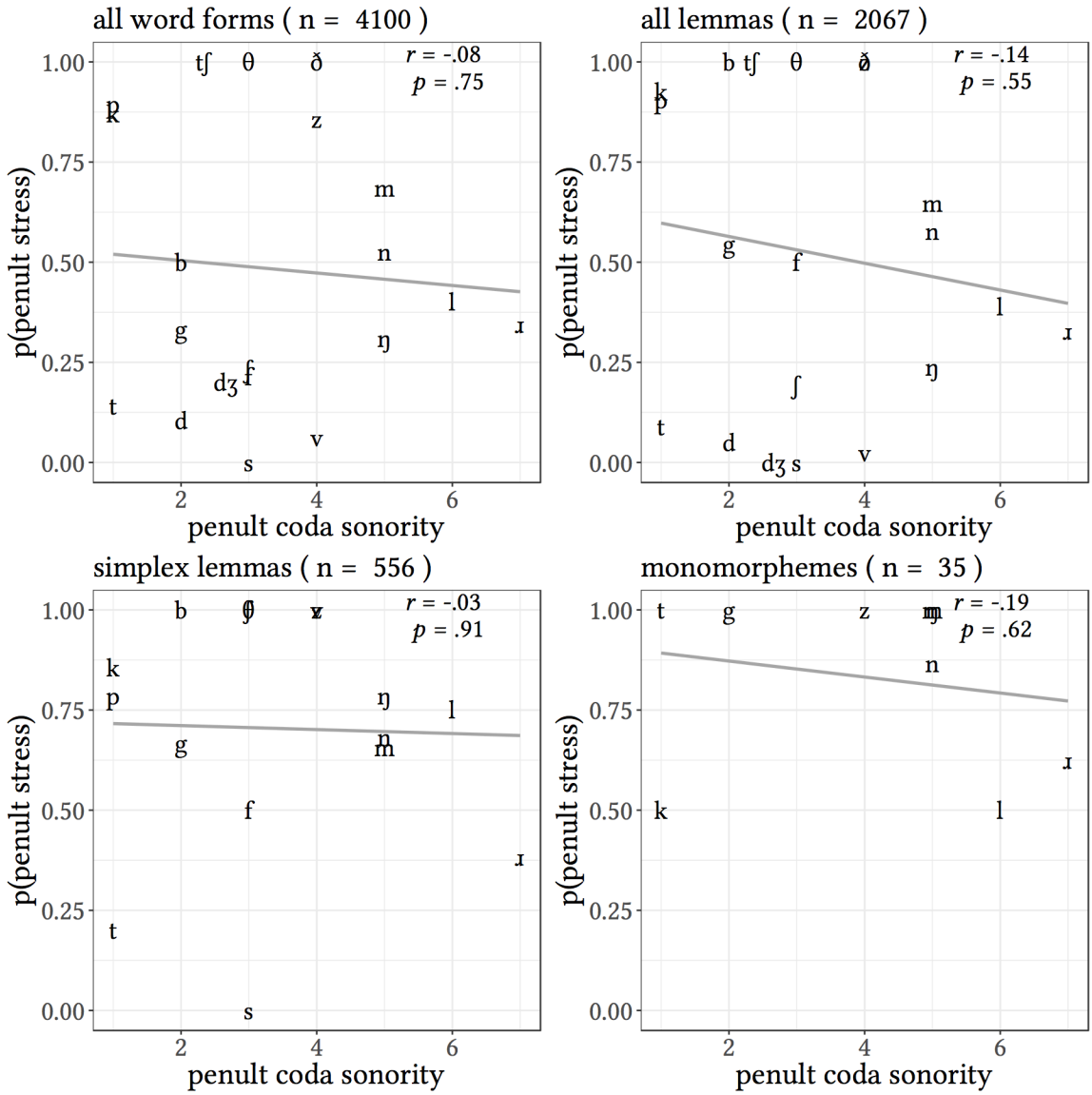

Figure 5.27. Penult stress as a function of penult coda sonority (V̆C rimes only) across different subsets of the lexicon (trisyllabic and longer words).

The figure shows no significant relationship between coda sonority and penult stress for any of the sublexicons. If anything, the regression lines slope downward, suggesting that as coda sonority increases, the likelihood of stress goes down. Figure 5.28 collapses across sonority levels, showing stress on obstruent and sonorant penult codas. The pattern is opposite from that expected by gradient weight: in all but the smallest of the sublexicons, obstruent codas are numerically *more* likely than sonorant

codas to attract stress. Mixed-effects logistic regressions revealed that this pattern was significant for all word words and all lemmas (both *p*s > .001), but not for the two smaller lexicons.

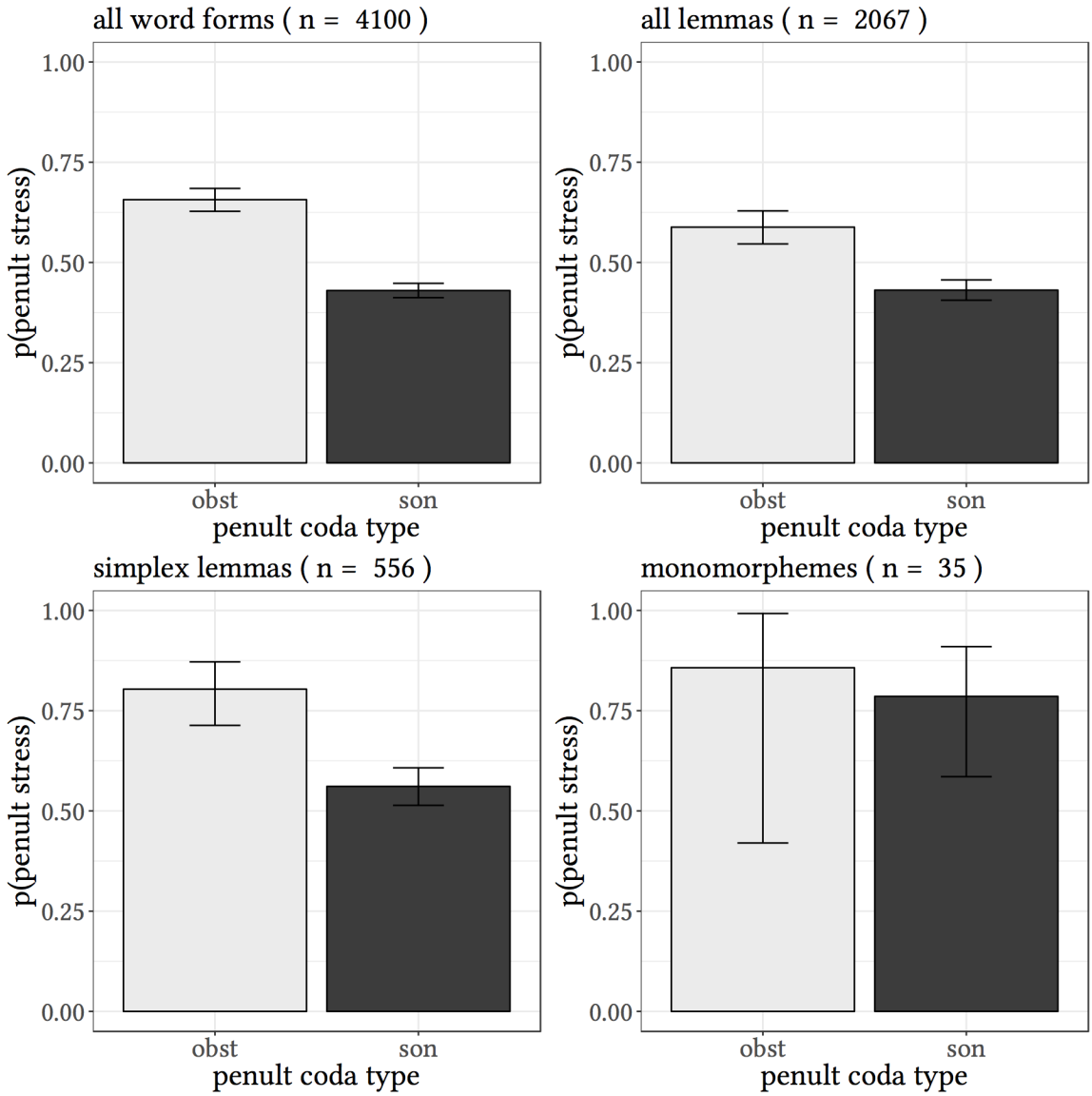

Figure 5.28. Penult stress in obstruent vs. sonorant codas (V̆C rimes only) across different subsets of the lexicon (trisyllabic and longer words).

Overall then, the categorically-parsed lexicon does not provide the learner with the kind of gradient weight generalizations necessary to account for the results of

Study 4. To be clear, this is not to say that English weight is binary, only that the categorical parsing model is inadequate. It may well be that Latin Stress generalizations emerge from the interaction of a gradient parser with gradient weight. Such an interaction might explain why the stress assignment results in Study 4 (Figure 5.22) were less sensitive to sonority than the hyphenation results in Study 1 (Figure 4.4): the parser might prefer to place sonorants into codas, but these attract less stress than obstruents in the same position.

5.4.4.1.2 Interval Theory

Interval Theory (Steriade, 2012) presents an alternative to the rime-based account of weight phenomena. Under this proposal, the metrical parser divides words into intervals rather than syllables, with intervals defined as the span of phonological material beginning with a vowel and ending at the onset of the following vowel or at the word boundary. The interval parse is categorical in nature, assigning all post-vocalic consonants to the preceding vowel. The examples in 5.6 illustrate the difference between an interval-based (a) and a maximal onset-based (b) parse of the word *constructionist*. Note that the interval parse strands word onsets:

5.6a. [<k>.ənstɹ.ʌkʃ.ən.ɪst] (Interval Theory)

5.6b. [kən.stɹʌk.ʃə.nɪst] (Maximal Onset Principle)

Under Interval Theory, intervals constitute the proper domain of weight computation. Steriade (2012) argues for a scalar treatment of interval weight, and

proposes a hierarchy based on a familiar combination of complexity and sonority (different languages make different uses of the scale, recognizing only some of the levels as distinct):

$$VVC > VV, VC_{[son]}C > VC_{[obst]}C > VC_{[son]} > VC_{[obst]} > V_{[+lo]} > V_{[-hi]} > V_{[\neq ə]} > ə$$

Applied to the pseudoword stimuli, the interval parse yields a <c>.VC.VC.VC output for singleton items and <c>.VC.VCC.VC for all others. In order to show that the participants relied on this type of parser, one must demonstrate that the weight of penultimate intervals predicts stress assignment better than the gradient syllable-based parser.

How can one compare the predictions of the two models? Both agree that singleton items should receive less penult stress that words with embedded clusters. For the gradient parser, this is mainly because C word onsets tend to be frequent and therefore parse as such in medial position, leaving an open penult. The interval explanation is simply that VC is lighter than VCC. Both models also largely agree that sonorant-initial clusters should attract more penult stress than obstruent-initial clusters: for Interval Theory, this is a stipulation (see hierarchy above); for the gradient parser, it falls out from a combination of word-edge statistics and fine-grained sonority information. But there was more gradience in the human behavior than can be captured by the $VC_{[son]}C > VC_{[obst]}C > VC$ generalization. At least as specified above, the phonological weight hierarchy is unable to account for much of the gradience observed within VCC intervals.

One way to incorporate additional gradience into a theory of intervals is to ground intervals in phonetic substance. Following Gordon's work on the phonetic basis of syllable weight (Gordon, 1999, 2002), Steriade (2012) suggests that intrinsic duration differences among consonants may have consequences for the assignment of weight to intervals (see also Hirsch, 2014). For example, she hypothesizes that, because [s] is intrinsically longer than [ɹ], the [Vks] interval in *aksa* is heavier than the [Vkɹ] interval in *akra* and should therefore attract initial stress more readily. To my knowledge, Lunden (2017) constitutes the only acoustic investigation of interval weight to date. Using pseudoword production data provided by native Norwegian speakers, Lunden compared the acoustic durations of intervals vs. rimes and found that both correlate with phonological complexity (i.e. vowel length and number of consonants). However, the study did not target the effect of intrinsic duration within intervals of the same phonological size (as in Steriade's *aksa* vs. *akra* example).

A phonetically-grounded theory of intervals makes testable, gradient predictions with respect to Study 4. Namely, the acoustic durations of the VC(C) penultimate intervals should predict the probability of stress assignment. A simple, relatively weak test of Interval Theory can be conducted with reference to the coarse-grained stress assignment results (see Figure 5.19 in Section 5.4.3.2). Recall that the proportion of penult stress followed the *singleton* < *attested* < *unattested* cline. If Interval Theory is the correct model, penult interval durations should at the very least follow the same pattern.

In order to test this prediction, all of the 3,970 error-free productions from Study 4 were re-parsed using the categorical interval model and the durations of the resultant penult intervals were measured. In order to normalize for individual differences in

speech rate, the raw values were divided by whole-word durations to obtain proportions. Because duration is also a correlate of stress, separate proportions were obtained for items coded as having antepenultimate, penultimate, final and ambiguous stress. The results are displayed in Figure 5.29.

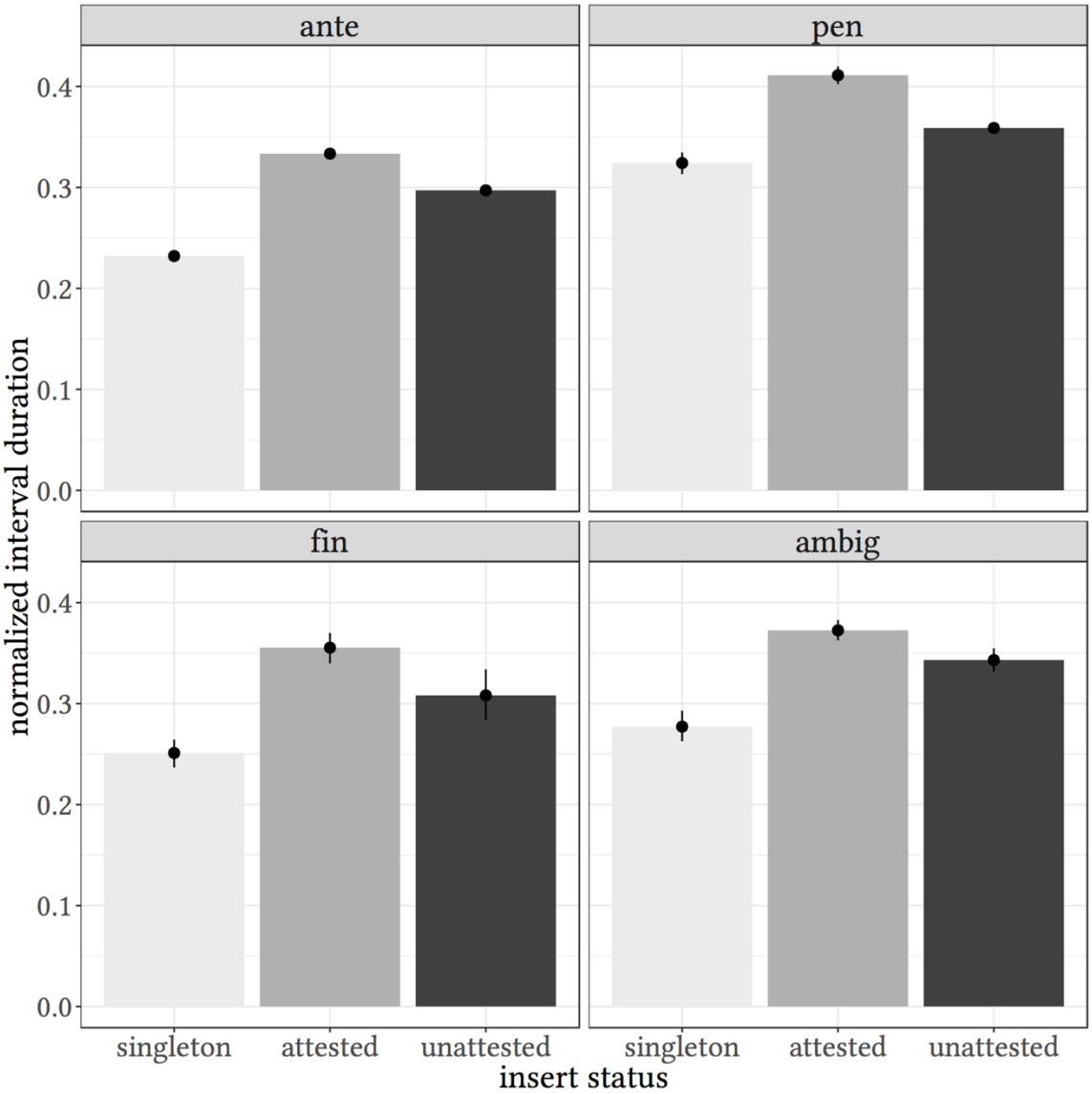

Figure 5.29. Penultimate interval durations as a function of insert status and coded stress. Error bars are confidence intervals obtained via nonparametric bootstrap.

As expected, productions of test items with medial singletons featured shorter penult intervals that those of words with embedded clusters. A series of maximal, mixed-effects linear models (with Helmert-coded insert status) supported this conclusion at each level of coded stress (antepenult: β = -.06 *S.E.* = .002, p < .001; penult: β = -.04 *S.E.* = .004, p < .001; final: β = -.05 *S.E.* = .006, p < .001; ambiguous: β = -.06 *S.E.* = .005, p < .001). Among the cluster-embedded words, however, the figure shows the opposite trend from that seen in Study 4: across stress patterns, productions of pseudowords with initially attested medial clusters featured *longer* penult intervals than productions of items with unattested CC inserts. This pattern was likewise supported across the board by mixed-effects regressions (antepenult: β = .015, *S.E.* = .002, p < .001; penult: β = .022, *S.E.* = .002, p < .001; final: β = .025 *S.E.* = .006, p < .001; ambiguous: β = .01 *S.E.* = .004, p < .05). These duration measures incorrectly predict that attested clusters should attract more penult stress than unattested clusters.

