The phonological grammar is probabilistic: New evidence pitting abstract representation against analogy

Claire Moore-Cantwell

Yale University

1 Introduction

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

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

This rule, like most generalizations in the English stress system, has exceptions. Words like *gálaxy* violate the first clause, and words like *vanílla* violate the second clause. However, speakers of English do demonstrate probabilistic knowledge of this and other generalizations (Olejarczuk, 2014; Domahs et al., 2014). This paper investigates speakers' knowledge of an exceptionful generalization in the English stress system. Participants in a wug-test extend this generalization probabilistically to novel words. One possible explanation for this behavior is that participants analogize to existing words in order to choose a stress pattern for a nonword. In this case, the statistics of the lexicon would 'automatically' be transferred to participants' production of nonwords. The production experiment directly tests for this possibility, but does not find evidence that participants use particular actual words to make their choice of stress on nonwords. I argue that the trend is abstractly represented as part of the phonological grammar.

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

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

2 Models of productivity

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

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

However, because the ability to generate a new form is used in day-to-day speech relatively rarely (compared to the ability to generate a new sentence, which is used almost constantly), it is tempting to posit a phonological system that is simplified compared to the syntactic system - a system which instead of a grammatical component uses independently necessary cognitive mechanisms to account for phonological productivity. For example, McClelland and Elman (1986) develop a model of the lexical access process which also predicts, essentially as a side effect, that generalizations in the lexicon should affect speakers' perception of novel words. Other strategies include 'statistical learning' mechanisms for phonology (Seidenberg et al., 2002 and references therein), which simplify the linguistic system by assuming that general cognitive mechanisms for pattern learning in a broad array of contexts apply also to a language's lexicon. Spreading-activation models (Rumelhart and McClelland, 1986) and analogical models (Skousen, 1989; Daelemans et al., 1994; Nakisa et al., 2001) do posit some extra mechanism beyond the lexicon, but one without explicit representation of any generalizations.

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

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

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

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

While grammatical models of variable phonology and agrammatical models both have strengths and weaknesses, it is surprisingly difficult to distinguish the predictions of the two types of model for probability matching behavior. Ernestus and Baayen (2003) consider six different models for their data, two of which are OT-type models, and four of which do not use a grammatical mechanism. All models achieve a close fit to the experimental data, and Ernestus and Baayen ultimately decide in favor of the agrammatical models on the grounds that they are simpler. However, they also point out that their simplicity metric is flawed in that it counts parameters of each model, but does not does not take into account other types of complexity such as the complexity of the learning process or the computational time required to produce a given output (which can be quite long for some analogical models).

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

In the experiment presented here, participants are given nonwords which are phonetically distant from any actual words. Following the methodology of Guion et al., participants are asked to provide potential analogical bases for each nonword. A system which does not represent the probabilistic trend grammatically predicts that such distant nonwords should either (a) be able to analogize to some word or group of words despite the phonetic distance, (b) vary between all possible stress patterns at chance (not following the generalization in the lexicon), or (c) be unstressable. Rather, I find that speakers can generalize the trend to nonwords, but that their chosen analogical bases do not relate to their choices of stress for the nonwords.

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

3 The stress generalization

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

For purposes of this search, a light syllable is defined as one (a) with no coda consonant, and (b) whose vowel is a monophthong (one of [ɑ, æ, ɛ, ɔ, ɪ, ʊ, i, u, ə, ɚ]). The maximal onset principle was used to determine syllabification, so e.g. 'palmistry' is treated as [pɔ.mɪ.stɹi], and included in the set of all-light words. However, a word like 'bonanza' ([bənæn.zə]) would count as heavy, since its first syllable must have a coda consonant - [nz] is not a legal onset in English.

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

The success of this morphological discrimination was assessed by randomly sampling 100 words from each category (morphologically simple, morphologically complex) for lengths of 2 syllables, 3 syllables, and 4 syllables. A native English speaker (the author) then checked these randomly sampled words (600 total) and noted the number of incorrect categorizations in each sample. The results of this categorization are shown in Table 1.

A total of 5101 words met the criteria of having a light penult, and being long enough to potentially take antepenultimate stress (at least trisyllabic). In these words, the most common final vowels were [ə] (1198 words), [i] (1531 words), [n̩/m̩/ŋ̍] (949 words), and [l̩] (674 words). Main stress was almost always penultimate or antepenultimate. A small number of words took preantepenultimate main stress (365) or final main stress (155). How likely each stress pattern was depended on the final vowel. Words ending in