Overall then, neither variant of Interval Theory is able to account for the stress assignment behavior observed in Study 4. On the one hand, an abstract version which employs a weight hierarchy based purely on phonological complexity lacks sufficient granularity to differentiate among $VC_{[obstr]}C$ intervals. On the other, the more fine-grained, phonetically-grounded variant of the theory makes incorrect predictions about items with embedded clusters. Indeed, this version was outperformed by the coarse-grained syllable parser and thus failed a relatively weak test.

5.4.4.1.3 Stress Without Syllables

Recall from section 5.1 that I have assumed the position that stress placement in pseudowords is *multiply determined*, with a number of generalizations being probabilistically extended from the lexicon to conspire (or compete) in determining stress location (see also chapter VIII). Because my focus has been on one family of generalizations – those related to word-edge phonotactics and the consequent syllable structure – the strategy has been to control for the others through design decisions (section 3.3) or by including them in the models as 'nuisance' predictors (section 3.4.2). As noted throughout this dissertation, one important strategy available to participants is analogical processing – extending the stress patterns of lexical neighbors to unfamiliar words. Recall for example the discussion of Baker & Smith, 1978 and Guion et al., 2003 in sections 3.4.2 and 5.1 – the findings of these studies motivated the inclusion of mean edit distance (as a measure of analogy) in the stress assignment models. In this section, I pursue a different measure of analogy, one that is localized to the medial clusters.

One important property of the unweighted edit distance measure used in the models of Studies 3 and 4 is that it assumes *lazy learning*. A learner based on edit distance does not privilege one part of the word over another: a difference found at the beginning of two strings counts the same as a difference in the middle or one at the end. This view of similarity is likely an an oversimplification: research has shown that linguistic creativity often involves task-specific weighting of sub-lexical features. For example, when attaching an English plural to a novel word, native speakers are more sensitive to the final consonant than to the rest of the word (Albright & Hayes, 2003).

Conversely, initial segments might be more important in prefixation (see Kapatsinski, 2014 for discussion of these issues).

In this section, I address the possibility that learners of English stress pay particular attention to the identity of the consonant(s) between the antepenult and penult vowels, and stress novel forms based on this generalization. Informally, the learning generalization can be expressed as follows:

> When the intervocalic insert is *ab*, stress the penult with probability *P*; when the intervocalic insert is *cd*, stress the penult with probability *Q*, etc.

Note that this generalization says nothing about syllables — in fact, it does not presuppose a metrical parse at all. Rather, it involves identifying and selectively attending to a particular position within a word for the purposes of stress assignment. It is in fact reminiscent of the strong/weak cluster distinction made in *SPE* (Chomsky & Halle, 1968), which also eschewed syllables. Unlike edit distance, which can be treated as independent of syllable structure (and indeed, it has been in the models), this generalization is directly in conflict with phonotactics: it is not the word-edge statistics of the inserts that matter, but rather their direct relationship with stress in the lexicon. For this reason, rather than adding the generalization (termed *insert ID* below) to the multivariate model alongside the phonotactic predictors, a separate model featuring it was constructed and compared with the gradient phonotactic parser.

In order to quantify the lexical basis for the *insert ID* generalization, I once again relied on the lexicon of trisyllabic and longer word forms. To match the relevant properties of the test probes, the lexicon was restricted to exclude words with long penult vowels. Furthermore, only words with the same C and CC inserts present in the

stimuli were kept (the overlap amounted to 61 inserts[11]). Figure 5.30 plots the correlation between penult stress in the lexicon and in the pseudowords, aggregated by insert.

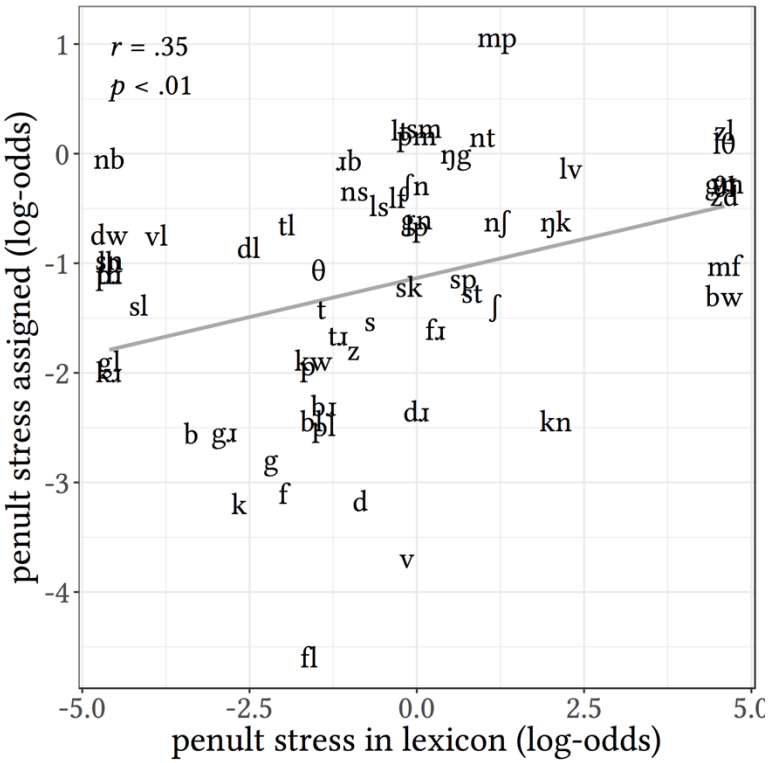

Figure 5.30. The *insert ID* generalization: correlation of penult stress assigned in Study 4 by penult stress in the lexicon, aggregated by the 61 shared (C)C inserts.

The correlation is significant and positive: the more often an insert is paired with penult stress in the lexicon, the more likely it was to trigger penult stress in the pseudowords. This insert-level generalization accounted for about 12% of the variance in the aggregated responses.

[11] In the 'all lemmas' and 'simplex lemmas' lexicons, the number of shared inserts was reduced to 56 and 42, respectively. This low degree of overlap with the stimuli motivated the decision to base the analysis on the word form lexicon.

In order to compare this account to the gradient parsing model, a maximal, mixed-effects logistic regression was fit to the stress assignment data. The probability of stressing the penult of a pseudoword was modeled as a function of the probability of its insert being paired with penult stress in the lexicon (*insert ID*). To facilitate the comparison, the model also featured edit distance and embedded word bias as nuisance predictors. All predictors were centered and scaled. The output is shown in Table 5.6.

Table 5.6. Insert-tracking model output, stress assignment task.

	Estimate (Std. Error)
Intercept	-2.529 (0.356)***
Insert ID	1.558 (0.515)**
Edit distance bias	-0.350 (0.188)
Embedded word bias	0.232 (0.189)
Observations	2,904
Log Likelihood	-1,218.707
Bayesian Inf. Crit.	2,628.787
Note:	*p<0.05; **p<0.01; ***p<0.001

As seen in the Table, insert-level stress in the lexicon significantly predicted stress assignment: with each unit increase in the predictor, the odds of stressing a pseudoword increased by a factor of 4.75. Neither edit distance nor embedded word bias were found to make significant, independent contributions to stress.

In order to facilitate a fair comparison, the gradient parsing model was refit to the same data set (i.e. to the pseudowords with the 61 inserts shared by the lexicon). The model's output is shown in Table 5.7.

Qualitatively, the gradient parser performed similarly on the reduced data set as it did on the full data set (cf. Table 5.5). Word onset frequency remained significant; with each standardized unit increase on the measure, the odds of stressing the penult

decreased by a factor of .38. A significant but smaller effect was also found for edit distance: as the similarity to antepenult-stressed words increased by one *z*-score, the odds of stressing the penult decreased by a factor of .76. Unlike in the full data however, the model also returned a significant effect of embedded words: with each standard unit increase in penult bias, the odds of penult stress rose by a factor of 1.32.

Table 5.7. Output of gradient parsing model fit to the same data as insert-tracking model.

	Estimate (Std. Error)
Intercept	-1.739 (0.330)***
Word Onset Frequency	-0.971 (0.131)***
Word Offset Frequency	-0.026 (0.167)
Sonority Slope	-0.003 (0.119)
Edit Distance Bias	-0.277 (0.114)*
Embedded Words Bias	0.281 (0.139)*
Observations	2,904
Log Likelihood	-1,118.282
Bayesian Inf. Crit.	2,619.309
Note:	*p<0.05; **p<0.01; ***p<0.001

The model comparison followed the same procedure adopted throughout this dissertation. First, the predictions of each model were aggregated by insert and correlated with the observed values. The scatterplots are presented in Figure 5.31.

As evident in the figure, the gradient parsing model outperformed the insert-tracking model. The mean squared deviation of the former was half that of the latter, with double the explained variance. Overall, the predictions of the insert-tracking model were distributed over a more restricted range of values. Furthermore, it appears that categorical word-initial phonotactics are not strongly paralleled by word-medial stress in the lexicon (as evidenced by the considerable overlap in predictions for the three insert types in the left panel).

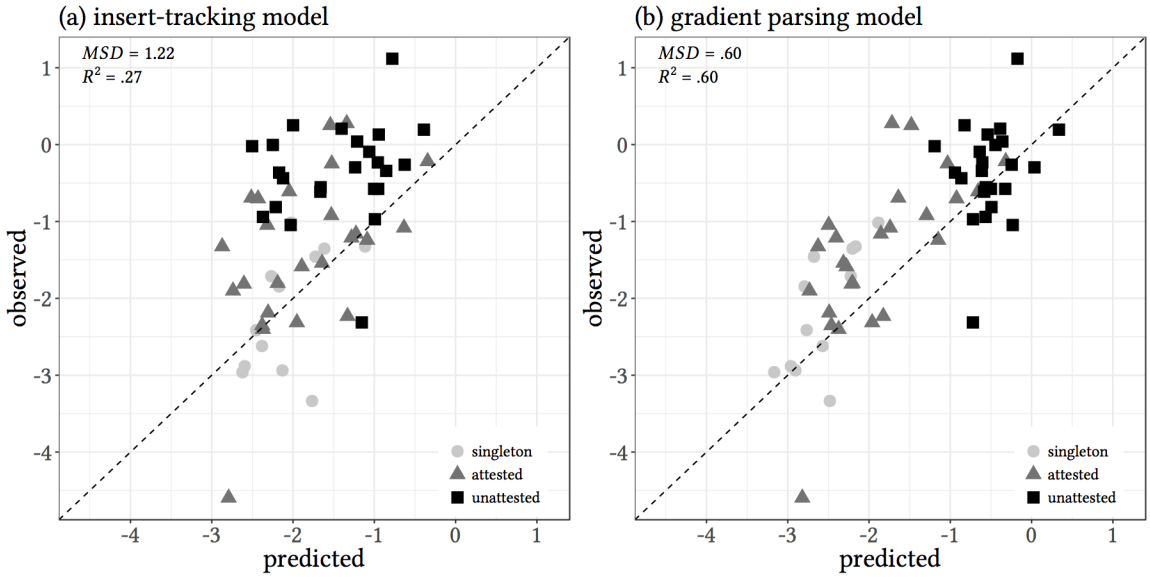

Figure 5.31. Comparison of insert-tracking vs. gradient parsing model predictions (stress assignment data). Values are in log-odds.

The second part of the comparison relied on *BIC* scores to guard against overfitting. As seen in Tables 5.6 and 5.7, the *BIC* score of the gradient parsing model was lower by 9.48 points. This translated to a Bayes Factor of 114.29, which in turn yielded a posterior probability of .991. In other words, given the choice of both models, an unbiased learner would almost always infer the gradient parser from the data.

5.5 Study 5: Production Accuracy

5.5.1 Overview

The probabilistic nature of the metrical parse has consequences not only for stress assignment, but potentially also for production accuracy. Consider again the pseudoword *vatabnick*: the cluster [bn] is initially unattested but nevertheless has a

non-zero probability of being syllabified as a complex onset, yielding the metrical parse [('væ.tə.)<bnɪk>]. Does such tautosyllabic treatment make this word more difficult to produce than splitting the cluster, as in [və.('tæb.)<nɪk>]? What about *vatablick*, which features an embedded, high-frequency word onset, or *vatadwick*, which contains a rare one? It is well-known that unfamiliar *word* onsets are prone to production errors (e.g. Davidson, 2006). Would the same hold for medial *syllable* onsets? Would production accuracy on the latter be a probabilistic function of their gradient well-formedness in onset position?

Prior research on speech errors induced in laboratory settings has found that lexical support is indeed implicated in production accuracy; for instance, target word and phoneme frequencies are inversely correlated with the probability of committing an error (Dell, 1990; Kupin, 1982; Levitt & Healey, 1985). At the same time, not all statistical asymmetries are reflected in error rates. For example, Davidson (2006) found that errors on novel CC word onsets were not significantly related to position-independent frequencies of these clusters in the English lexicon. This suggests that phonotactic well-formedness may reference syllable structure rather than purely sequential dependencies (contra Blevins, 2003; Steriade, 1999). In other words, there is reason to hypothesize that medial clusters which are parsed as syllable onsets might be subject to the same onset effects on production which have been observed at word edges.

In Study 5, I examine the extent to which production accuracy in Study 4 paralleled stress assignment, preferences and hyphenation in providing converging evidence for gradient well-formedness of the medial clusters. Specifically, I ask what happens to accuracy when the metrical parse treats the clusters as complex onsets to

the final syllable. The hypothesis is that, if the same phonotactic well-formedness cline subserves both syllabification and ease of articulation of onsets, then the probability of committing an error on pseudowords with antepenult stress (which indicates the tautosyllabic parse) should be better captured by the gradient than by the categorical phonotactic model (i.e. it should be predicted by the same lexical support measures as syllabification). Crucially, this prediction does not hold for penult-stressed items: since the medial cluster is split by this metrical parse, it should not be subject to onset-specific production constraints.

5.5.2 Typology of the Speech Errors

There were 955 total production errors committed by the participants, constituting just under 19% of the total trials. The errors were of several kinds, including epenthesis, substitutions, deletions and pauses. Table 5.8 provides a breakdown of errors by type.

Table 5.8. Typology of production errors in the stress assignment task.

Error Type	Example			Count (%)
deletion (insert C)	*tamapmish*	→	tamapish	104 (10.9)
deletion (V)	*tamapish*	→	tampish	6 (0.6)
deletion (other)	*lidigmeph*	→	ligmeph	1 (0.1)
epenthesis (insert C)	*sipalbesh*	→	sipalblesh	62 (6.5)
epenthesis (V)	*sipalbesh*	→	sipaləbesh	103 (10.8)
epenthesis (other)	*sanankep*	→	sansankep	32 (3.4)
metathesis (insert CC)	*sipalbesh*	→	sipablesh	43 (4.5)
metathesis (other)	*nepantep*	→	neptanep	50 (5.2)
substitution (insert C)	*zepazriss*	→	zepadriss	60 (6.3)
pause	*zepazriss*	→	zepaz…riss	257 (26.9)
multiple	*zepazriss*	→	zepalidrilis	200 (20.1)
null response	*zepazriss*	→	…	37 (3.9)
TOTAL				955 (100)

By far, the largest proportion of errors fell into the 'pause' and 'multiple' categories. The former consisted of cases where participants would fail to produce an item under a unified prosodic contour, inserting one or more pauses into the middle of the word. The 'multiple' category consisted of productions that deviated from the expected output by more than one error. For example, a pause could be inserted and an extra syllable added in the same production. The remaining errors were distributed among various deletions, insertions, substitutions and metatheses.

In what follows, rather than restricting the analysis only to the obvious, 'classic' phonotactic repairs (deletion, epenthesis, etc.), all errors are considered together. The reason for this was two-fold. First, initially-attested inserts are already well-formed, and so strictly speaking, they cannot be repaired. An analysis of obvious repairs would have to exclude these items and thus be unable to model word-edge statistics (since these do not vary in unattested onsets). Second, repairs can manifest in ways other than segmental rearrangement. For example, the most common location for a pause by far was between the penultimate and final syllables. This error essentially repaired the pseudowords by turning each into two shorter ones, a disyllable followed by a monosyllable. Ignoring such pauses might therefore overlook an important insight.

5.5.3 Results

5.5.3.1 Coarse-Grained Phonotactics

In this section, I analyze how coarse-grained phonotactics interact with stress in predicting the likelihood of a speech error. Of the 1,137 attempts at penult stress, 270 (23.7%) resulted in errors. In contrast, out of 2,944 attempts at antepenult stress, the number of errors was 474 (16.1%). In other words, trying to stress the penult resulted in a higher overall probability of committing an error relative to trying to stress the antepenult. Figure 5.32 reveals how these probabilities were further modulated by insert type. The left panel displays the phonotactic effect within antepenult-stressed items. Again, under this stress pattern, the inserts are assumed to be parsed as onsets to the final syllable. Note that the pattern of errors appears to reflect well-formedness effects observed in word-initial position: singleton onsets are relatively easy to produce (5.6% errors), attested clusters somewhat less so (11.1% errors), and unattested clusters appear to be markedly more difficult than the others (30.1% errors). The situation is quite different among penult-stressed items, where the C1 of each insert is assumed to close the penult coda. Here, the differences are less pronounced, and the numerical trend is in the opposite direction. Singleton items are the *most* likely to be mispronounced (27.9%), followed by attested clusters (26.5%) and unattested clusters (21.4%).

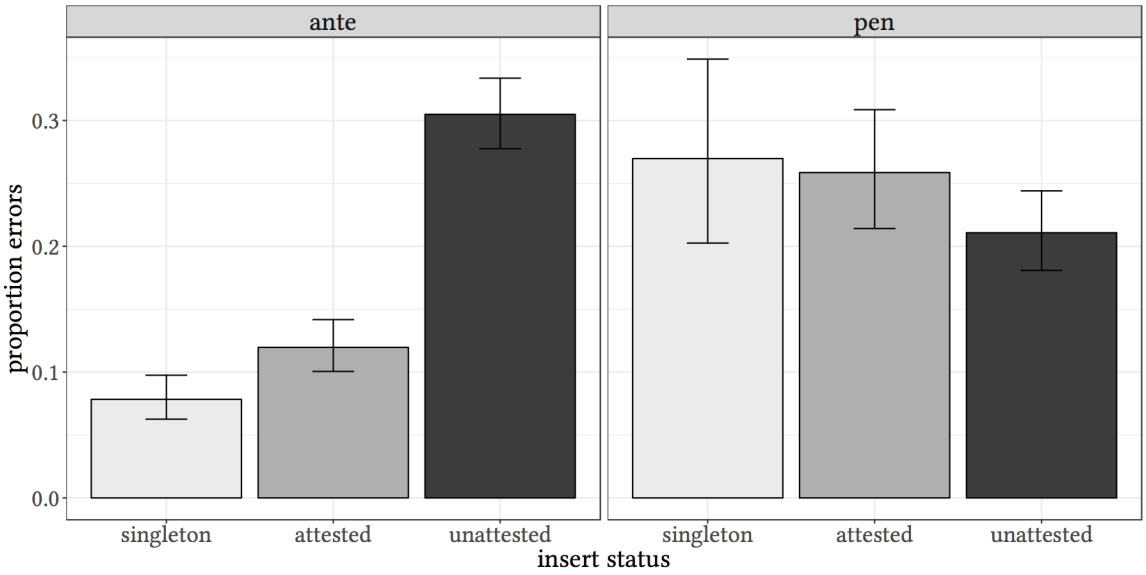

Figure 5.32. Proportion of speech errors by insert type and stress pattern.

To test the significance these patterns, a maximal, mixed-effects logistic regressions was fit to the data. The model contained the main effects of stress and insert status as well as the interaction between these predictors. This model significantly improved fit over a main effects-only version ($\chi^2(24) = 104.61$, $p < .001$), which in turn outperformed a null model ($\chi^2(3) = 50.85$, $p < .001$). These results indicate that the effects of stress pattern and insert type depended on each other in predicting errors.

In order to explore this interaction, a number of follow-up models investigated simple effects and contrasts. First, the effect of insert status was investigated separately for each level of stress. The output of these two models is listed together in Table 5.9.

The model predicting antepenult-stressed errors significantly outperformed the null hypothesis ($\chi^2(2) = 58.91$, $p < .001$). Relative to singletons, the odds of mispronouncing attested and unattested items were significantly higher by factors of 2.12 and 12.43, respectively. A follow-up comparison further revealed a significant

difference between the two cluster types (β = 1.73, *S.E.* = .24, p < .001), with the odds ratio of erring on unattested items higher by a factor of 5.62.

Table 5.9. Categorical models within stress levels.

	Estimate (Std. Error)	
	ante model	pen model
Intercept (Status = singleton)	-3.366 (0.321)***	-1.433 (0.371)***
Status = attested	0.752 (0.294)*	0.051 (0.356)
Status = unattested	2.520 (0.265)***	-0.313 (0.349)
Observations	2,944	1,137
Log Likelihood	-1,054.523	-565.733
Bayesian Inf. Crit.	2,228.860	1,237.009
Note:	*p<0.05; **p<0.01; ***p<0.001	

In contrast, the model predicting the penult-stressed errors failed to significantly improve fit over an intercept-only model ($\chi^2(2) = 1.81$, $p = .40$), offering no evidence that insert status had an impact on the production accuracy of these items.

The second set of comparisons predicted the effect of stress at each level of insert type. The output of the three models is listed in Table 5.10.

Table 5.10. Coarse models within insert status.

	Estimate (Std. Error)		
	singleton model	attested CC model	unattested CC model
Intercept (Stress = antepenult)	-3.480 (0.346)***	-2.647 (0.330)***	-0.840 (0.181)***
Stress = penult	2.121 (0.395)***	1.228 (0.307)***	-1.057 (0.234)***
Observations	1,074	1,309	1,698
Log Likelihood	-265.525	-477.331	-896.349
Bayesian Inf. Crit.	586.884	1,012.078	1,852.196
Note:	*p<0.05; **p<0.01; ***p<0.001		

In all three models, the effect of stress was statistically significant. For singletons and attested clusters, committing an error was more likely with penult- than antepenult-stressed items (the error odds were higher by a factor of 8.34 in singletons and 3.41 in attested word onsets). In contrast, words with initially unattested clusters were less likely to be mispronounced when paired with penult stress than with antepenult stress (odds ratio = .35).

Taken together, these results suggest a number of conclusions. First, when the stress pattern points to a tautosyllabic onset parse of the medial inserts, these inserts behave like word onsets. That is, their production accuracy in medial position depends on their relative well-formedness in word-initial position, with legal word onsets easier to pronounce than illegal word onsets. On the other hand, when stress suggests a closed penult parse, the inserts no longer behave like word onsets, with production accuracy being independent of word-initial status. Interestingly, 'splitting' the clusters with the metrical parse did not make them easier to pronounce across the board: legal word onsets (both singletons and clusters) suffered when paired with penult stress, suggesting that the closed-penult parse of these items is dispreferred by the production system.

To sum up, much like the hyphenation and stress assignment results, the error rates provide converging evidence for the gradient well-formedness of the medial clusters. I now turn to the question of whether production accuracy is also sensitive to more fine-grained onset phonotactics.

5.5.3.2 Fine-Grained Phonotactics

In this section, I analyze how fine-grained phonotactics interact with stress in predicting the likelihood of a speech error. I begin by examining the correlations between each gradient predictor and the insert-level error probabilities, separately for each stress pattern. The data for word onset frequency are plotted in Figure 5.33. As in the other studies, illegal word onsets were excluded in order to facilitate a more stringent test of frequency. Each panel contains the same 40 data points (12 singletons, 28 attested word onsets). The left panel shows the onset frequency effect when stress fell on the antepenultimate syllable. The correlation is negative and significant, with frequent word onsets leading to fewer errors when syllabified as such. Word onset frequency captured about 23% of the variance in the aggregated errors. The right panel plots the data for penult-stressed errors. Here, the error rates, though higher overall, appear to be independent of onset frequency.

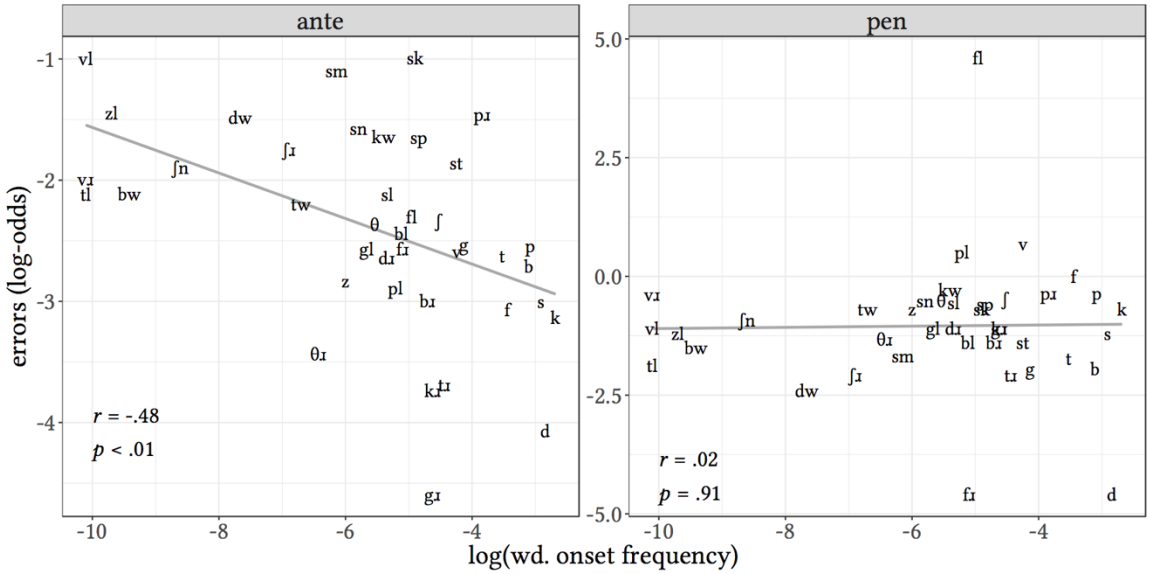

Figure 5.33. Log-odds of production errors by stress and word-onset frequency of each embedded insert (singletons, attested CC onsets).

Follow-up models explored the onset frequency effect among singletons and attested items at each level of stress. For antepenult-stressed items, onset frequency was significant (β = -.49 *S.E.* = .10, p < .001). As onset frequency increased by one z-score within legal onsets, the log-odds of committing an error decreased by a factor of .61. This effect persisted even after the removal of marginal word onsets (β = -.40 *S.E.* = .14, p < .01). For penult-stressed items, the onset frequency effect failed to reach significance (β = .11 *S.E.* = .15, p = .43).

Figure 5.34 plots the interaction between word offset frequency and stress, with the data averaged within the initial consonant of each insert. The left panel, which plots the antepenult-stressed errors, shows a positive and significant between offset frequency and error rates: the more likely a consonant is encountered as a word offset, the more likely placing it in the syllable onset (via stress) resulted in a production error. Offset frequency accounted for about 25% of the variance in the aggregated errors. For penult-stressed items, the correlation was in the opposite direction: the more likely a consonant was word-finally, the less likely parsing it in the coda (via stress) led to a mispronunciation. However, this correlation failed to reach significance.

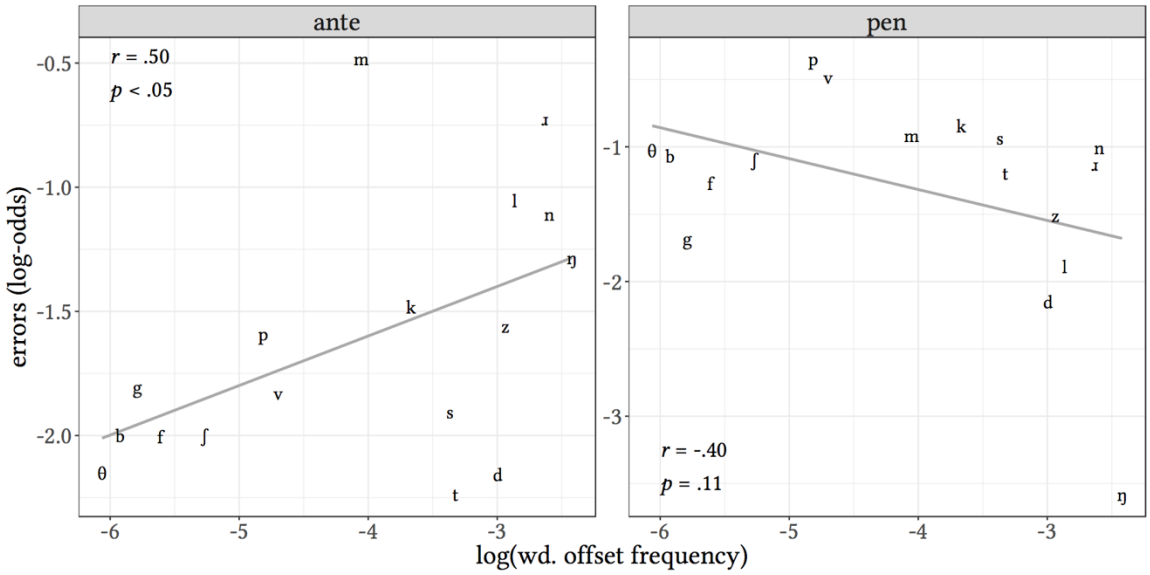

Figure 5.34. Log-odds of production errors by stress and word-onset frequency of the C1 of each embedded insert.

The correlations were explored with mixed-effects models predicting errors by offset frequency at each level of stress. For antepenult-stressed errors, the offset frequency effect was significant (β = .41 *S.E.* = .12, p < .001). With each standard unit increase in offset frequency, the odds of committing an error increased by a factor of 1.50. The effect persisted even after /ŋ/ was removed from the data (β = .41 *S.E.* = .12, p < .001), indicating that the effect was not driven by the categorical prohibition against this segment in onset position. For penult-stressed errors, the frequency effect was also significant (β = -.23 *S.E.* = .10, p < .05). The effect was in the opposite direction from the antepenult-stressed errors: as offset frequency increased, the error odds decreased by a factor of .80. However, the effect was no longer significant after /ŋ/ was removed from the data (β = -.17 *S.E.* = .10, p = .076), indicating that some of the effect was driven by the very low error rates observed when penultimate stress syllabified this segment into the coda.

The interaction of stress and sonority in predicting aggregate errors in items with unattested word onsets is plotted in Figure 5.35. In both panels, the relationship is in the positive direction, with rising sonority leading to more errors than falling sonority regardless of stress. This is surprising from the perspective of the SSP, which predicts a negative correlation for the antepenult-stressed items. That said, neither correlation reached statistical significance.

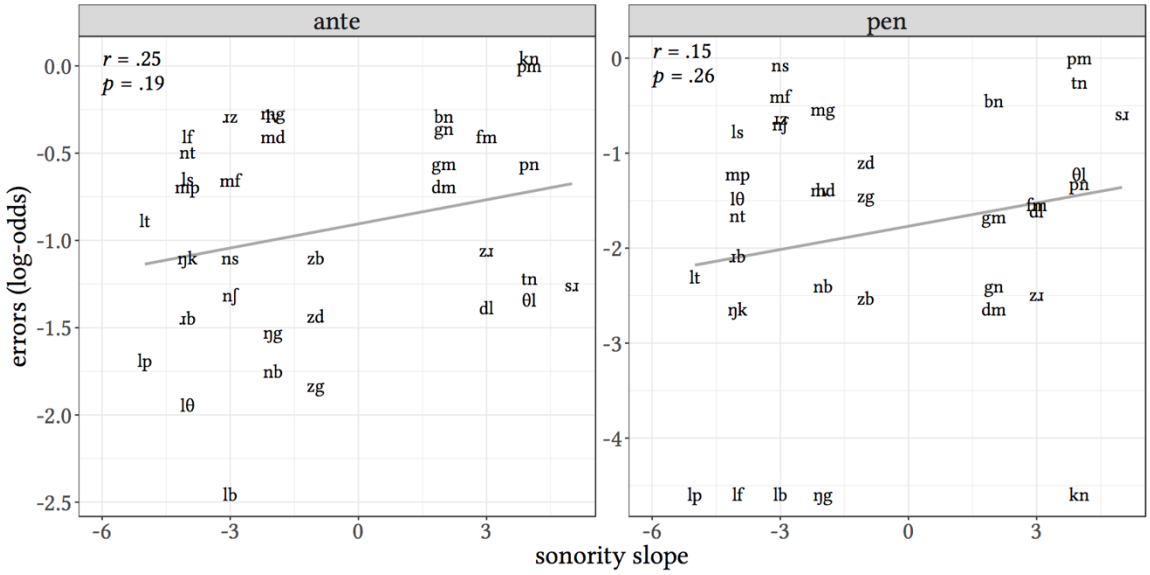

Figure 5.35. Log-odds of production errors by stress and sonority slope of each embedded insert (unattested CC onsets).

Mixed-effects regression models fit to the raw observations of unattested clusters supported the conclusions suggested by the correlations. The effect of sonority failed to reach significance for both antepenult-stressed items (β = .13 *S.E.* = .13, p = .34) as well as penult-stressed items (β = .16 *S.E.* = .24, p = .49). In other words, there was no evidence that, for pseudowords with medially-embedded illegal word onsets, the probability of committing a speech error was dependent on the sonority profile of the insert.

The joint influence of the gradient predictors and stress on error rates was tested in a multiple regression model fit to the entire set of observations. The model included main effects of onset frequency, offset frequency, residualized sonority and stress, as well as the two-way interactions between stress and each of the three phonotactic predictors. The maximal model converged and significantly outperformed the main-effects only version ($\chi^2(39) = 119.6$, $p < .001$), which in turn performed significantly above the intercept model ($\chi^2(3) = 52.51$, $p < .001$). These results indicate that at least some of the phonotactic predictors depended on stress in predicting errors.

Two follow-up models investigated the significant interaction by examining the effects of the predictors at each level of stress. The results of both models are listed in Table 5.11.

Table 5.11. Gradient models within stress levels.

	Estimate (Std. Error)	
	ante model	pen model
Intercept	-2.369 (0.261)***	-1.586 (0.245)***
Word Onset Frequency	-1.169 (0.126)***	0.117 (0.113)
Word Offset Frequency	-0.035 (0.125)	-0.179 (0.119)
Sonority Slope	-0.002 (0.094)	0.036 (0.152)
Observations	2,944	1,137
Log Likelihood	-1,035.517	-553.452
Bayesian Inf. Crit.	2,262.735	1,275.772
Note:	*p<0.05; **p<0.01; ***p<0.001	

The model predicting antepenult-stressed errors significantly outperformed the null hypothesis ($\chi^2(3) = 56.68$, $p < .001$). As seen in the table, only word onset frequency was significantly associated with error rates on these items. For each standard unit

increase in word onset frequency of the insert, the odds of mispronouncing an antepenult-stressed item decreased by a factor of .31. Neither offset frequency nor sonority contributed to error rates. For the penult-stressed items, the model failed to improve fit over the intercept-only model ($\chi^2(3) = 4.13$, $p = .25$), with none of the predictors emerging as significant.

Taken together, these results support the conclusion that fluency probabilistically benefits when the metrical parse places well-formed onsets in onset position, with well-formedness defined on the same cline as in Studies 1-4. Specifically, when antepenult stress indicates a .(C)C parse, frequent word onsets are produced with fewer errors than rare word onsets. This fluency advantage disappears when the speaker uses penult stress, indicating C.(C) syllabification of the insert.