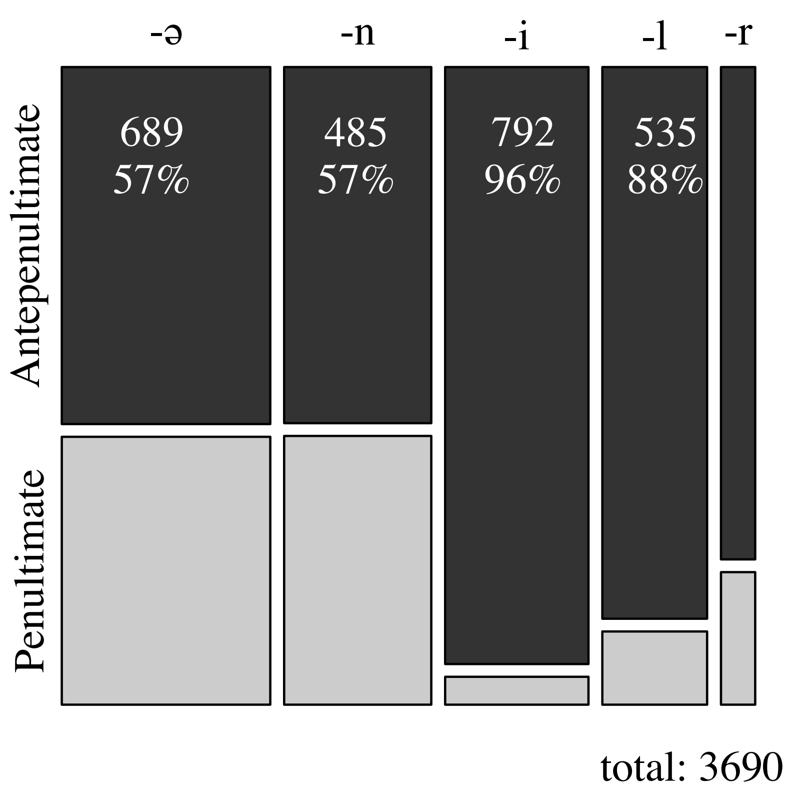
[1] exceptions include *gálaxy, cháracter, ádjective*

Table 1: Number of incorrect categorizations in each random sample of 100 words

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

[-i] or [-l̩] were more likely to take antepenultimate stress than penultimate stress, but words ending in [-ə] or syllabic nasals were roughly equally likely to take penultimate or antepenultimate main stress. These distributions are recorded in Figure 1.

Figure 1: Words with light penultimate syllables and the five most common final syllable nucelii, from the CMU pronouncing dictionary (North American English).

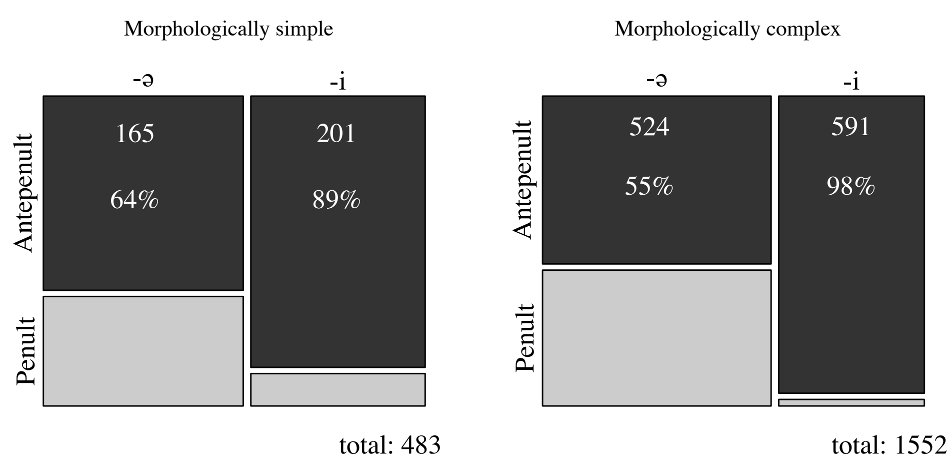


Word-final [i], [l̩], and also [ɹ̩] have been treated as special by Chomsky and Halle (1968); Liberman and Prince (1977); Hayes (1982), all of whom noticed that words ending in these segments were more likely to have exceptional preantepenultimate stress, e.g. *álligator, állegory, párticiple.* In these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

The corpus data in Figure 1 demonstrates that words with final [i/l̩/ɹ̩] also tend to have much more antepenultimate stress than penultimate stress. This would be predicted if these segments tend to be extrametrical - a three syllable word like ćannery would be treated as a two syllable word, [kæ.nə], and take initial stress.

Many [i/l̩/ɹ̩]-final words in English are morphologically complex, and specifically contain a suffix which is stressless or which demands antepenultimate stress, e.g. -y (yellow ∼ yellowy), -er (yellow ∼ yellower), -ity (absúrd ∼ absúrdity), -able (avért ∼ avértable). However, the statistical differences between these three final vowels and final schwa persists in monomorphemic words. Figure 2 illustrates this with just -i and -ə.

Figure 2: Words with light penultimate syllables, classified as morphologically simple or complex, from the CMU pronouncing dictionary (North American English). Only words which end in either [i] or [ə] are included.



The experiment presented below will directly compare participants' behavior on items with final -i vs.

final -ə. Items are exactly three syllables long, and have only light syllables.

4 Analogical models of Productivity

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

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

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

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

If a listener perceiving the ambiguously-stressed nonword [dækæθi] arrived at a state where 'ápathy' received the most activation (let's call this the case where 'ápathy' is *accessed*), that listener would then be more likely to produce the word as [dǽkəθi], with antepenultimate stress. On the other hand, if she arrived at a state where 'decáthlon' received the most activation ('decáthlon' is accessed), she would be more likely to produce [dəkǽθi], with penultimate stress. If the structure of the lexicon is such that 'ápathy' or other antepenultimately-stressed words are accessed more often than 'decáthlon' or other penultimately stressed words, then in aggregate more speakers will choose antepenultimate stress for [dækæθi]. For some nonwords, there will be a single actual word which most speakers will access, while for others there may be multiple actual words which are equally similar in segmental content to the nonword, so that different speakers will access different ones. In either case, across many nonwords, the accessed word will be more likely to have the stress pattern which prevails in the lexicon than to have the minority stress pattern, and therefore in aggregate speakers will produce more of the prevailing stress pattern.