5.5.3.3 Model Comparison

The analysis in the preceding two sections revealed that, when inserts are parsed as medial onsets, both coarse-grained and fine-grained phonotactic generalizations affect the probability of producing errors. This section directly compares the ability of these two types of generalizations to account for the data. Because phonotactics did not affect errors in penult-stressed items, the comparison is restricted to trials where antepenult stress was attempted.

Three phonotactic models were compared. The first was the categorical model containing insert status and maximal random effects (see left column of Table 5.9 in section 5.5.3.1 for model output). The second was the gradient model containing all three gradient predictors (Table 5.11, section 5.5.3.2, left column). Finally, a reduced

version of the gradient model that excluded the sonority predictor was also added to the comparison. The rationale for including this model was as follows. Recall from the preceding section that, on its own, sonority was unable to account for the error rates in antepenult-stressed items containing unattested items (in contrast, both onset frequency and offset frequency were significant when fit to their respective data subsets). In other words, there is reason to believe that sonority slope is a spurious predictor that would add unnecessary complexity to the gradient model. For this reason, the reduced gradient model contained only the two frequency measures. Indeed, dropping sonority from the gradient model revealed no significant difference in fit ($\chi^2(1) = .002$, $p = .97$), indicating that sonority slope was not a good predictor. The output of the reduced model is shown in Table 5.12. As in the full gradient model, only onset frequency emerged as a significant predictor. Furthermore, the effect size was unchanged from that in the full model: with each standard unit increase in onset frequency the odds of committing a speech error decreased by a factor of .31.

Table 5.12. Reduced gradient model, antepenult-stressed errors.

	Estimate (Std. Error)
Intercept	-2.315 (0.250)***
Word Onset Frequency	-1.163 (0.126)***
Word Offset Frequency	-0.031 (0.122)
Observations	2,944
Log Likelihood	-1,049.277
Bayesian Inf. Crit.	2,218.367
Note:	*p<0.05; **p<0.01; ***p<0.001

The predictive performance of the three models is visualized in Figure 5.36, which plots the aggregate observations against each model's predictions and lists the mean squared deviations and R^2 values.

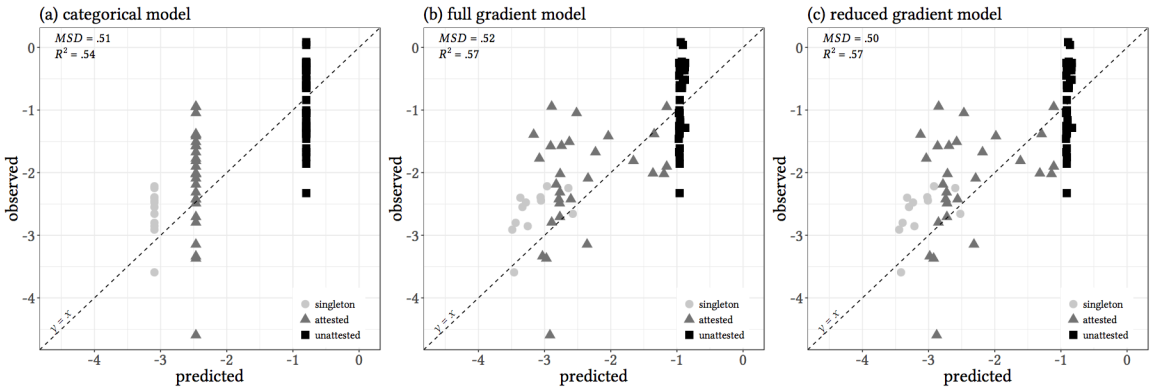

Figure 5.36. Comparison of model predictions (production error data). Values are in log-odds.

Relative to the categorical model, the predictions of the two gradient models are more distributed, albeit only among the singleton and attested items. In other words, all three models largely agreed on their predictions of unattested items (for the gradient models, this confirms that word onset frequency is doing the bulk of the work). Panels (b) and (c) are virtually identical; this is an expected result given that the removal of sonority was largely undetected by the likelihood ratio test. Overall, the gradient models appear to have a slight edge over the categorical model, but the differences in R^2 values are so small that aggregate predictions alone cannot adjudicate between the three models.

A comparison of the BIC scores somewhat clarifies the picture. The scores were as follows: categorical model, 2,229; full gradient model, 2,263; reduced gradient model, 2,218. Thus, the categorical model had an advantage over the full gradient model which translated to a Bayes Factor of over 2.2×10^7 and a posterior probability of essentially 1. However, the smallest score was featured by the reduced gradient model. Compared to

the categorical model, the Bayes Factor was about 190, indicating a posterior probability of .995.

To sum up, under the assumption that learners balance predictive power with complexity, the data support fine-grained sensitivity to frequency-driven (but not sonority-driven) phonotactics.

5.5.4 Discussion

Study 5 revealed that production accuracy was sensitive to the same gradient phonotactics as syllabification. Specifically, 'bad' medial clusters lead to more errors than 'good' medial clusters, but only when antepenultimate stress indicated that the metrical parse placed them into complex onsets. Crucially, the notions 'good' and 'bad' lay on a continuum captured by onset frequency. In other words, production errors provided evidence that the same well-formedness cline that effects syllabification also affects production accuracy. This result bolsters the argument that phonotactic knowledge is gradient and demonstrates that the same knowledge is implicated in a number of diverse linguistic behaviors.

This result was somewhat different from that reported in Davidson (2006), where error rates in the production of novel CC word onsets were not predicted by frequencies of the clusters in other positions. It appears that the relationship between initial and medial phonotactics is asymmetrical: word-initial well-formedness transfers to medial onsets, but the reverse does not hold.

CHAPTER VI

CORRELATING THE RESULTS

6.1 Overview

As discussed throughout the previous two chapters, the results of Studies 1-5 point to the conclusion that the metrical parse follows gradient phonotactic knowledge. In section 5.4.4, I discussed some differences in task sensitivity across the studies and offered a few explanations for why the base rate of closed penults differed between hyphenation and stress assignment. In this chapter, I focus on the similarities by comparing the insert-level and item-level responses across the experiments. This procedure will shed further light on the extent to which the behavior observed in these studies was dependent on the task.

Comparing the responses across different experimental paradigms is especially illuminating for two reasons. First, Studies 1, 4 and 5 (and to a lesser extent Study 3) employed the same test stimuli. The extent to which the stress and error results parallel hyphenation on an item-by-item (or cluster-by-cluster) basis can thus serve as strong evidence that both processes tapped into the same kind of knowledge. Second, because Study 2 used very different test items, we can observe how robust the insert-level phonotactic generalizations are in different environments. Significant pairwise correlations among the results are by no means a foregone conclusion. Working with native Russian speakers, Côté & Kharlamov (2011) used the same set of nonword stimuli to examine five different syllabification tasks: first-syllable repetition, second-

syllable repetition, pause insertion, hyphenation and a Likert-scale rating of alternative parses. Fewer than half of the pairwise correlations were statistically significant. For example, the results of first-syllable repetition did not correlate significantly with those of any other task, yielding more closed syllable responses than the others (the authors interpreted this as a word minimality bias, see section 4.1). The notion that stress assignment will correlate with hyphenation cannot be taken for granted.

6.2 Results and Discussion

Figure 6.1 plots the correlation matrix of the responses in Studies 1-5, aggregated by insert. The values are in log-odds, with positive numbers indicating higher than 50% rates of closed penults, as observed directly (hyphenation study and Eddington et al. reanalysis) or inferred indirectly (stress preference, production and error studies). The error data are for productions with antepenultimate stress. In addition to these responses, the matrix includes one additional, relevant data set: the well-formedness judgments of word-initial CC clusters reported in Scholes (1966, Experiment 5). In that study, 33 seventh-graders rated the grammaticality of nonsense monosyllables featuring 66 unique CC onsets. The results are included here because this data set has been used as a test case for a number of recent, influential models of phonotactic learning (Albright, 2009; Daland et al., 2011; Hayes & Wilson, 2008).

Each scatterplot in the lower triangle of the matrix is fitted with a smoother. The upper triangle shows Pearson's coefficients along with the significance levels. Recall that the hyphenation, binary preference, stress assignment, and error datasets

featured the same 75 unique inserts, while the Eddington et al. (2013a,b) study shared 67 inserts with these studies. In contrast, the Scholes study only had 25 word onsets in common with the insert pool. Of these, none were singletons and only 3 were unattested (fm, sɪ, zɪ); the remaining 22 clusters were legal word onsets.

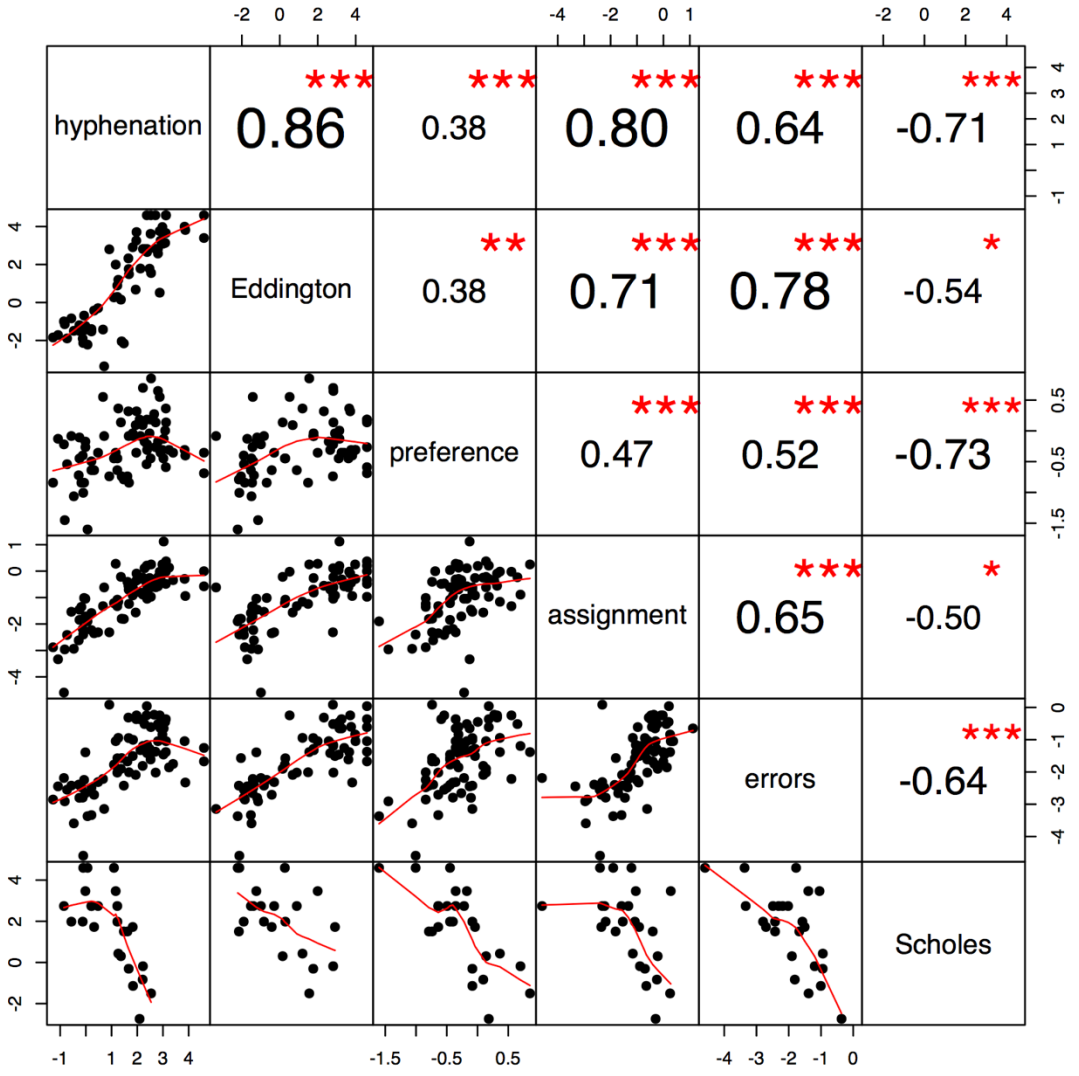

Figure 6.1. Correlation matrix of the responses in Studies 1-4, production errors in Study 4, and Scholes (1966) well-formedness judgments. The data are aggregated by insert and converted to log-odds.

Several conclusions can be drawn from the correlation matrix. The first, and most obvious, is that the Scholes (1966) judgments correlate negatively with all other data sets. Despite the small number of shared data points, the correlations are significant. At least for the 25 CC clusters shared among these studies, perceived well-formedness in initial position appears to be a good, gradient predictor of medial syllabification — whether metalinguistic or inferred from stress — and of production accuracy. The better a CC cluster is as a word onset, the less likely it is to be split by hyphenation or stress, and the less likely it is to be mispronounced when stress treats it as a medial onset.

The second conspicuous observation is that all of the remaining correlations are positive and statistically significant, suggesting that each task tapped into the same underlying mechanism (i.e. the parse). This is of course expected given the success of the gradient model at predicting each experiment.

The strongest correlation was between the two metalinguistic parsing tasks: the hyphen insertion experiment and the forced-choice syllabification in the Eddington et al. (2013a,b) study shared a remarkable 74% of the variance. This is somewhat surprising, given that, like the Scholes experiment, the Eddington et al. study used entirely different test items. The fact that responses to real, disyllabic words so closely matched the treatment of trisyllabic pseudowords again suggests that the two tasks employed the same underlying mechanism. Importantly, their close correspondence cannot be attributed only to the fact that both tasks were metalinguistic: both studies were strongly correlated with the stress assignment task, which shared 64% of the variance with hyphenation and over 50% with Eddington et al. (2013a,b). In fact, the production study correlated more strongly with these two tasks than it did with the

other stress-based study, the preference task. The fact that stress assignment mirrored hyphenation behavior is the strongest piece of evidence that underlyingly, phonotactic knowledge dictated behavior independently of task-specific effects.

The preference task yielded the weakest correlations with all studies other than Scholes (1966), sharing only 14% of the variance with the hyphenation and Eddington et al. (2013a,b) study, 22% with the stress assignment results, and 27% with the errors. This is also an anticipated result: recall that the responses in Study 3 were less sensitive to phonotactic predictors and generally noiser (note also the reduced range in the preference responses). Furthermore, this task employed only half of the CVCV__VC frames used in the hyphenation and production studies, giving more weight to item-level effects.

Somewhat surprisingly, the preference study featured the strongest correlation with the Scholes data (over 53% shared variance), and antepenult-stressed errors correlated most strongly with the Eddington data (nearly 61% shared variance). These findings were unexpected given that the studies in question did not share the same test items. There is no obvious explanation for the strength of the error:Eddington relationship. As for the preference:Scholes correlation, one possible explanation for this may lie in task effects. Both of these studies asked for well-formedness judgments, whether relative (preference task) or absolute (Scholes); it is thus likely that task similarity played a role in bolstering the correlation. That said, task effects obviously cannot be the whole story: the Scholes data are significantly correlated with the other studies, suggesting common reliance on underlying phonotactics.

Figure 6.2 plots the correlation matrix of the by-item aggregated responses and errors among the four studies that shared pseudowords. Because the preference task

used half of the frames as the other two experiments, there are only 85 points in the scatterplots involving this study.

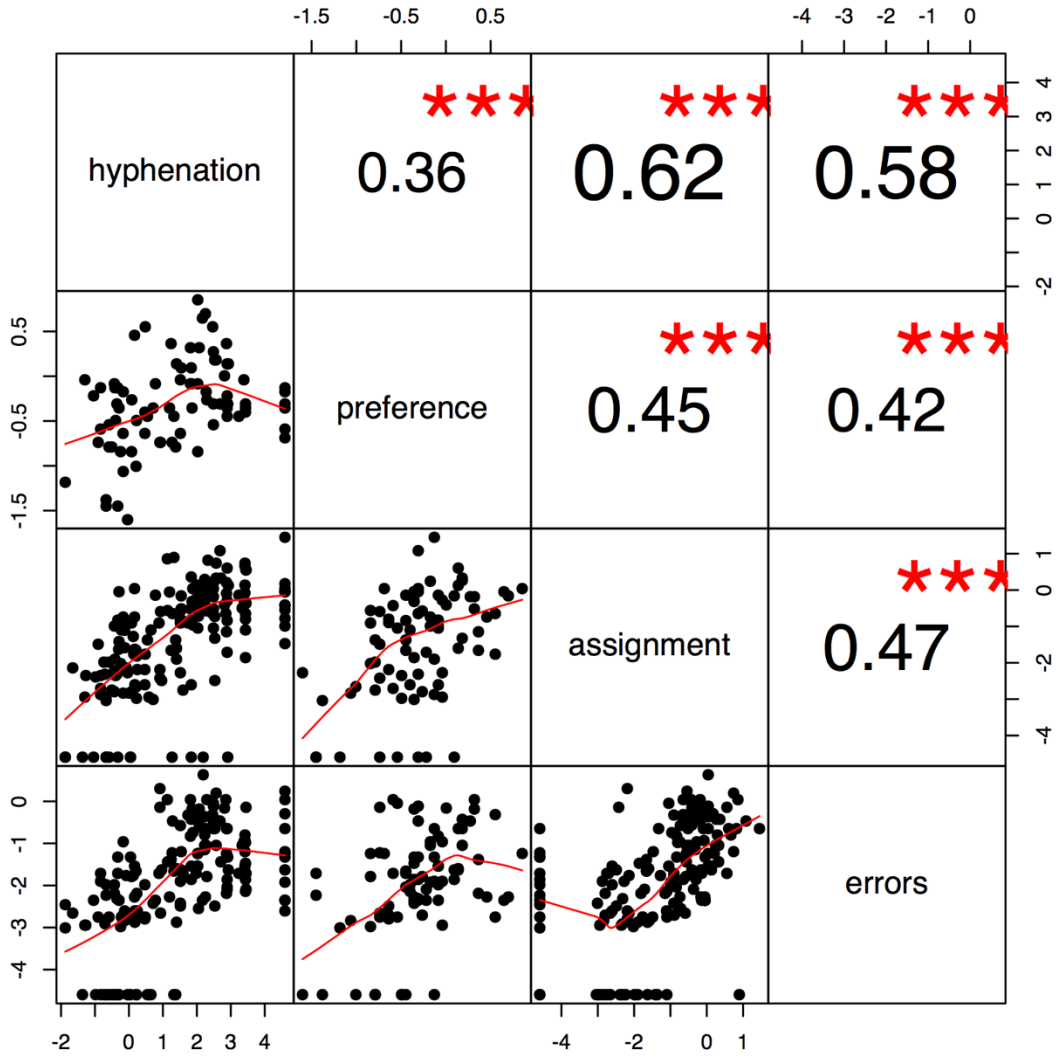

Figure 6.2. Correlation matrix of the response data (in log-odds) across Studies 1-4, aggregated by pseudoword.