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

5 Production Experiment

5.1 Introduction

This experiment was modeled after Guion et al. (2003), in which the productivity of certain trends in the English stress system was tested by asking participants to pronounce novel English words. The challenge for a production task for the English stress system is that English orthography is non-transparent, and different participants may interpret one orthographic string in many different ways. Vowels are especially difficult to represent unambiguously in English orthography, which is problematic because the quality of a vowel is one factor which can affect the stress of a word. On the other hand, it is difficult to present a novel word auditorily without giving it some stress pattern. Guion et al. solved this problem by presenting auditorily strings of individual syllables, each pronounced as a separate prosodic word, and asking participants to string the syllables together into a word. I copy that methodology here.

5.2 Methods

5.2.1 Participants

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

5.2.2 Items

Items were a subset of the items used in the ERP experiments, so that all were three syllables long, consisting only of light syllables (codaless syllables whose vowel is a monophthong). Both novel words and real words were used. Because participants were asked to pronounce a novel word after listening to three isolated syllables, real words were included in order to encourage participants to make their productions as like real English words as possible - in particular to encourage them to reduce unstressed vowels in their productions. Real words used in the experiment were evenly split among the four word shapes used in the ERP experiments (-i final antepenultimate stress, -i final penultimate stress, -əfinal antepenultimate stress, -əfinal penultimate stress).

Each item consisted of three auditorily presented individual syllables, and two auditorily presented versions of the full word, with different stress patterns (Antepenultimate and Penultimate). Participants first heard the syllables, then pronounced the word, then heard the two stress option, and chose between them. An example item is shown in Figure 3.

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

🔊 [bæ] [mæ] [ki]

RECORD...

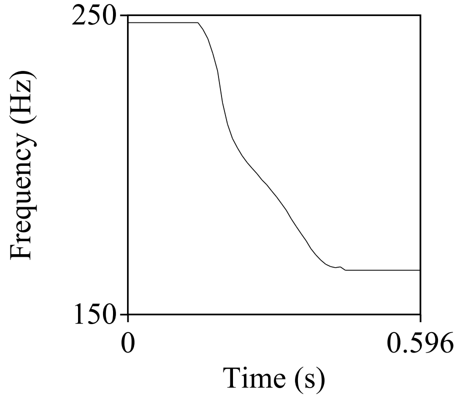
What did you say?

🔊 [bǽməki] 🔊 [bəmǽki]

☐ ☐

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

Figure 4: Pitch contour for individual syllables presented to participants - all syllables were resynthesized to have this contour



stressed vowel.

For the two stress versions of each item, the stimuli were transcribed into the international phonetic alphabet (IPA) and pronounced in a random order by a male native speaker of American English, in the frame sentence "Say X again." The words were then spliced out of the frame sentence. Eight real words were also included - two words for each stress x vowel quality condition. In each case, the stressed vowels of the correctly stressed and misstressed versions of the word matched.

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

The isolated syllables were constructed in the following way: A female native speaker of American English (the author) read a list of individual syllables written in IPA. These recordings were then resynthesized in Praat (Boersma and Weenink, 2011) so that each vowel was approximately 400ms long, and faded into silence over the final 100ms. The pitch contour of the syllables was also resynthesized to be identical (a H* pitch accent followed by a H-H% boundary tone, shown in Figure 4). The intensity of the syllables was also normalized.

The same 32 nonwords that were used in the production task were used in the analogical base task.

Participants were divided into two groups: Group 1 heard full word pronunciations, where one stress pattern or the other was randomly selected for each item. Group 2 heard the individual syllable prompts.

5.2.3 Procedure

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

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

5.3 Results

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

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

5.3.1 Production results

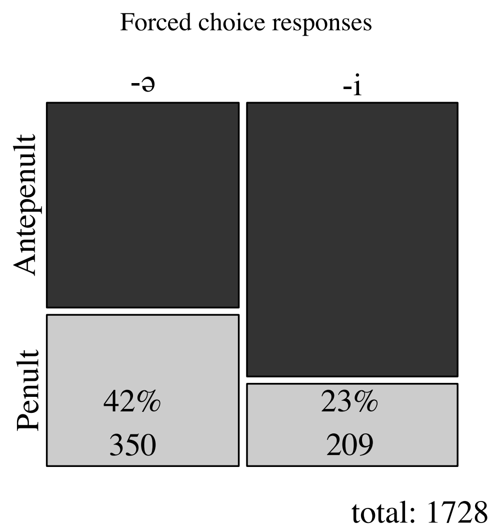
Both groups of participants did the same production task, so their results were analyzed together. Overall, participants’ productions of nonwords probabilistically obeyed the strong i-final generalization, and did not obey the weak ə-final generalization.

Figure 5 shows the counts of each type of stress response for each type of final vowel. These counts are responses in the forced choice task, but recall that participants are only included if their choice of stress agreed with their produced stress at least 90% of the time. This means the counts can be thought of as production counts. Additionally, trials in which participants failed to listen to both stress options before responding were excluded.

Overall, participants preferred antepenultimate stress for both ə-final nonwords and i-final nonwords, but this preference was slight in the ə-final case, and relatively strong in the i-final case. This pattern matches the distribution in the lexicon among all three-syllable and longer words. Note also that participants responded with a probabilistic rather than a categorical preference for antepenultimate stress on i-final nonwords.

A mixed effects logistic regression was fitted to this data, with produced stress as the dependent variable, final vowel as the predictor, and including random slopes and intercepts for both subjects and items. Penul-

Figure 5: Counts of stress choices for each final vowel in nonwords



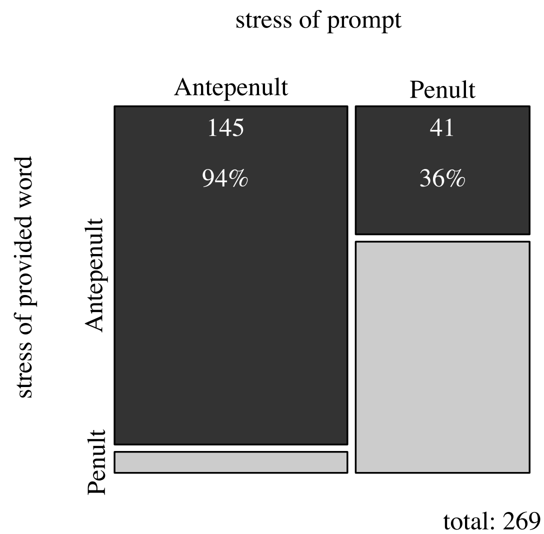
timate stress was the baseline value, so negative coefficients indicate a preference for antepenultimate stress. There was a slight preference for antepenultimate stress over penultimate stress over all (Intercept=-0.35, p=0.049), and a stronger preference when the final vowel was [i] (β= -1.27, $p<0.001$).

5.3.2 Analogy results

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

In group 1, participants mostly responded with two and three syllable words. Their three-syllable responses tended to match the stress pattern of the prompt they were given. If the prompt was an antepenultimately stressed sequence like [bǽməki] they were more likely to write an antepenultimately stressed word in the blank, while if the prompt was penultimately stressed, like [bəmǽki], they were more likely to write down a penultimately stressed word in the blank. Figure 6 shows the relationship between the stress of the prompt and the stress of the response, for all three syllable responses.

Figure 6: Agreement between the stress of the prompt and the stress of the word participants give in response: only three-syllable responses are included here.

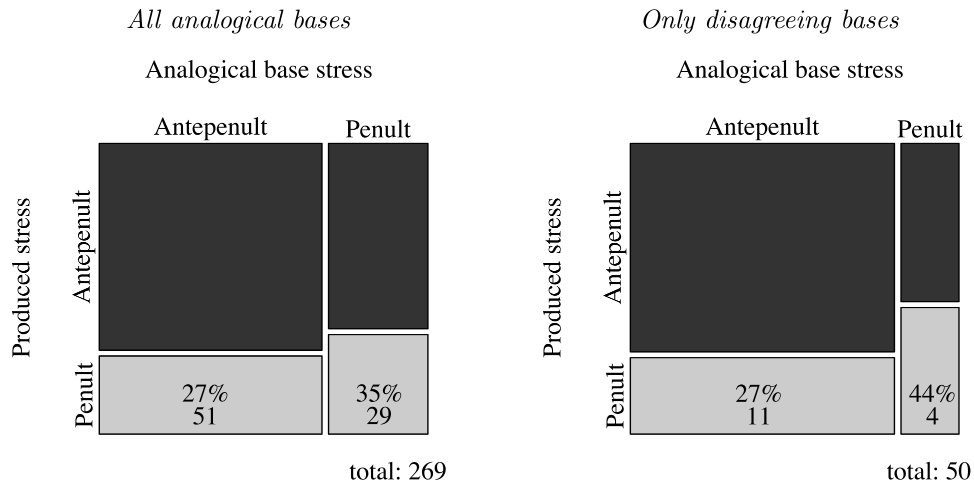


When the prompt is antepenultimately stressed, the vast majority of responses were also antepenultimately stressed. When the prompt was penultimately stressed, the majority of words were penultimately stressed, but the trend for agreement was much less strong.

The stress of the analogical bases provided by group 1 did not in general predict their choices of stress in the production task. Figure 7 shows the relationship between the stress of the word participants wrote down in the analogical base task and the stress they produced for that item in the production task. The majority of produced stresses were antepenultimate, regardless of the stress of the word given in the analogical base task, and there does not appear to be any preference for a participant's produced stress to match the stress of their chosen analogical base. This is also true if only bases that disagree with the stress of the prompt are examined, although in this latter case the numbers are quite small.