As expected, the correlations were somewhat more noisy than those among the by-insert responses. Nevertheless, they remained statistically significant, with the

patterns unchanged. In the strongest correlation, hyphenation and stress assignment shared about 38% of their variance.

Overall, the correlation matrices reinforce the idea that syllabification and production accuracy are influenced by medial onset well-formedness, which is in turn largely driven by gradient word onset phonotactics.

CHAPTER VII

SIMULATIONS

Portions of the work presented in this chapter will be published as a coauthored article: Olejarczuk, P. & Kapatsinski, V. The metrical parse is guided by gradient phonotactics. To appear in *Phonology*.

7.1 Background

The evidence presented in chapters 4-6 converges on the conclusion that fine-grained phonotactic generalizations are involved in determining syllable boundaries in English. This raises the question of why this would be the case. The idea of phonotactic knowledge as being highly detailed is by now largely accepted by phonologists, but why would this level of detail be relevant to the metrical parse? After all, as good as humans are at tracking the statistics of their linguistic environment, some patterns and dependencies appear to go unnoticed (e.g. Becker et al., 2011).

From the functional perspective, it would seem that the categorical parsing grammar would be preferable, for at least three reasons. First, such a grammar is much simpler and thus may hold an advantage in acquisition. Recent work on language learning in laboratory settings strongly suggests that formal simplicity correlates with learnability: phonological patterns are easy to learn to the extent that they can be expressed with elegant notational mechanisms (Moreton & Pater, 2012). Second, a model that yields deterministic syllable boundaries would facilitate efficient

phonological processing because accumulating frequency information over units is presumably much easier when those units are stable (i.e. clearly defined). Recall that syllable frequency effects have been observed in nonword judgments (Vitevitch et al., 1997) and production latencies (Cholin, Levelt & Schiller, 2006), indicating that this unit is indeed being tracked by learners; it would therefore seem adaptive to evolve an efficient parsing system.

The third reason why a categorical parser may seem to be preferable to a gradient one is that, in addition to being potentially easier to learn, simple grammars appear to be more robust against variability in individual lexicons and thus more transmittable across successive generations of learners. This point was made by Pierrehumbert (2001), who investigated four statistical regularities in the adult lexicon. In increasing level of granularity, these were as follows: (i) the preference of antepenultimate to penultimate stress across all trisyllables, (ii) the relative well-formedness of five different nasal-obstruent in word-medial position, (iii) a conjunction of (i) and (ii), where the cluster well-formedness was constrained to trisyllables with initial stress, and (iv) the relative well-formedness of word-final, stressed /gɹi/ and /kɹi/ (as in *agree* and *decree*, respectively, see also Moreton, 1997), the former of which is considerably more frequent as a token, but the difference is much smaller when considering types.

Pierrehumbert (2001) investigated the statistical robustness of these four generalizations by conducting a series of simulations intended to resemble vocabulary acquisition. A number of learning agents acquired vocabularies of various sizes by frequency-weighted subsampling from the English lexicon. At each vocabulary size, the agent lexicons were checked for the presence of the four generalizations which

characterize the adult lexicon. A pattern was considered robust against individual variability to the extent that it was acquired earlier and by more agents. The results indicated that general, coarse-grained generalizations like (i) were more robust than specific, fine-grained generalizations like (iii) and (iv). Pierrehumbert (2001) argued that only transmittable (i.e. sufficiently robust) patterns belong in the grammar because grammatical uniformity across speaker/listeners is required for both correct phonetic encoding in production and efficient processing in perception.

In this chapter, I examine the relative robustness of the categorical and gradient metrical parsing models with respect to stress assignment. Relying on vocabulary simulations inspired by (though somewhat different from) Pierrehumbert (2001), I investigate the extent to which both models are acquired by agents at different stages of lexical development.

7.2 Method

As noted above, the categorical parser seems to have a learning advantage due to its simplicity: rather than having to estimate the frequency of each individual insert, the learner simply needs to recognize them as previously encountered word edges. On the other hand, learners are of course quite good at keeping track of frequency information. Indeed, they can't seem to help but learn and match probabilities in the input, especially when that information ends up being useful for discovering linguistic units like phonetic categories or words (e.g. Harmon & Kapatsinski, 2017; Kapatsinski, 2010; Maye et al., 2002; Olejarczuk et al., to appear; Saffran et al., 1996). The question

then becomes not just about simplicity, but about its trade-off with predictive success. In other words, it's a question of statistical model selection.

How well does each parsing model capture the stress facts of English? Recall from Figures 5.1, 5.2 and 5.3 (section 5.2.2) that the categorical parser does a relatively good job of predicting Latin Stress: the heavy penults it produces consistently attract stress across different sections of the lexicon. But what of the gradient parser? After all, our participants appear to have preferred it when stressing nonce forms — would it also outperform the categorical model when applied to their own lexicons? If so, would the improvement be worth the trade-off in complexity? A second, related question concerns the time course of learning. Specifically, at what point during acquisition does the learner have a large enough vocabulary to support the relationship between the parser and stress? As the lexicon grows, do the data consistently prefer the same parsing model for predicting stress, or is there a point at which learners should abandon one in favor of the other?

I approach these questions with a series of vocabulary simulations inspired by Pierrehumbert (2001). One hundred simulated agents learned vocabularies of different sizes by random sampling from the adult English lexicon consisting of 48,951 word forms (defined in section 3.2). To approximate the order of acquisition, the sampling was weighted by SUBTLEXus counts so that the probability of learning a word was proportional to its token frequency. This way, frequent words had a greater chance to be learned early on. Vocabulary size began at 175 words and grew to over 46,000 in 20% increments. After sampling their lexicon, each agent examined the word edges and extracted both categorical and gradient phonotactic information. The former meant simply assigning an "attested" label to each word onset. The latter consisted of

recording word onset and word offset frequencies, as well as the sonority slopes of each onset. As in the models in Studies 1-5, the sonority slopes were residualized against the two measures of lexical support to control collinearity and capture the contribution of phonetic properties of the clusters to syllabification. Having learned their (idiosyncratic) word-edge phonotactics, the agents attempted to predict stress in trisyllabic and longer words using two parsing models. The categorical model syllabified each word based on the Maximal Onset Principle, with each agent relying on its own set of attested onsets to determine legality. The gradient model contained three predictors: the two word-edge frequency measures as well as the onset sonority slopes. Both models were formalized as mixed-effects logistic regressions with random intercepts for individual words.

At each of the 25 increments in vocabulary size, model performance was compared. The assessment method differed from that in Pierrehumbert (2001); there, learnability was operationalized as the Spearman's rank correlation between the relative well-formedness of each of the four phonotactic patterns in the agent's lexicon and the adult lexicon, averaged across all agents of the same vocabulary size. In other words, the learning target was the adult-like rankings of the relevant constraints. Here, the research question is somewhat different in that it concerns the within-learner competition between two alternative models of the same phonological process. Assessment thus involved conducting two different model comparisons within each individual and then averaging within the developmental stages (i.e. vocabulary sizes). First, each regression model was tested against its null (intercept-only) counterpart using a likelihood ratio test. This test indicated whether the predictors significantly improved fit over a baseline, i.e. whether the parser predicted stress assignment in that

individual's lexicon. Second, for the two models of a given lexicon, the values of the Bayesian Information Criterion (*BIC*) were directly compared and posterior probabilities of each model were calculated. As described in section 4.2.3.3, this test penalizes model complexity.

7.3 Results

Figure 7.1 plots the results of the likelihood ratio tests for each model at different stages in lexical development. By the time they learned 2,000 words, virtually all agents acquired both parsing models. Earlier in the simulated development, however, the gradient model was consistently supported by a larger proportion of lexicons. This advantage was not always statistically significant in this sample of 100 agents, but the numerical pattern never showed a reversal (i.e. the majority of lexicons never supported the categorical parser). At the very least, this indicated that, at the individual level, the simpler model is not more learnable.

Figure 7.2 shows the results of the *BIC* comparisons across vocabulary size. The top panel plots the *BIC* difference calculated by subtracting the score of the gradient model from that of the categorical alternative; positive numbers thus indicate an advantage for the gradient parser. The bottom panel shows the posterior probability of the gradient model (the posterior probability of the categorical model is calculated by simply subtracting this value from 1).

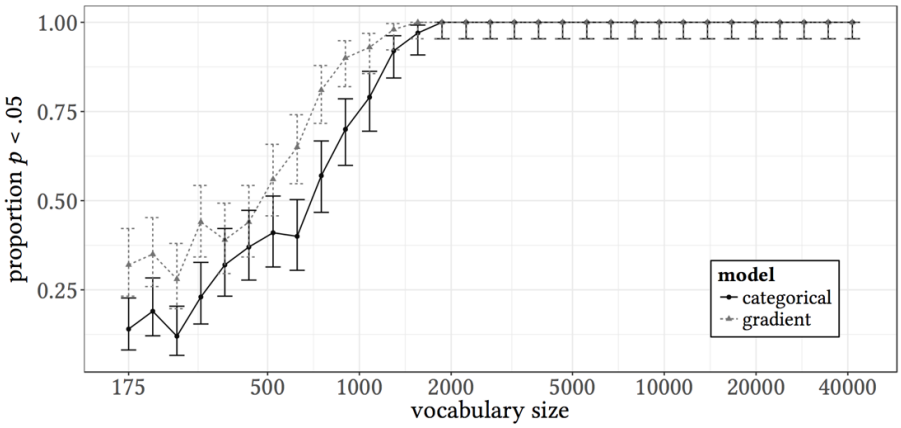

Figure 7.1. Proportion of lexicons where the relevant parsing models significantly outperformed their intercept-only alternatives according to the likelihood ratio test, across vocabulary sizes.

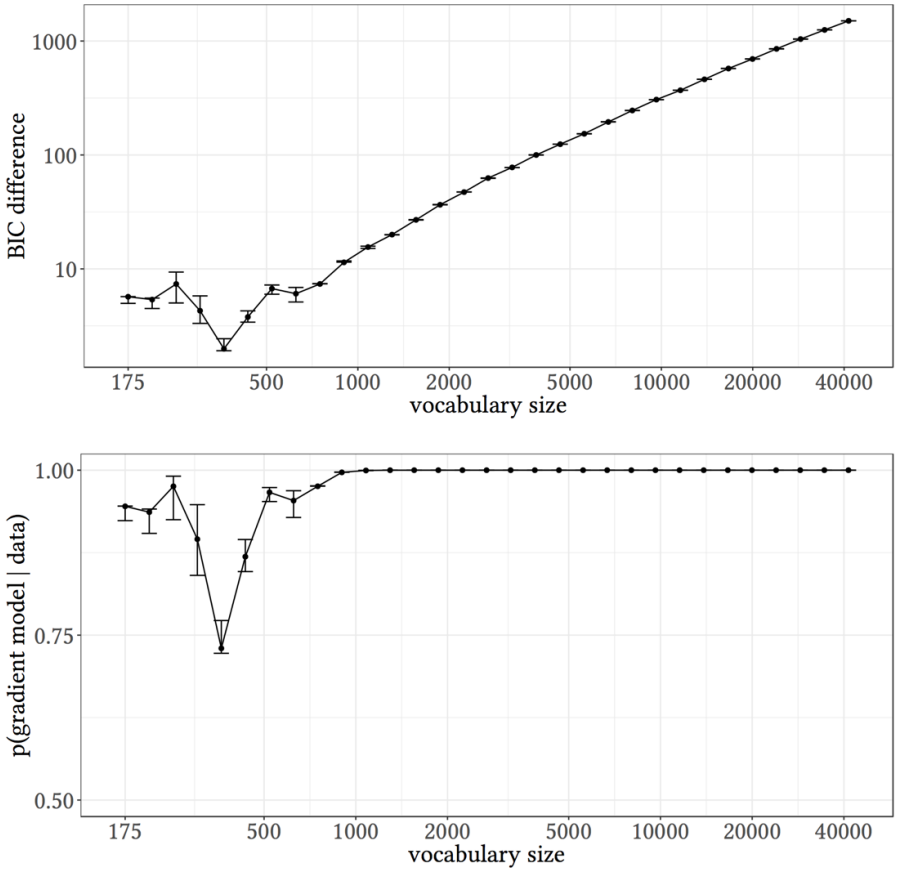

Figure 7.2. BIC score advantage (top) converted to posterior probability (bottom) of the gradient relative to the categorical parsing model, across vocabulary size.

As seen in the top panel, the *BIC* difference is always in favor of the gradient model. The advantage is fairly steady at the first few development stages, dips somewhat (but never reverses) around 360 words and increases exponentially thereafter. The bottom panel reveals that the mean posterior probability of the gradient model never drops below .70, essentially reaching 1 around 1,000 words.

Taken together, the simulation results thus suggest that, when learning phonotactic parsing models, the cost of complexity is surpassed by gains in performance. From early on, stochastic phonotactic knowledge offers a predictive advantage over coarse binning. There is no point during development at which this advantage does not hold, allowing learners to retain the same parsing model as they gather new data. To the extent that the phonological learner favors the most predictive grammar, a learner of the English stress system is expected to acquire a gradient metrical parser.

CHAPTER VIII

CONCLUSIONS

8.1 Summary of the Results and Contributions

In this dissertation, I have proposed that syllabification, or what I have called the *metrical parse*, is a probabilistic process which is guided in large part by gradient well-formedness of potential sub-syllabic constituents. This well-formedness is in turn computed with reference to word-edge statistics and sonority. In other words, the proposal unifies syllabification theory with modern phonotactic theory.

The gradient metrical parser was supported by converging evidence from several studies. In Study 1, participants hyphenated trisyllabic pseudowords. In Study 2, I reanalyzed the results of Eddington et al. (2013a,b), where participants chose from among syllabified alternatives of real English disyllables. Both studies supported the idea that gradient rather than categorical phonotactics guide the parse. This idea was further strengthened in Studies 3, and 4, all of which employed the same nonword stimuli (or a subset thereof) as Study 1. Unlike in Studies 1 and 2, the tasks in the latter experiments were not metalinguistic, relying instead on the productive extension of a real phonological process of Latin Stress assignment.

After establishing baseline expectations about the productivity of Latin Stress by analyzing a lexical database, Study 3 found that preferences for penultimate stress were modulated by gradient phonotactics of medial clusters in the same way as hyphenation behavior. This was supported even more strongly in Study 4, which analyzed stress

location in nonwords produced by the participants. Importantly, behavior in both Study 3 and 4 was influenced by gradient phonotactics independently of analogical factors. Furthermore, the findings of Study 4 could not be explained by three alternative accounts: the phenomenon of gradient weight, Interval Theory or a syllable-free association of stress to individual clusters. Study 5 analyzed speech errors committed by participants of the stress assignment task, and found that, when the metrical parse syllabified the insert as an onset to the final syllable, the likelihood of committing an error was predicted by gradient phonotactics of the insert. This relationship did not obtain for penult-stressed items, where inserts were never parsed as complex onsets. The error results demonstrated that the same lexicon-derived sources of well-formedness which guide the gradient parse also affect production accuracy of medial onsets, demonstrating that the influence of phonotactic knowledge permeates throughout language behavior.

Taken together, the evidence for the gradient parser was diverse and robust. In chapter 6, this was further reinforced by strong insert-level and item-level correlations among the responses of the five studies. The responses also correlated well with Scholes (1966), a seminal study which has served as the test case for a number of state-of-the-art phonotactic models. Vocabulary simulations presented in chapter 7 showed that the gradient parsing model is available to any unbiased learner of English, and that it is preferable to the categorical alternative at all stages of acquisition.

Incorporating gradient phonotactics into syllabification is desirable for theoretical reasons. As discussed throughout chapter II, evidence for fine-grained knowledge of sound sequences is by now overwhelming. Syllabification was one of the few remaining areas of phonology which resisted gradient phonotactics. The evidence

provided in this dissertation argues that both phenomena can and should be modeled under the same probabilistic assumptions. To be clear, stochastic grammars are able to handle categorical as well as gradient behavior (Berent & Shimron, 1997; Berent et al., 2001; Coetzee, 2009; see section 2.4) – indeed, they are preferable exactly because of this flexibility. Nevertheless, demonstrating that human behavior is in fact gradient in some domain constitutes the strongest argument for such grammars. Integrating syllabification with phonotactics under the same modeling assumptions has the desired outcome of lending coherence to the phonological system as a whole.

As noted in section 4.3.4, a common critique of hyphenation studies is that they are susceptible to extra-grammatical sources of knowledge like orthographic conventions, and that they might target word rather than syllable properties (Côté & Kharlamov, 2011; Goslin & Floccia, 2007; Smith & Pitt, 1999; Titone & Connine, 1997; Treiman et al., 2002). Extra-grammatical knowledge is in turn often cited as the locus of frequency effects by approaches that assume a hard distinction between linguistic performance and competence (see e.g. Newmeyer, 2003). Study 5 provides crucial evidence in favor of the usage-based position: unlike hyphenation, stress assignment is a phonological process that is part of natural behavior of speakers of languages with lexical stress. It would be difficult to explain away the results of that study (and the correlation matrices in chapter VI) by appealing to performance factors.