Although the task given to group 1 was relatively natural - listen to a felicitous nonword and write down a near actual word neighbor - the bases provided in this task may not be the same bases which would be accessed during the production task. From participants' responses, it is clear that the stress pattern of the

Figure 7: The relationship between the stress of the participant-provided analogical base, and that participant's produced stress for that item: The left figure shows all analogical bases, while the right shows only those which disagree in stress with the prompt



prompt in the analogical task partially determined their choice of real word, but in the production task the prompts did not have a stress pattern. The words that a participant accessed while listening to three separate monosyllables with no stress relationship between them might be different than the words accessed while listening to a nonword with a particular stress pattern.

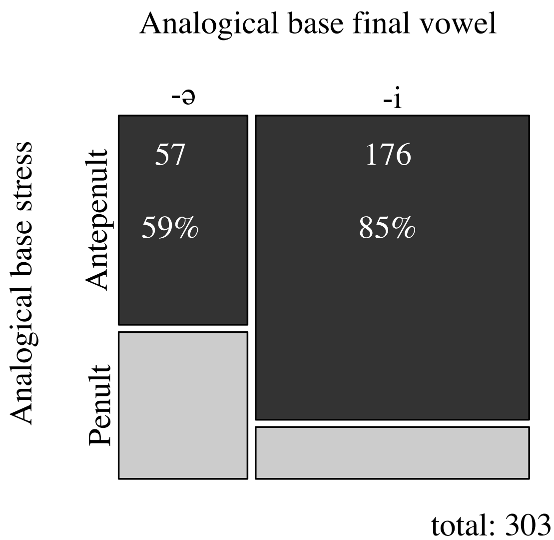
Participants in group 2 were given the exact same prompts in the analogical task as in the production task. Words given in response to these prompts varied in how long they were, with three-syllabled words being the most common. Table 2 shows the number of words produced with in each length category.

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

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

Analogical bases provided by the participants typically had the same final vowel as the nonword prompt. For i-final prompts, the given analogical base ended in [i] 63% of the time overall, and 92% of the time in three-syllable bases. For ə-final prompts, the given analogical base ended in [ə] 36% of the time overall, and 50% of the time in three-syllable bases. In all cases, agreement was the most common outcome. The analogical bases provided by participants tended to be antepenultimately stressed overall, but especially when they ended in [i]. Figure 8 illustrates this.

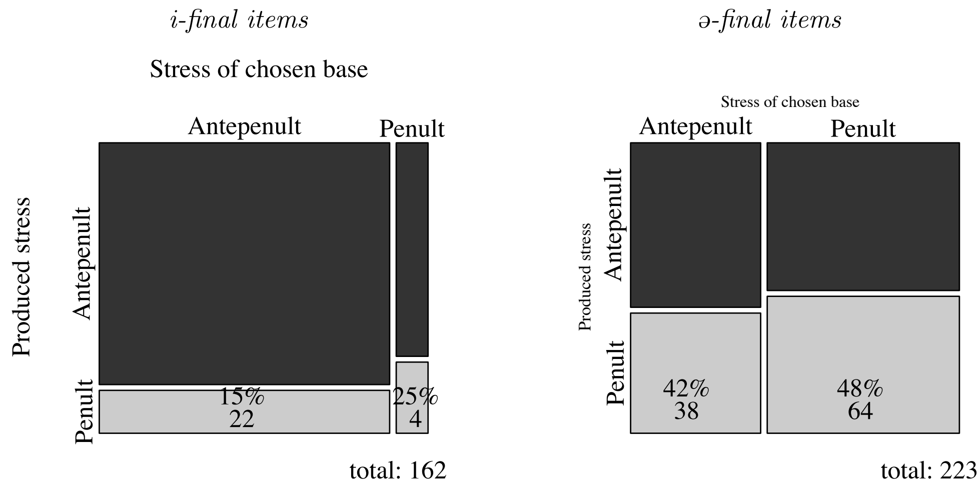
Figure 8: Stress of provided analogical bases which ended in [i] or [ə]: final vowel indicates the final vowel of the actual provided base, not the final vowel of the nonword prompt.



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

Figure 9 shows the relationship between the stress of the analogical base provided for an item by a participant, and that participant's produced stress pattern on that item. Only analogical bases long enough to host antepenultimate stress (3 syllables long or longer) are included. In both the produced stress patterns and in the choice of analogical bases, participants give more antepenultimate stress on i-final words than on ə-final words, but there does not appear to be a direct relationship between the stress of the given base and the produced stress on that item. The items are divided by final vowel, and in neither i-final nor ə-final items is there a clear relationship.

Figure 9: The relationship between the stress of the participant-provided analogical base, and that participant's produced stress for that item: The left figure shows i-final items, and the right figure shows ə-final items. Only analogical responses 3 syllables long and longer are included.



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

Table 3: Logistic regression with two factors: negative coefficients mean greater chance of antepenultimate stress, while positive coefficients mean greater chance of penultimate stress

Model: *Produced Stress* ~ *Final Vowel* + *Analogical Base Stress*

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

In the model with both factors, reported in Table 3, the final vowel of the stimulus has a large coefficient and is highly significant. The stress of the analogical base has only a small coefficient and is not significant. In order to assess the relative contributions of the two factors to the overall model fit, each one was dropped from the model and the fit of that simpler model was compared to the fit of the model with both factors.

The results shown in Table 4 demonstrate that if the stress of the analogical base is removed from the model, the fit is not worsened. On the other hand, when final vowel is removed as a predictor, the fit of the

Table 4: Each factor was dropped from the model, and that resulting simpler model was compared to the full model.

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

model is significantly worse. Based on these results, it is possible to conclude that the stress of the analogical base is not related to the produced stress of that item.

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

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

Table 5 lists the different bases that participants wrote down for the item [rɛ mɛ nə]. Two-thirds of the bases had antepenultimate stress and one-third had penultimate stress. Only bases long enough to take antepenultimate stress (3 syllables or longer) are analyzed here, although results including two-syllable bases as well are very similar. This was about half of the responses.

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

Table 5: Example list of analogical base responses for a stimulus. For this stimulus, a base with antepenultimate stress was given 2/3 of the time.

Stimulus	[rɛ mɛ nə]			
Analogical Base	banána	rémedy	hármony	pánama
no. Responses	2	2	1	1
67% Antepenult, 33% Penult				

For most items, the most common stress pattern among the given analogical bases was antepenultimate. This was also modulated by the item's final vowel. Table 6 shows that more ə-final stimuli than i-final stimuli had majority penultimately stressed bases. This is expected, since there are more ə-final penultimately stressed words in the English lexicon than i-final penultimately stressed words. This table shows the same effect as Figure 8 - namely that participants' choices of analogical base generally tend towards being antepenultimately stressed but otherwise mimic the trends found in the lexicon.

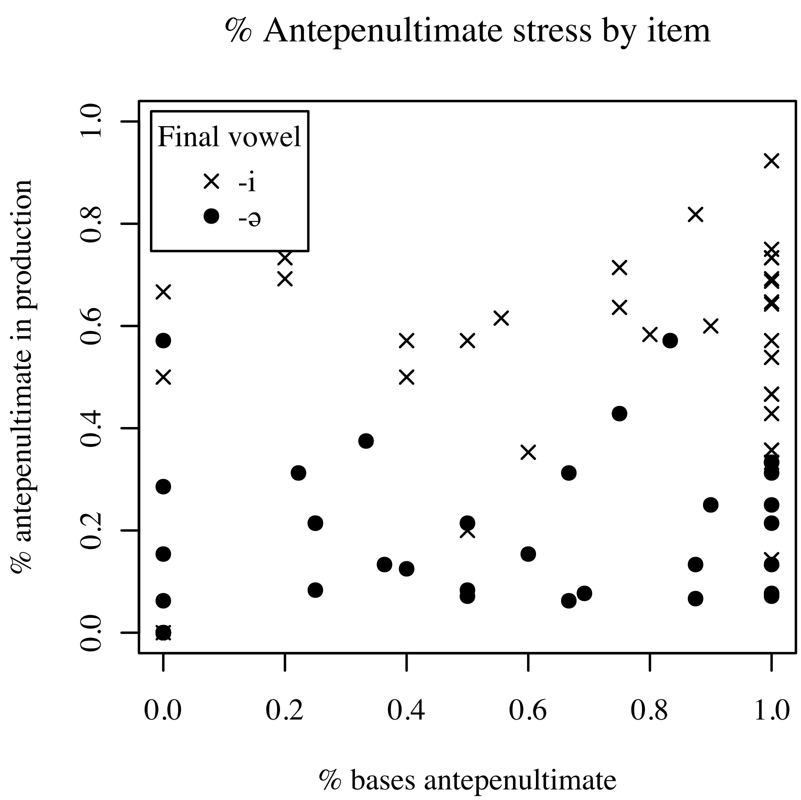
Table 6: Most common stress among the given analogical bases, for each experimental item. Items are not included if no given base for them was 3 syllables long or longer.

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

Figure 10 shows that there is no correlation between how many bases given for an item were antepenultimately stressed and how likely that item was to be given antepenultimate stress in the production task. Each point in the plot is one nonword item. Overall, i-final items take higher percentages of antepenultimate stress in production than ə-final items do, but items in both categories vary widely in the percentage of antepenultimate stress among their analogical bases. Whenever stress was not antepenultimate it was almost always penultimate, so the relationship between the two rates of penultimate stress is similarly noncorrelative. There were just a few given analogical bases which were neither antepenultimate nor penultimate, but pre-antepenultimate. Ten nonword items got at least one pre-antepenultimate base - 8 i-final and 2 ə-final, but only one got majority pre-antepenultimate bases, namely [dɛ lɛ si], for which 8 participants wrote 'délicacy'.

To summarize the findings of the analogical base task: The analogical bases provided by participants in aggregate followed the lexical trends: i-final bases were antepenultimately stressed more often than ə-final bases were. This same trend was obeyed in production, but there is no evidence that participants were

Figure 10: For each nonword item: the percentage of provided bases which had antepenultimate stress vs. the percentage of productions of that item with antepenultimate stress. There is no correlation between these two percentages.



using their lexicon directly to perform the production task. Individual participants' choice of analogical base did not correlate with their choice of stress on that item in production, and the composition of the set of analogical bases given for an item did not predict the distribution of stress patterns on that item in production.

5.4 Discussion

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

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

In the analogical base task, participants are explicitly asked to retrieve a real word based on the given nonword input. In this case, the preference for finding real words with antepenultimate stress may be due to the importance of a word's initial syllable in lexical retrieval (Nooteboom, 1981; Horowitz et al., 1968). Participants may be more likely to find actual words whose initial syllable matches the stimulus's first syllable. Since the stimulus's first syllable was always stressed, and contained a full vowel, the retrieved word would then typically be stressed on the first syllable.