8.2 Implications for Speech Perception and Production

As reviewed in section 2.1.3, the status of the syllable as a unit of speech segmentation is quite controversial, especially in stress languages like English. Cutler et al. (1986) hypothesize that this may be because English is characterized by ambisyllabicity, making syllable-based segmentation inefficient.[12] The present results are certainly consistent with this idea; if syllable boundaries are probabilistic rather than stable, they would make unreliable segmentation cues. At the same time, Study 3 showed that listers *can* infer syllable boundaries in perception — not from allophonic cues but from stress — and judge the resultant parse according to gradient phonotactics. It may be the case that syllable structure is indeed largely ignored in speech segmentation, but nevertheless available for evaluation in a judgment task.

On the speech production side, reliance on the syllable is also not universal: as outlined in section 2.1.2, there is a lively debate about the role of the syllable as a unit of planning or motor execution during spoken word production. Shattuck-Hufnagel (2011) argues that much of the evidence in support of the syllable is consistent with larger planning units; for instance, it may be the case that entire words or even larger phrases constitute production targets (see also Redford, 2015). Does the gradient parsing model I have proposed have any bearing on this debate?

In fact, the results presented in this dissertation may have little relevance for the production of real words. It may well be the case that, rather than stringing together syllable-sized units, 'real world' speech proceeds by activating the largest motor

[12] Kapatsinski & Radicke (2009) note that the stimuli used by Cutler et al. (1986) and similar studies often encourage ambisyllabic interpretation because then tend to feature post-vocalic sonorants.

program that is associated with the intended meaning (both semantic and pragmatic). Nevertheless, I have shown that the syllable as a unit does appear to surface when the speaker is faced with producing an unfamiliar word for which there is no stored plan. In other words, while I advocate for the syllable's existence as a mental object, I accept that its role may be limited.

That said, the present results can be used to evaluate specific claims about the syllable in production models that do employ it. One of the most influential of these theories is presented in Levelt et al. (1999). This model relies heavily on the notion of the syllable at multiple levels. At the level of phonological encoding, the production system constructs phonological words by combining the segmental and metrical components of word forms retrieved from the lexicon. Recall that a phonological word (prosodic word in Figure 1.1) may consist of a single lexical word or a clitic group. The segmental and metrical components of each word form are stored separately in the lexicon. The former are merely strings of phonemes, while the latter are metrical frames which are highly underspecified, usually containing only the number of syllables in each word. For Levelt et al. (1999), stress location is only stored with frames that bear non-initial stress; initial stress is considered the default English pattern and thus assumed to be computed by the grammar. Note that neither the segmental nor metrical components contain any reference to syllable weight or structure (this is in contrast to earlier assumptions, eg. in Levelt, 1992; Levelt & Wheeldon, 1994). Syllabification is a process that operates on the phonological word, i.e. once the segmental strings have been concatenated and the metrical frames merged and recomputed (this allows for syllabification across lexical word boundaries, a major assumption of the model). Crucially, the syllabification process is explicitly assumed to

proceed according to categorical phonotactics. Once the string has been parsed, syllable-sized motor programs are retrieved from the mental syllabary at the level of phonetic encoding. The existence of these motor programs has been supported by frequency effects observed in Dutch repetition latencies (Cholin et el., 2006; Cholin & Levelt, 2008; Levelt & Wheeldon, 1994).

The results presented here challenge some of the assumptions in the Levelt et al. (1999) model. First and most obviously, I have provided new evidence that the metrical parse is gradient rather than categorical. This demands an adjustment to the Levelt et al. (1999) syllabification stage. In and of itself, the adjustment seems minor and easily accommodated by the model. However, it also affects downstream assumptions about the syllabary. If syllabification is gradient, parses like *e.ni.gma* should have non-zero output probabilities. Does this mean that the mental syllabary contains *gma*? If so, then it must also contain a huge number of gestural scores corresponding to all kinds of unconventional (and yet possible) syllables. This seems hardly efficient. If fact, Levelt et al. (1999:5) admit that speakers must be equipped to compose novel syllables without retrieving pre-assembled motor programs from the syllabary, but they argue that occasions that would necessitate this are rare. If the metrical parse is probabilistic (and relevant to real word production), unconventional syllables would surface much more frequently than Levelt et al. (1999) assume, necessitating their addition to the syllabary.

The second major issue is that of lexical storage. Because weight sensitivity can be probabilistically extended from the lexicon and projected onto pseudowords, real word forms must be stored along with information about their syllable structure. Otherwise, English-speaking participants would not be able to probability match the statistics of Latin Stress. Yet, in the Levelt et al. (1999) theory, the stored metrical

frames cannot provide the basis for weight-based generalizations because they are highly impoverished. Furthermore, syllabification is a downstream process, making it impossible to extend Latin Stress without retrieving all the word forms and proceeding to the phonological encoding stage to arrive at the syllabified forms. Encoding a huge number of phonological words without intending to actually produce them surely seems like a wasteful effort. A better solution would be to simply allow for probabilistic syllabification to proceed at two levels: once in the lexicon (making syllable structure available for weight generalizations), and again during phonological encoding (to account for resyllabification across word boundaries). To some, the complications introduced by these adjustments may constitute an argument in favor of abandoning the syllable altogether from real word production models.

8.3 Toward a Model of English Stress

As noted throughout this dissertation, a complete picture of English stress will require substantial effort beyond the present scope. In this section, I sketch out a basic framework for such a model and identify few of the issues that must be addressed. As a starting point, let us maintain the assumption that wug tests are the proper technique for probing the nature of grammatical knowledge (section 3.3). Faced with the task of producing an unfamiliar form, what sort of knowledge is recruited by a native English speaker in order to assign stress?

My own view, hinted at in sections 5.1 and 5.4.4.1.3, is that the phonological grammar is a system of generalizations over the lexicon at multiple levels of organization. The induction process is guided by constraints on production, perception and memory, and the resultant generalizations are stored as part of the mental lexicon itself. This means that so-called 'analogical' and 'grammatical' processing both come from the same source, and the two differ only in degree of generality: what is called analogy is just generalization over low-level features, whereas grammatical processing involves recognizing structural similarities at higher levels.

The idea that stress assignment in pseudowords in multiply determined (section 5.1) is a coherent consequence of this general view. This is due to two corollaries. First, multiple levels of generality are simultaneously available to speakers attempting the stress assignment task in Study 4: in principle, they can choose the overall most common stress pattern in the language (i.e. initial stress, see Cutler & Carter, 1987), or else restrict the search in a number of ways — by lexical class, morphological makeup, number of syllables, syllable weight, segment-level or feature-level similarity to *n*-nearest lexical neighbors, and so on (Baker & Smith, 1978; Guion et al., 2003). Second, these different generalizations are in competition for outputs. For example, the German (initial) and Latin (weight-sensitive) patterns conspire in supporting first-syllable stress in pseudowords with light penults but compete in forms with heavy penults. Similarly, low-level, segment-based similarity to the word *cinema* might compete with the higher-level Latin pattern for *cinempa* (Baker & Smith, 1978). The outcome of such competition is stochastic, decided according to a system of weights on the different generalizations. These weights reflect not only the strength with which each pattern is represented in the lexicon (c.f. 'adjusted confidence scores' in Albright & Hayes, 2003), but also real-

time fluctuations in accessibility (e.g. stress patterns of recently encountered forms might prime the treatment of subsequent forms, see e.g. Harmon & Kapatsinski, 2017).

As noted in section 3.4.2, the focus on phonotactics necessitated the treatment of other generalizations as nuisance covariates in order to isolate the effects of syllable structure. While this terminological choice was justified in the present case, I consider a complete account of the competition among generalizations to be the ultimate goal of the stress modeling effort.

One problem that must be addressed by the probabilistic framework is that of simultaneous acquisition of the gradient parser and of weight sensitivity. Because the shifty nature of syllable boundaries makes it difficult to accumulate frequency counts, this can have cascading effects on the learning of probabilistic associations between syllable structure and stress. For instance, how does a learner probability match syllable weight from the lexicon if syllabification is variable? One promising possibility, suggested by Claire Moore-Cantwell (p.c.) is that weight is estimated from the evidence provided by long vowels. Here, syllable weight can be computed without the need to have a fully-developed model of boundary locations. Acquisition of Latin stress would then proceed as follows. First, learners would simultaneously begin learning syllable edges from word edges while acquiring the probabilistic relationship between stress and CVV+ penults. Having acquired the probabilistic parse, they would then notice that CVC+ penults tend to behave like CVV+ penults, and conclude that these are also heavy. Thus, the model predicting penult stress on the pseudoword *vatablick* would be along the lines of (8.1):

(8.1) $$p(PenStress \mid vatablick) = p(b.l) \times p(PenStress \mid H),$$

where *p(b.l)* reflects the probability of splitting the /bl/ cluster and *p(PenStress | H)* is the probability of stressing penults with long vowels in the lexicon. The final model could further be extended to accommodate gradient weight by adding a term assigning different weights to different penult structures. Developing such a model presents considerable challenges (for instance, to avoid circularity, weight should be defined independently of stress) and falls outside of the present scope. However, as noted above, a comprehensive model of stress assignment must somehow account for the interaction of all relevant generalizations, including gradient phonotactics and gradient weight.

8.4 What is the Syllable?

When linguistics began to be viewed as a branch of psychology in the 1950s, abstract phonological units like the syllable instantly acquired cognitive status without much scrutiny. The decades that followed were less kind to the syllable, with mixed results from perception and production experiments leading to controversies and disagreements (recall sections 2.1.2 and 2.1.3). Nevertheless, the syllable has remained prominent in psycholinguistics. Our most influential models of speech production employ it in their machinery (e.g. Dell, 1986; Levelt, 1989). It features in our theories of how children develop reading and spelling skills (Ferreiro, 2009; Snow et al., 1998). It even plays a role in how we understand speech disorders (Aichert & Zeigler, 2004). But

does the syllable really exist at some level of the mental grammar? And if so, what does it look like?

The evidence provided in this dissertation points to the conclusion that a sublexical unit like the syllable does indeed exist in the internal phonologies of English speakers. As for its shape, it appears to reflect generalizations over word edges in the lexicon. This seems like nothing new: as discussed in chapter 2, the relationship between word margins and syllable margins has been acknowledged in some form since the very beginnings of phonological analysis. Steriade (1999) provides somewhat more recent evidence for the idea that inferences of syllable boundaries are guided by knowledge of word edges. What *is* new here is that the word-edge generalizations that define syllable boundaries are probabilistic rather than deterministic. For the learner, this makes for a complex internal model of the phonological system. As noted in section 7.1, stochastic grammars are at their core based on tracking the frequencies of different units. If the units themselves are probabilistically defined, this makes the learning task that much more complex. Nevertheless, it seems clear from the evidence provided here that such a task is not beyond human learning capabilities.

If syllable margins reflect generalizations over word margins, there is still the question of whether all word margins matter. Recall that many theories have argued for extra-syllabic appendices in words like *masks*, *slammed* and *spice* (Fujimura & Lovins, 1977; Kaye et al., 1990; Treiman et al., 1992), and that the number of attested medial consonant sequences does not reflect all possible combinations of word onsets and offsets (Pierrehumbert, 1994). In this dissertation, complex word edges were largely ignored. Because the medial inserts in the test items were either singletons or biconsonantal, there was no way to incorporate the statistical properties of long word

onsets and offsets into the models. As a result, only (C)C onsets and C offsets were counted in the lexicon, leaving the issue of appendices unresolved. The application of the probabilistic parsing model to long medial sequences remains an area for future work.

The finding that syllabification reflects gradient word-edge phonotactics invites the criticism that syllables are epiphenomenal sublexical chunks with no real cognitive status. Indeed, such a proposal has been advanced by some researchers. For example, Dziubalska-Kołaczyk (2009) argues that phonotactics are best explained by reference to intersegmental cohesion determined by a gradient, sonority-like scale based on perceptual distance between adjacent consonants and vowels. In her view, syllables simply emerge as a result of universal attractive forces between segments. In other words, intersegmental cohesion determines syllable structure rather than the other way around. Some evidence for this proposal is provided in a syllabification study conducted by Bertinetto et al., (2007) with Polish speakers. These results are not incompatible with the studies conducted in this dissertation, since intersegmental cohesion measure closely resembles sonority (in fact, the two are highly collinear in the insert set used here). However, the argument that syllables have no cognitive status whatsoever is challenged by the results of the stress assignment experiment (Study 4). The fact that online stress placement responds to the same units as hyphenation indicates that these units are not mere inferences; they are active participants in phonological productivity.

That said, it remains to be seen whether the present results generalize to other languages. All else being equal, the strong claim is that they should because statistical learning is a general property of the human species (Saffran, Aslin & Newport, 1996). However, all else is never equal; languages differ in word-edge possibilities and

statistics, and factors other than phonotactics may influence syllable division tasks in language-specific ways. Kharlamov (2009) found some effects of word-edge statistics on the well-formedness ratings of medial onsets, but the effects were much weaker than those reported here. Bertinetto et al., (2007) reported that Polish speakers were more sensitive than Italian speakers to the intersegmental cohesion scale in their syllabifications. The authors argued that the difference was due to the fact that Polish has richer phonotactics, providing more learning data (see also Steriade, 1999 for a similar point when comparing English and Arrente). The generalizability of the gradient parsing model to other languages thus remains another open area for future research.

APPENDIX A

STIMULI

List of stimuli with their values on the nuisance predictors. All items were used in Studies 1, 4 and 5. Only items marked with (*) were used in Study 3. For words analyzed in Study 3, see to Eddington et al. (2013a,b).

test item	edit distance (penult bias)	embedded words (antepenult bias)
belesesh	-0.1	0
beleskesh	0.1	0
belezgesh	-0.1	-1
benesid	0.5	-2
benestid	0	-1
benezdid	0	-2
dakadmuth	0.3	1
dakaduth	-0.1	1
dakadwuth	0.2	1
dakamduth	0.5	1
debampab*	-0.6	2
debapab*	-0.5	-1
debapmab*	-0.4	-1
debaprab	-0.3	-1
depansish	-0.9	3
depasish	-0.4	2
depasnish	-1	2
depavrish	-0.8	1
falageck	0.3	-1
falaskeck	-0.5	2
falazgeck	-0.1	-2
fazabish*	0.5	1
fazablish*	0.2	1
fazabnish*	0	1
fazanbish*	0.4	1
fibagath*	0.5	0
fibagnath*	0.1	0

test item	edit distance (penult bias)	embedded words (antepenult bias)
fibagrath*	0.3	0
fibangath*	0	2
gidikwop*	0.4	-1
gidirzop*	0.6	-1
gidizop*	0.7	-1
gidizrop*	0.6	1
hadaseph	-0.3	-2
hadaspeph	-0.7	-1
hadazbeph	-0.5	-3
kapalthiss	0.7	4
kapathiss	0.5	4
kapathliss	0.5	4
kapathriss	0	4
kenadlozz*	0.3	1
kenadozz*	0.4	1
kenadrozz*	0	1
kenalbozz*	0.3	0
kiniltem	0.2	-1
kinitem	0.1	0
kinitlem	0	0
kinitrem	0.1	0
lapanshup*	-0.2	2
lapashnup*	-0.1	2
lapashrup*	0.2	2
lapashup*	0	2
lekagnop	0.1	0
lekagop	0.2	0
lekagrop	0.1	0
lekangop	0.1	1
lepabazz	0.1	2
lepablazz	0	2
lepabnazz	0.1	2
lepanbazz	0	2
lidigeph	0.6	-1
lidigleph	0.9	-1
lidigmeph	0.9	-1
lidimgeph	1	-1
madalpazz*	0	-1
madapazz*	-0.2	-3

test item	edit distance (penult bias)	embedded words (antepenult bias)
madaplazz*	0	-3
madapnazz*	-0.2	-3
menelsuss*	0.3	-1
menesluss*	0.5	-2
menesruss*	0.5	-2
menesuss*	0.1	-2
naragish*	0	1
naraglish*	0.2	1
naragmish*	-0.1	1
naramgish*	-0.1	2
nepantep	-0.6	6
nepatep	-0.9	3
nepatnep	-0.2	3
nepatwep	-0.2	3
nibifim*	0.3	0
nibifmim*	0.1	0
nibifrim*	0.1	0
nibimfim*	0.1	-1
nibisozz	-0.5	0
nibispozz	-1.1	0
nibizbozz	-0.1	0
pimalvib	-0.2	2
pimasmib	-0.2	2
pimavib	0.1	1
pimavlib	-0.1	1
pimintoth*	0.2	3
pimitnoth*	0.1	1
pimitoth*	1.1	1
pimitwoth*	0.7	1
redalthosh*	0.2	0
redathlosh*	-0.5	-1
redathosh*	0	-1
redathrosh*	0.2	-1
sakansud*	-0.1	1
sakasnud*	-0.5	1
sakasud*	-0.1	1
sakavrud*	-0.1	0
sanakep	0.2	-3
sanaknep	0	-2

test item	edit distance (penult bias)	embedded words (antepenult bias)
sanakrep	0	-3
sanankep	0	-3
sebinshaph	0.6	2
sebishaph	0.4	1
sebishnaph	0.7	1
sebishraph	0.5	1
shepidmoph*	0.4	-1
shepidoph*	1.1	-1
shepidwoph*	0.6	-1
shepimdoph*	0.4	-2
shigalpeff	0.5	3
shigapeff	0.7	1
shigapleff	0.6	1
shigapneff	0.2	1
shimabeph*	0.8	1
shimabreph*	0.2	1
shimabweph*	0.2	1
shimarbeph*	0.3	0
sibidoss	0.6	2
sibistoss	-0.1	1
sibizdoss	-0.1	1
sipadesh	0.8	2
sipadlesh	0.2	2
sipadresh	0	3
sipalbesh	0.2	2
tabalvub*	0.5	-1
tabasmub*	-0.2	-1
tabavlub*	0.3	-2
tabavub*	0.2	-1
tamampish	0.1	0
tamapish	0.2	1
tamapmish	0	1
tamaprish	0.3	1
thanabiss	0.3	-1
thanabriss	0.3	-2
thanabwiss	0.2	-2
thanarbiss	0.5	-4
thibifar	0.8	1
thibiflar	0.5	-1

test item	edit distance (penult bias)	embedded words (antepenult bias)
thibilfar	0.4	0
thibizlar	0.6	1
vatafiss*	0.5	-2
vatafliss*	0.9	-2
vatalfiss*	1	1
vatazliss*	0.9	-2
vemiknoph*	0.1	0
vemikoph*	0.4	0
vemikroph*	0.7	0
veminkoph*	0.1	3
wabaltiss*	0.6	1
wabatiss*	-0.2	1
wabatliss*	0	1
wabatriss*	-0.1	1
wibilseph	0.4	0
wibiseph	0	1
wibisleph	0.9	1
wibisreph	0.9	1
zedafmup	-0.1	-2
zedafrup	0	-2
zedafup	0.3	-2
zedamfup	0	-1
zepakwiss	0	1
zeparziss	0.2	2
zepaziss	0.3	1
zepazriss	0	1