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

The preference for antepenultimate stress on i-final words is probabilistic - participants did not observe it with every production, but rather they observed it in aggregate 77% of the time, a rate which is comparable to the generalization's rate of observance in the lexicon of English. Recall (Figure 2) that in the lexicon i-final words took antepenultimate stress between 89% (monomorphemic words) and 98% (morphologically complex words) of the time.

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

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

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

On the other hand, the lexical access process on a nonword may be non-deterministic. In this case, there would not be a single lexical item which becomes activated each time a participant heard a particular stimulus, but instead all words in the lexicon would have some probability of being accessed upon each exposure to the stimulus. Most of the probability would fall on some relatively small set of lexical items, though - words which are similar to the nonword in some way. If a nonword's set of most-similar words contained mostly antepenultimately stressed words, participants would be most likely to access an antepenultimately stressed word, and therefore give that nonword antepenultimate stress. If that set contained mostly penultimate stress, they would be more likely to give the nonword penultimate stress. In the analysis section, I examined the set of real words provided by participants for each item in Group 2[2]. The percentage of provided analogical bases with a particular stress pattern did not predict the rate of occurrence of that stress pattern in production.

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

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

The nonword items in this experiment were specifically designed to not be particularly similar to any actual words. All had sparse neighborhoods according to the generalized neighborhood model (Bailey and Hahn, 2001), and participants in the experiment had trouble coming up with words in the analogical base task. The average number of valid responses per item in this task was 10, even though 32-33 participants saw each item. These results therefore do not rule out the possibility that participants' *could* analogize to a particular lexical item in producing a nonword if the nonwords were constructed so as to be closer to some particular lexical item, or to have dense neighborhoods. Baker and Smith (1976) specifically test nonwords that are very close to actual words (the example they give is 'cinempa'), and find in fact that the behavior of nearby words affects the behavior of nonwords more than things like syllable weight and part of speech (specifically, noun vs. verb). In an experiment very similar to the one presented here, Guion et al. (2003) found with two-syllable nonwords that the stress pattern of participants' choice of analogical base predicted their choice of stress on a nonword, but not as reliably as syllable weight and part of speech. Guion et al neither specifically manipulated their nonwords' similarity to particular words nor specifically designed their words to be dissimilar to any actual words. In general, two-syllable words will have denser neighborhoods than three-syllable words, just because there are many more two-syllable than three-syllable words in English. Thus, Guion et al's nonwords likely had denser neighborhoods than the nonwords used in the experiment presented here.

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

6 Grammatical representation of the i-final trend

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

Main stress in English typically occurs within a three-syllable window at the right edge of the word. This is predicted by the joint action of an alignment constraint demanding that main stress be as close as possible to the right edge of the word (Align-R), and a nonfinality constraint demanding that the final syllable of a word be unfooted (Foot-NonFinality) (Pater, 2000, p. 240, Alcántara, 1998, p. 120-121).

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

(3) Foot-NonFinality (NonFin): Assign a violation if the final syllable of the word is parsed into a foot.

Figure 11: Violations assigned by Align-R and NonFin to several candidate stress patterns.

/σσσσ/	NonFin	Align-R
→ a. $\sigma\sigma\sigma(\acute{\sigma})$	1	
(→) b. $\sigma\sigma(\acute{\sigma}\sigma)$	1	1
→ c. $\sigma\sigma(\acute{\sigma})\sigma$		1
(→) d. $\sigma(\acute{\sigma}\sigma)\sigma$		2
e. $\sigma(\acute{\sigma})\sigma\sigma$		2
f. $(\acute{\sigma}\sigma)\sigma\sigma$		3
g. $(\acute{\sigma})\sigma\sigma\sigma$		3

In Table 11, candidates d-g violate Align-R too many times and are harmonically bounded by candidate c, while candidate b violates both NonFin and Align-R and is harmonically bounded by both a and c. However, candidates a and c, each with a single-syllable foot, will only satisfy FootBinarity when the foot contains a single heavy syllable. In words with light penultimate and final syllables, candidates b and d will not be harmonically bounded, since they satisfy FootBinarity.

In Hayes (1995), English is analyzed with moraic trochees, which can consist of a single heavy syllable (Ḣ) or two light syllables (ĹL) but not of a single light (*Ĺ) or a heavy plus a light (*ĹH, *ḢL). Foot-Binarity penalizes these disallowed foot shapes. Kager (1999) discusses several examples of other stress patterns for which such a constraint is necessary.

(4) Foot-Binarity (FtBin): Assign a violation to any foot which does not contain exactly two moras (which is not binary).

If this constraint is high enough ranked, parses such as a, c, and e in Table 11 would only be allowed if the syllable bearing the foot is heavy, whereas parses like b and d would only be allowed if the two syllables in the foot were both light. Words of shape LHL could only be parsed with a foot on the penultimate syllable: L(Ḣ)L, *(Ĺ)HL, *(ĹH)L. Words with only light syllables, such as those in the experiment, could only have penultimate stress with a right-aligned (ĹL) trochee, as in b, or antepenultimate stress, as in d. Penultimate stress would be preferred when Align-R ≫ NonFin, and antepenultimate stress would be preferred when NonFin ≫ Align-R.