APPENDIX B

INSERTS

List of C(C) inserts with their values on the phonotactic predictors.

insert status	insert IPA	log(wd. onset freq.)	log(wd.offset freq.), C1	sonority slope
singleton	b	-3.11	-5.92	6
singleton	d	-2.84	-2.99	6
singleton	f	-3.44	-5.61	7
singleton	g	-4.13	-5.79	6
singleton	k	-2.68	-3.67	8
singleton	p	-3.08	-4.81	8
singleton	s	-2.91	-3.37	7
singleton	ʃ	-4.52	-5.27	7
singleton	t	-3.53	-3.33	8
singleton	v	-4.24	-4.7	5
singleton	z	-6	-2.94	5
singleton	θ	-5.54	-6.06	7
attested	bl	-5.12	-5.92	3
attested	bɹ	-4.71	-5.92	4
attested	bw	-9.41	-5.92	5
attested	dɹ	-5.35	-2.99	4
attested	dw	-7.66	-2.99	5
attested	fl	-4.96	-5.61	4
attested	fɹ	-5.09	-5.61	5
attested	gl	-5.67	-5.79	3
attested	gɹ	-4.62	-5.79	4
attested	kɹ	-4.62	-3.67	6
attested	kw	-5.4	-3.67	7
attested	pl	-5.21	-4.81	5
attested	pɹ	-3.84	-4.81	6
attested	sk	-4.9	-3.37	-1
attested	sl	-5.34	-3.37	4
attested	sm	-6.14	-3.37	3

insert status	insert IPA	log(wd. onset freq.)	log(wd.offset freq.), C1	sonority slope
attested	sn	-5.79	-3.37	3
attested	sp	-4.85	-3.37	-1
attested	st	-4.25	-3.37	-1
attested	ʃn	-8.6	-5.27	3
attested	ʃɹ	-6.89	-5.27	5
attested	tl	-10.11	-3.33	5
attested	tɹ	-4.44	-3.33	6
attested	tw	-6.7	-3.33	7
attested	vl	-10.11	-4.7	2
attested	vɹ	-10.11	-4.7	3
attested	zl	-9.7	-2.94	2
attested	θɹ	-6.43	-6.06	5
unattested	bn	-10.8	-5.92	2
unattested	dl	-10.8	-2.99	3
unattested	dm	-10.8	-2.99	2
unattested	fm	-10.8	-5.61	3
unattested	gm	-10.8	-5.79	2
unattested	gn	-10.8	-5.79	2
unattested	kn	-10.8	-3.67	4
unattested	lb	-10.8	-2.87	-3
unattested	lf	-10.8	-2.87	-4
unattested	lp	-10.8	-2.87	-5
unattested	ls	-10.8	-2.87	-4
unattested	lt	-10.8	-2.87	-5
unattested	lv	-10.8	-2.87	-2
unattested	lθ	-10.8	-2.87	-4
unattested	md	-10.8	-4.05	-2
unattested	mf	-10.8	-4.05	-3
unattested	mg	-10.8	-4.05	-2
unattested	mp	-10.8	-4.05	-4
unattested	nb	-10.8	-2.6	-2
unattested	ns	-10.8	-2.6	-3
unattested	nʃ	-10.8	-2.6	-3
unattested	nt	-10.8	-2.6	-4
unattested	ŋg	-10.8	-2.42	-2
unattested	ŋk	-10.8	-2.42	-4
unattested	pm	-10.8	-4.81	4
unattested	pn	-10.8	-4.81	4

insert status	insert IPA	log(wd. onset freq.)	log(wd.offset freq.), C1	sonority slope
unattested	ɹb	-10.8	-2.63	-4
unattested	ɹz	-10.8	-2.63	-3
unattested	sɹ	-10.8	-3.37	5
unattested	tn	-10.8	-3.33	4
unattested	zb	-10.8	-2.94	-1
unattested	zd	-10.8	-2.94	-1
unattested	zg	-10.8	-2.94	-1
unattested	zɹ	-10.8	-2.94	3
unattested	θl	-10.8	-6.06	4

REFERENCES CITED

Aichert, I. & Ziegler, W. (2004). Syllable frequency and syllable structure in apraxia of speech, *Brain and Language, 88*(1), 148-159.

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

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

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

Anderson, J., & Jones, C. (1974). Three theses concerning phonological representation. *Journal of Linguistics 10*, 1-26.

Arnold, H.S., Conture, E.G. & Ohde, R.N. (2005). Phonological neighborhood density in the picture naming of young children who stutter: Preliminary study. *Journal of Fluency Disorders, 30*, 125–148.

Baayen, R.H., Piepenbrock, R. & Gulikers, L. (1995). The CELEX lexical database. Release 2 [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Baertsch, K. (2012). Sonority and sonority-based relationships within American English monosyllabic words. In Steve Parker (ed.), *The Sonority Controversy*, 3-39. Mouton: Berlin.

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

Baker, R. G., & Smith, P. T. (1976). A psycholinguistic study of English stress assignment rules. *Language and Speech*, *19*(1), 9–27.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates D, Maechler M, Bolker B and Walker S (2014). lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.7, Available at http://CRAN.R-project.org/package=lme4.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, R.H. (2015). Parsimonious mixed models. arXiv:1506.04967 [stat].

Becker, M., Ketrez, N., & Nevins, A. (2011). The Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternations. *Language, 87*(1), 84–125.

Beckman, M. E., & Pierrehurnbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook 3*, 255-309.

Berent, I., Everett, D. & Shimron, J. (2001). Do phonological representations specify formal variables? Evidence from the Obligatory Contour Principle. *Cognitive Psychology, 42*, 1-60.

Berent, I., Lennertz, T., Jun, J., Moreno, M. A., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(14), 5321–5325.

Berent I, Lennertz T, Smolensky P, Vaknin-Nusbaum V. (2009). Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology, 26*. 75–108.

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the Obligatory Contour Principle. *Cognition, (64)*1, 39-72.

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition, 104*(3), 591–630.

Berg, T., & Niemi, J. (2000). Syllabification in Finnish and German: Onset filling vs. onset maximization, *Journal of Phonetics 28*(2), 187–216.

Berko, J. (1958). The child's learning of English morphology. *Word,14*:150-177

Bertinetto, P. M., Scheuer, S., Dziubalska-Kołaczyk, K., & Agonigi, M. (2007). Intersegmental cohesion and syllable division in Polish. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1953-6.

Blevins, J. (2003). The independent nature of phonotactic constraints: an alternative to syllable-based approaches. In Caroline Féry and Ruben van de Vijver (eds.). *The syllable in optimality theory*. Cambridge: Cambridge University Press. 375-403.

Boersma, P. (1997). How we learn variation, optionality, and probability. In R. J. J. H. van Son (ed.), *Proceedings of the Institute of Phonetic Sciences, Amsterdam, 21*, 43–58. Amsterdam: University of Amsterdam, Institute of Phonetic Sciences.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International 5*:9/10, 341-345.

Boersma, P. & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

Broselow, E. (2003). Marginal phonology: phonotactics on the edge. *The Linguistic Review, 20*(2-4), 159–193.

Browman, C. P. & Goldstein, L. G. (1995) Gestural syllable position effects in American English. In F. Bell-Berti & L. J. Raphael, (eds)., *Producing Speech: Contemporary Issues (for Kathering Safford Harris)*. Woodbury, NY: AIP Press, 19-33.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109*, 204-223.

Bruck, M., Treiman, R., & Caravolas, M. (1995). Role of the Syllable in the Processing of Spoken English: Evidence From a Nonword Comparison Task. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 469-479.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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.

Bybee, J. (2001). *Phonology and Language Use (Cambridge Studies in Linguistics, 94).* Cambridge UK: Cambridge University press.

Carpenter, A. (2010). A naturalness bias in learning stress. *Phonology, 27*, 345-392.

Carpenter, A.C. (2016). The role of a domain-specific mechanism in learning natural and unnatural stress. *Open Linguistics, 2*, 105-131.

Cholin, J., & Levelt, W.J.M. (2008). Effects of syllable preparation and syllable frequency in speech production: Further evidence for syllabic units at a post-lexical level, *Language and Cognitive Processes, 24*:5, 662-684

Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition, 99*, 205-235.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English.* New York: Harper & Row.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In M. Beckman (ed.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 282–333). Cambridge: Cambridge University Press.

Clements, G. N. & Keyser, S.J. (1983). *CV Phonology: a Generative Theory of the Syllable.* MIT Press, Cambridge, MA.

Coetzee, A. W. (2009). Grammar is Both Categorical and Gradient. In Steve Parker (ed.), *Phonological Argumentation*, Equinox, London.

Coleman, J. & Pierrehumbert, J.B. (1997). Stochastic phonological grammars and acceptability. In John Coleman (ed.) *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*. Somerset, NJ: Association for Computational Linguistics. 49–56.

Content, A., Kearns, R. K., & Frauenfelder, U. H. (2001). Boundaries versus Onsets in Syllabic Segmentation. *Journal of Memory and Language, 45*(2), 177–199.

Côté, M.-H., & Kharlamov, V. (2011). The impact of experimental tasks on syllabification judgments: a case study of Russian. In C. Cairns & E. Reimy, *Handbook of the Syllable* (pp. 271–294). Boston: Brill.

Crompton, A. (1982). Syllables and segments in speech production. *Linguistics, 19*, 663–716.

Croot, K., & Rastle, K. (2004). Is there a syllabary containing stored articulatory plans for speech production in English? *Proceedings of the 10th Australian International Conference on Speech Science and Technology*. 376-381.

Cser, A. (2012). The role of sonority in the phonology of Latin. In Steve Parker (ed.), *The Sonority Controversy*, 39-65. Mouton: Berlin.

Cutler, A. (1997) The Syllable's Role in the Segmentation of Stress Languages. *Language and Cognitive Processes, 12*:5-6, 839-846.

Cutler, A. (2005). Lexical Stress. In D. B. Pisoni & R. E. Remez, *The handbook of speech perception* (pp. 264–289).

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133–142.

Cutler, A., Mehler, J., Norris, D.G., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*, 385–400.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*(2), 197–234.

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech, Language, and Hearing Research, 11*, 707–721.

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics*, *34*(1), 104–137.

Davis, S. (1985). *Topics in syllable geometry*. Ph.D. thesis, University of Arizona.

Davis, S. (1989). On a non-argument for the rhyme. *Journal of Linguistics, 25*(1), 211-217

Dell, G. S. (1986) A spreading-activation theory of retrieval in sentence production. *Psychological Review 93*:283–321.

Dell, G. S. (1990). Effects of Frequency and Vocabulary Type on Phonological Speech Errors. *Language and Cognitive Processes, 5*(4), 313-349.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1355–1367.

Domahs, U., Plag, I., & Carroll, R. (2014). Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics, 17*, 59-96.

Draper, M. H., Ladefoged, P. & Whitteridge, D. (1959) Respiratory muscles in speech. *Journal of Speech and Hearing Research, 2,* 16-27.

Dupoux, E. (1994). The time course of prelexical processing: The syllable hypothesis revisited. In G.T.M. Altmann & R.C. Shillcock (eds), *Cognitive models of speech processing: The Second Sperlonga Meeting*, pp. 81–123. Cambridge, MA: MIT Press.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1568-1578.

Dziubalska-Kołaczyk, K. (2009). NP Extension: B&B Phonotactics. *Poznań Studies in Contemporary Linguistics, 45*(1), 55–71.

Eddington, D., Treiman, R., & Elzinga, D. (2013a). Syllabification of American English: Evidence from a Large-scale Experiment. Part I. *Journal of Quantitative Linguistics*, *20*(2), 45–67.

Eddington, D., Treiman, R., & Elzinga, D. (2013b). Syllabification of American English: Evidence from a Large-scale Experiment. Part II. *Journal of Quantitative Linguistics*, *20*(2), 75–93.

Ernestus, M., & Neijt, A. (2008). Word length and the location of primary word stress in Dutch, German, and English. *Linguistics 46*(3). 507–540.

Ettlinger, M., Finn, A. S., & Hudson Kam, C. L. (2011). The Effect of Sonority on Word Segmentation: Evidence for the Use of a Phonological Universal. *Cognitive Science*, *36*(4), 655–673.

Fallows, D. (1981). Experimental Evidence for English Syllabification and Syllable Structure. *Journal of Linguistics*, *17*(2), 309–317.

Ferrand, L., & Segui, J. (1998). The syllable's role in speech production: Are syllables chunks, schemas, or both? *Psychonomic Bulletin & Review, 5*, 253–258.

Ferreiro, E. (2009). The transformation of children's knowledge of language units during beginning and initial literacy. In J. V.Hoffman & Y. Goodman (Eds.), *Changing literacies for changing times: An historical perspective on the future of research reading research, public policy, and classroom practices*. New York: Routledge, 61-75.

Frauenfelder, U.H., & Kearns, R.K. (1996). Sequence monitoring. *Language and Cognitive Processes, 11*, 665–673.

Frisch, S. A. (2000). Temporally organized lexical representations as phonological units. In J. B. Pierrehumbert & M. B. Broe (eds.), *Acquisition and the Lexicon: Papers in Laboratory Phonology V.* Cambridge: Cambridge University Press, 283–298.

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

Frisch, Stefan A. and Zawaydeh, Bushra Adnan. (2001) The psychological reality of OCP-Place in Arabic. *Language 77*:91--106.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27-52.

Fudge, E. C. (1969). Syllables. *Journal of Linguistics, 3,* 253–286.

Fujimura, O. & Lovins, J. (1977). Syllables as concatenative phonetic units. *Paper presented at Symposium on Segment Organization and the Syllable*, Boulder, CO, Oct. 21 - 23, 1977, published in 1982 by the Indiana University Linguistics Club, Bloomington, Indiana.

Garcia, G. D. (2017). Weight gradience and stress in Portuguese. *Phonology, 34*(1):41–79.

Giles, S. B. & Moll, K. L. (1975) Cinefluorographic study of selected allophones of English /l/. *Phonetica, 31*, 206-227.

Gillis, S., Daelemans, W., & Durieux, G. (2000). 'Lazy learning': a comparison of natural and machine learning of word stress. In P. Broeder & J. Murre (eds.), *Models of Language Acquisition.* Oxford University Press, 76-99.

Goldsmith, J. (2011). The syllable. In J. Goldsmith, J. Riggle & A. C. L. Yu (eds.). *The Handbook of Phonological Theory, 2nd ed.* Wiley Blackwell. pp. 164-196.

Gordon, M. (1999). *Syllable weight: Phonetics, phonology, and typology*. Doctoral dissertation, UCLA.

Gordon, M. (2002). A phonetically-driven account of syllable weight. *Language 78*, 51-80.

Goslin, J., & Floccia, C. (2007). Comparing French syllabification in preliterate children and adults, *Applied Psycholinguistics, 28*(02), 341–367.

Goslin, J., & Frauenfelder, U. H. (2001). A Comparison of Theoretical and Human Syllabification. *Language and Speech, 44*(4), 409–436.

Guion, S. G., Clark, J. J., Harada, T., & Wayland, R. P. (2003). Factors Affecting Stress Placement for English Nonwords include Syllabic Structure, Lexical Class, and Stress Patterns of Phonologically Similar Words. *Language and Speech*, *46*(4), 403–426.

Gussenhoven, C. (1986). English plosive allophones and ambisyllabicity. *Gramma 10*. 119-141.

Gussenhoven, C. (1992). Intonational phrasing and the prosodic hierarchy. *Phonologica 1988*, 89-99.

Hall, T.A. (2004). English syllabification as the interaction of markedness constraints. *ZAS Papers in Linguistics 37, 2004*: 1 – 36.

Halle, M. (1959). *The sound pattern of Russian*. The Hague: Mouton.

Halle, M. (1998). The stress of English words: 1968–1998. *Linguistic Inquiry, 29*(4), 539–568.

Halle, M., Keyser, S.J. (1971). *English Stress, its Form, its Growth, and its Role in Verse*. Harper & Row, New York.

Halle, M., Vergnaud, J.-R. (1987). *An Essay on Stress*. MIT Press, Cambridge, MA.

Hammond, M. (1995). Syllable parsing in English and French. ROA-58, Rutgers Optimality Archive, http://roa.rutgers.edu/

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

Hammond, M. (1999). *The phonology of English: a prosodic optimality-theoretic approach*. Oxford University Press, USA.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, *98*, 22-44.

Harris, J. (1994). *English Sound Structure*. Oxford: Blackwell Publishers.

Hay, J., Pierrehumbert, J., & Beckman, M. (2003). Speech perception, well-formedness, and the statistics of the lexicon. Cambridge, UK.: *Papers in Laboratory Phonology VI*, 58-74.

Hayes, B. (1980). *A Metrical Theory of Stress Rules*. Doctoral dissertation, MIT.

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

Hayes, B. (1989a). The prosodic hierarchy in meter. In P. Kiparsky and G. Youmans (eds.), *Phonetics and Phonology, Vol 1: Rhythm and Meter.* San Diego: Academic Press. pp. 201-260.

Hayes, B. (1989b). Compensatory lengthening in moraic phonology. *Linguistic Inquiry 20*, 253-306.

Hayes, B. (1995). *Metrical Stress Theory: Principles and case studies*. Chicago: University of Chicago Press.

Hayes, B. & White, J. (2013). Phonological Naturalness and Phonotactic Learning. *Linguistic Inquiry 44*(1), 45-75.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39, 379-440.

Hirsch, A. (2014). What is the domain for weight computation: the syllable or the interval? (pp. 1–12). *Proceedings of Phonology 2013*.

Hitchcock, L. & Greenberg, S. (2001). Vowel height is intimately associated with stress-accent in spontaneous American English discourse. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, 79-82.

Hoard, J. (1971). Aspiration, tenseness, and syllabification in English. *Language, 47.* 133-I40.

Hooper, J. B. (1972). The syllable in phonological theory. *Language*, 48(3), 525-540.

Hooper, J. B. (1976). *An introduction to natural generative phonology.* New York: Academic Press.

Hooper, J. B. (1978). Constraints on schwa deletion in American English. In J. Fisiak (ed.). *Recent developments in historical phonology.* The Hague: Mouton. 183-207.

Hulst, H.G. van der. (1984). *Syllable structure and stress in Dutch*. Dordrecht: Foris.

Hulst, H.G. van der, & Ritter, N. (1999). Theories of the syllable. In: Hulst, H.G. van der & N. Ritter (eds.). *The syllable: views & facts*. Berlin: Mouton de Gruyter, 13-52.

Hyman, L. M. (1985). *A Theory of Phonological Weight*. Dordrecht: Foris.

Inkelas, S., & Zec, D. (1993). Auxiliary reduction without empty categories: A prosodic account. *Working Papers of the Cornell Phonetics Laboratory 8*, 205-253.

Itô, J. (1989) A prosodic theory of epenthesis, *Natural Language and Linguistic Theory*, *7*(2), 217-259.

Jespersen, Otto. (1904). *Lehrbuch der Phonetik*. Leipzig and Berlin: Teubner.

Kager, R. (1989). *A metrical theory of stress and destressing in English and Dutch*, Dordrecht: Foris.

Kahn, D. (1976). *Syllable Based Generalizations in English Phonology*. PhD. dissertation, MIT, Cambridge, MA.

Kapatsinski, V. (2009). Testing theories of linguistic constituency with configural learning: The case of the English syllable. *Language, 85*(2), 248-277.

Kapatsinski, V. (2010). Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology*, *1*(2), 361-393.

Kapatsinski, V. (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language, 89*(1), 110-48.

Kapatsinski, V. (2014). What is grammar like? A usage-based constructionist perspective. *Linguistic Issues in Language Technology, 11*(1), 1-41.

Kapatsinski, V., & J. Radicke. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In R. Corrigan, E. Moravcsik, H. Ouali, & K. Wheatley (eds). *Formulaic Language. Vol. II: Acquisition, loss, psychological reality, functional explanations*, 499-520. Amsterdam: John Benjamins. (Typological Studies in Language 83).

Kaye, J., Lowenstamm, J., & Vergnaud, J. -R. (1990). Constituent structure and government in phonology. *Phonology Yearbook, 7*, 193–231.

Kehoe, M. (1998). Support for metrical stress theory in stress acquisition. *Clinical Linguistics & Phonetics 12*, 1-23.

Kelly, M. H. (2004). Word onset patterns and lexical stress in English. *Journal of Memory and Language, 50*, 231-244.

Kessler, B., & Treiman, R. (1997). Syllable Structure and the Distribution of Phonemes in English Syllables. *Journal of Memory and Language*, *37*, 295–311.

Keuleers, E. (2013). vwr: Useful functions for visual word recognition research. R package version 0.3.0. Available at http://CRAN.R-project.org/package=vwr.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator, *Behavior Research Methods, 42*(3), 627–633.

Kharlamov, V. (2009). Speakers' notion of the syllable: the role of statistical factors in onset wellformedness judgments. *Proceedings of the 2009 annual conference of the Canadian Linguistic Association*, 1-12.

Kiparsky, P. (1979). Metrical structure assignment is cyclic. *Linguistic Inquiry, 10*, 421–441..

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*(5), 1208–1221.

Krakow, R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics, 27*. 23–54.

Kučera, H., & Francis, W. (1967). *Computational analysis of present- day American English.* Providence, RI: Brown University Press.

Kupin, J. J. (1982). Tongue twisters as a source of information about speech production. Bloomington: Indiana University Linguistics Club.

Kuryłowicz, Jerzy. 1948. Contribution à la théorie de la syllabe. Reprinted in his *Esquisses linguistiques, 2nd ed.*, 1973. München: Wilhelm Fink. Vol. 1. 193-220.

Lahiri, A., Riad, T., & Jacobs, H. M. G. (1999). Diachronic Prosody. In H. van der Hulst (ed.), *Word Prosodic Systems in the Languages of Europe* (pp. 335-422). Berlin: Mouton de Gruyter.

Lee, Y. (2006). *Sub-syllabic constituency in Korean and English.* PhD. Dissertation, Northwestern University.

Legate, J. A. & Yang, C. (2012). (2012) Assessing child and adult grammar. In Berwick & Piattelli-Palmarini (eds.) *Rich Languages from Poor Inputs.* Oxford: Oxford University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press.

Levelt, W. J. M. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition 42*, 1–22.

Levelt, W. J. M. & Wheeldon, L. (1994) Do speakers have access to a mental syllabary? *Cognition 50*:239–69.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences 22*: 1–38.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Levitt, A. G., & Healy, A. F. (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language, 24*. 717-733.

Liberman, M. Y. (1975). *The intonational system of English.* Doctoral dissertation, MIT

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8*, 249–336.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon.* Doctoral dissertation, Indiana University, Bloomington, IN.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1–36.

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language, 52*(3), 436-459.

Lunden, A. (2017). Syllable weight and duration: A rhyme/intervals comparison. In *Proceedings of LSA 2017*, vol. 2, 1-12.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

McCarthy, J. (2010). An introduction to harmonic serialism, *Language and Linguistics Compass 4*(10), 1001- 1018.

McCarthy, J. & Prince, A. (1986). Prosodic morphology. ms. Amherst: University of Massachussetts.

McQueen, J. M. (2004). Speech perception. In K. Lamberts & R. Goldstone (eds.), *The handbook of cognition.* London: Sage, 255-275.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior, 20*, 298–305.

Moore, B.C.J. (2013). *An Introductin to the Psychology of Hearing, 6th edition.* Boston: Brill.

Moore-Cantwell, C. (2016). *The representation of probabilistic phonological patterns: neurological, behavioral, and computational evidence from the English stress system.* PhD. Thesis, UMass Amherst.

Moreton, E. (1997). Phonotactic rules in speech perception. *Abstract 2aSC4, 134th Meeting of the Acoustical Society of America*, San Diego, CA, Dec. 1–5.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology, 25*(1), 83–127.

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning. Part I: Structure. *Language and Linguistics Compass, 6*(11), 686–701.

Morrill, T. (2012). Acoustic correlates of stress in English adjective–noun compounds. *Language and Speech, 55*(2), 167–201

Morton, J., Marcus, S. & Frankish, C. (1976). Perceptual centers (P-cen-ters). *Psychological Review, 83*: 405–8.

Murray, R. W. & Vennemann, T. (1983). Sound change and syllable structure in Germanic phonology. *Language, 59.* 514–528.

Nespor, M., & Vogel, I. (1986). *Prosodic Phonology.* Dordrecht: Foris Publications.

Newmeyer, F.J. (2003). Grammar is Grammar and Usage is Usage. *Language,* 79, 682-707.

Norris, D., McQueen, J.M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology 34*, 191–243.

Ohala, D. K. (1999). The influence of sonority on children's cluster reductions, *Journal of Communication Disorders*, *32*(6), 397–422.

Olejarczuk, P. (2014). The productivity and stability of competing generalizations in stress assignment. *Poster presented at the 14th Conference on Laboratory Phonology*, Tokyo, Japan.

Olejarczuk, P. & Kapatsinski, V. (in revision) The role of surprisal in phonological learning: the case of weight-sensitive stress.

Olejarczuk, P., Kapatsinski, V. & Baayen, R.H. (to appear). Distributional learning is error driven: the role of surprise in the acquisition of phonetic categories.

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition, 83,* B13–B23.

Pallier, C., Sebastian-Galle´s, N., Felguera, T., Christophe, A., & Mehler, J. (1993). Attentional allocation within the syllabic structure of spoken words. *Journal of Memory and Language, 32*, 373–389.

Parker, Stephen G. (2002). *Quantifying the sonority hierarchy.* PhD dissertation, University of Massachusetts, Amherst.

Pierrehumbert, J. B. (1994). Syllable structure and word structure: a study of triconsonantal clusters in English. In P. A. Keating (ed.), *Phonological structure and phonetic form: Papers in Laboratory Phonology III* (pp. 168–190). Cambridge, U.K.: Cambridge University Press.

Pierrehumbert, J. B. (2001). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes*, *16*(5/6), 691–698.

Pierrehumbert, J., & Nair, R. (1995). Word games and syllable structure. *Language and Speech, 38*(1), 77–114.

Pike, K. (1947). *Phonemics : a technique for reducing languages to writing.* Ann Arbor : Univ. of Michigan Press

Pike, K., & Pike, E. (1947). Immediate constituents of Mazatec syllables. *International Journal of American Linguistics, 13*, 78–91.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication, 4*, 75–95.

Pitt, M.A. and McQueen, J.M. (1998) Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39*, 347–370

Prince, A. (1991). Quantitative Consequences of Rhythmic Organization. In K. Deaton, M. Noske and M, Ziolkowski (eds.), *Proceedings of the Chicago Linguistic Society 26*(2).

Prince, A., & Smolensky, P. (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar.* Technical Report, Rutgers University and University of Colorado at Boulder, 1993, Rutgers Optimality Archive 537, 2002, Revised version published by Blackwell 2004, New York, NY.

Pulgram, E. (1970) *Syllable, Word, Nexus, Cursus*, Mouton, The Hague.

Raffelsiefen, R. 1999. Phonological constraints on English word formation. In: G. Booij and J. van Marle (eds.) *Yearbook of Morphology 1998.* Kluwer. 225-287.

Redford, M. A. (2008). Production constraints on learning novel onset phonotactics. *Cognition*, *107*, 785–816.

Redford, M.A. (2015). Unifying speech and language in a developmentally sensitive model of production. *Journal of Phonetics, 53*, 141-152.

Redford, M.A. & Oh, G.E. (2015). Children's abstraction and generalization of English lexical stress patterns. *Journal of Child Language*, Available on CJO 2015 doi:10.1017/S0305000915000215.

Redford, M. A., & Randall, P. (2005). The role of juncture cues and phonological knowledge in English syllabification judgments. *Journal of Phonetics*, *33*(1), 27–46.

Ryan, K. M. (2011a). *Gradient weight in phonology.* PhD. dissertation, UCLA.

Ryan, K.M. (2011b). Gradient syllable weight and weight universals in quantitative metrics. *Phonology,28*:413–454

Ryan, K. M. (2014). Onsets contribute to syllable weight: Statistical evidence from stress and meter. *Language 90*(2), 309-341.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-olds. *Science, 274*, 1926-1928.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime. Pittsburgh, PA: Psychology Software Tools.

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

Selkirk, E. (1978). On prosodic structure and its relation to syntactic structure. In T. Fretheim (ed), *Nordic Prosody II*, Trondheim: TAPIR.

Selkirk, E. (1982). The syllable. In H. van der Hulst & N. Smith (eds.), *The structure of phonological representations* (pp. 337–383). Dordrecht: Foris.

Shademan, Shabnam (2006). Is phonotactic knowledge grammatical knowledge? In D. Baumer, D. Montero, and M. Scanlon (eds.). *Proceedings of the 25th West Coast Conference on Formal Linguistics,* 371–379.

Shattuck-Hufnagel, S. (1992). The role of word structure in segmental serial ordering. *Cognition, 42*(1), 213-259.

Shattuck-Hufnagel, S. (2011). The role of the syllable in speech production in American English: a fresh consideration of the evidence. In C.E. Cairns & E. Raimy (eds). *Handbook of the Syllable.* Boston; Brill. pp. 197-224.

Shattuck-Hufnagel, S., & Turk, A. (1996) A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research 25*(2). 193-247.

Shelton, M., Gerfen, C., & Gutiérrez Palma, N. (2012). The interaction of subsyllabic encoding and stress assignment: A new examination of an old problem in Spanish. *Language and Cognitive Processes*, *27*(10), 1459–1478.

Shport, I. A. (2011). *Cross-linguistic perception and learning of Japanese lexical prosody by English listeners.* Doctoral dissertation, University of Oregon.

Sievers, E. (1881). *Grundzüge der Phonetik*. Leipzig: Breitkopf and Hartel.

Smith, K. L., & Pitt, M. A. (1999). Phonological and Morphological Influences in the Syllabification of Spoken Words. *Journal of Memory and Language*, *41*(2), 199–222.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry, 31*(3), 575-582.

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing, 18*(2), 89-99.

Steriade, D. (1999). Alternatives in syllable-based accounts of consonantal phonotactics. In O. Fujimura, B. Joseph & B. Palek (eds.). *Proceedings of LP Vol. I.* Prague: Charles University and Karolinum Press. pp. 205-2446.

Steriade, D. (2012). Intervals vs. syllables as units of linguistic rhythm. Handouts, *EALING*, Paris.

Stetson, R. H. (1951) *Motor Phonetics: A Study of Speech Movements in Articulation* (2nd Ed.). Amsterdam: North Holland.

Stockall, L., & Marantz, A. (2006). A single route, full decomposition model of morphological complexity: MEG evidence. *The Mental Lexicon*, *1*(1), 85-123.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research, 49*(6), 1175-1192.

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review, 18*(3), 605-611.

Tesar, B. & Smolensky. P. (2000). *Learnability in Optimality Theory.* Cambridge, MA: MIT Press.

Titone, D., & Connine, C. M. (1997). Syllabification strategies in spoken word processing: Evidence from phonological priming. *Psychological Research*, *60*, 251–263.

Trager, G. L., & Bloch, B. (1941). The syllabic phonemes of English. *Language, 17*, 223–46.

Treiman, R. (1983). The structure of spoken syllables: evidence from novel word games. *Cognition, 15*, 49 - 74.

Treiman, R. & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language*, *27*, 87-104.

Treiman, R., & Zukowski, A. (1990). Toward an understanding of English syllabification. *Journal of Memory and Language, 29*(1), 66–85.

Treiman, R., Bowey, J. A., & Bourassa, D. (2002). Segmentation of spoken words into syllables by English-speaking children. *Journal of Experimental Child Psychology, 83*, 213–238.

Treiman, R., Gross, J., & Cwikiel-Glavin, A. (1992). The syllabification of /s/ clusters in English. *Journal of Phonetics, 20*, 383–402.

Treiman, R., Straub, K., & Lavery, P. (1994). Syllabification of bisyllabic nonwords: evidence from short-term memory errors. *Language and Speech*, *37* (1), 45–60.

Turk, A. E., Jusczyk, P. W., & Gerken, L. (1995). Do English-learning infants use syllable weight to determine stress? *Language and Speech*, *38*(2), 143–158.

Vennemann, T. (1972). On the theory of syllabic phonology. *Linguistische Berichte* 18: 1-18.

Vennemann, T. (1988). *Preference laws for syllable structure and the explanation of sound change: With special reference to German, Germanic, Italian, and Latin.* Berlin: Mouton de Gruyter.

Vitevitch, M. & Luce, P.A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9*, 325-329.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: implications for the processing of spoken nonce words. *Language and Speech*, *40*(1), 47–62.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.

Walch, M. L. (1972). Stress rules and performance. *Language and Speech, 15*, 279 – 287.

Warker, J.A., & Dell, G.S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*(2), 387–398.

Watkins, L. (1984). *A grammar of Kiowa*. Lincoln: University of Nebraska Press.

Weide, Robert L. (1994). CMU Pronouncing Dictionary. Available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Whalen, C. A., & Dell, G. S. (2006). Speaking outside the box: Learning of non-native phonotactic constraints is revealed in speech errors. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society, 2371-2374.* Mahwah, NJ: Erlbaum.

White, J. (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language, 93*(1), 1-36.

Whitney, W.D. (1874). *Oriental and Linguistic Studies*. Scribner, Armstrong.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science, 30*, 945–982.

Wright, R. A. (2004). A review of perceptual cues and cue robustness**.** In B. Hayes, R. Kirchner, & D. Steriade (eds.), *Phonetically based phonology* (pp. 34-57). Cambridge; New York: Cambridge University Press.

Yang, C. (2005). On productivity. *Yearbook of Language Variation, 5*, 333-370.

Yarkoni, T., Balota, D. A., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979.

Yi, K. (1999). The internal structure of Korean syllables. *2nd International Conference on Cognitive Science and the 16th Annual Meeting of the Japanese Cognitive Science*

Society Joint Conference, 978–981. Tokyo: The Japanese Cognitive Science Society.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, *12*(1), 85–129.

Zhang, J. (2002). *The effects of duration and sonority on contour tone distribution--A typological survey and formal analysis.* Routledge, New York.