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

(5) Foot-NonFinality-i (NonFin-i): Assign a violation if a word-final [i] is parsed into a foot.

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

Figure 12 shows an OT grammar which predicts that ə-final words should vary between antepenultimate and penultimate stress, and that i-final words should take antepenultimate stress only. In this grammar, FtBin is high-ranked, ruling out candidates with a single light syllable as a foot. Align-R and NonFin are unranked with respect to each other, but are both outranked by NonFin-i. For i-final inputs, only antepenultimate stress is allowed, but for inputs with other final vowels, antepenultimate and penultimate stress are equally grammatical.

Participants' behavior in the experiment was stochastic - they produced both penultimate and antepenultimate stress on nonwords. There was a slight preference for antepenultimate stress in ə-final items, and a strong but still probabilisitic preference for antepenultimate stress in i-final items. Based on the results

Figure 12: OT grammar predicting that ə-final words should vary between antepenultimate and penultimate stress, while i-final words should only take antepenultimate stress. Only inputs with light penultimate syllables are considered here. 'X' indicates either a light or a heavy initial syllable, while 'L' indicates a light penult.

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

of the analogical base task, I argue against a model in which the distinction between final [i] and final [ə] obtains because participants analogize to particular real words when choosing nonword stresses. I argue that instead, participants' phonological grammar must be able to represent the i-final trend probabilistically.

In Maximum Entropy grammar (Goldwater and Johnson, 2003), interacting constraints predict a probability distribution over candidate outputs. Constraints are not strictly ranked but are assigned weights. The negative dot product of those weights with the violation profiles of each candidate is the harmony score ($\mathcal{H}$) of the candidate. A probability distribution can then be calculated over those candidates using the exponential of $\mathcal{H}$, regularized over all candidates for a given input. Figure 13 shows a set of weights over the four constraints in Figure 12 which predicts a 50-50 distribution over penultimate and antepenultimate stress for ə-final inputs with a light penultimate syllable, and a 70-30 distribution in favor of antepenultimate stress for i-final inputs with a light penultimate syllable.

In this tableau, FtBin has a high weight, so that candidates with a foot consisting of a single light syllable are assigned a vanishingly small probability. The other constraints all have equal weight, but candidates with a final [i] parsed into a foot incur a violation of both NonFin and NonFin-i, leading them to have a lower harmony score than candidates which have a final ə parsed into a foot. If the weight on NonFin-i were lower, the distribution over i-final candidates would be more similar to the distribution over ə-final candidates (if its weight were zero, the two distributions would be equivalent).

If NonFin were given a weight higher than the weight of Align-R, then the model would predict a higher probability of antepenultimate stress than penultimate stress for both types of input. The greater

Figure 13: MaxEnt grammar predicting a 50-50 distribution over antepenultimate and penultimate stress for ə-final words, and an approximately 70-30 distribution for i-final word, favoring antepenultimate stress. $\mathcal{H}$ indicates harmony scores for each candidate, while P indicates the model's predicted probability for that candidate. Only inputs with light penultimate syllables are considered. 'X' indicates either a light or a heavy initial syllable, while 'L' indicates a light penult. Weights are listed immediately beneath constraint names, and are fitted by hand. Note that probabilities listed as zero are actually nonzero but very very small (on the order of 10^{-5}).

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

the difference between the two weights, the greater the preference for antepenultimate stress. Likewise, if Align-R had a higher weight than NonFin, the model would predict a higher probability on penultimately stressed candidates.

Each actual word of English has a single correct stress pattern, and does not vary: *cánnery* is always pronounced with antepenultimate stress, and *canáry* is always pronounced with penultimate stress, despite the fact that they both have a phonological shape for which the grammar predicts 70% antepenultimate stress. In a MaxEnt framework, a high weighted faithfulness constraint can ensure that all real words are pronounced correctly. One must assume that all words in the English lexicon are underlyingly specified for their stress pattern. Only nonwords with no prespecified stress pattern are subject to within-word variation. Figure 14 includes a high-weighted faithfulness constraint, and illustrates the behavior of a real word input, specified for an exceptional stress pattern compared to the behavior of a nonword input in which stress is not specified.

In typical language use, inputs with no specified stress pattern are the exception - the vast majority of words uttered are from the lexicon and are specified for their stress pattern. The most important fact for a learner of English to acquire is that Faith-Stress is high weighted, and that individual stress patterns must be stored. Given this, one might wonder what is the mechanism by which the markedness constraints and their relative weightings are acquired. Zuraw (2000, 2010) examines a similar case, in which each

Figure 14: MaxEnt grammar with high weighted faithfulness, illustrating the different behavior of an input from the lexicon, whose stress pattern is pre-specified, and a nonword input without a specified stress pattern. Only inputs with no pre-specified stress pattern have variable outputs. FAITH-STRESS ensures that when there is stress in the input, it is preserved. Weights are listed immediately beneath constraint names, and are fitted by hand. Note that probabilities listed as zero are actually nonzero but very very small (on the order of 10^{-5}).

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

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

7 Conclusion

Models of phonological grammatical knowledge such as Maximum Entropy grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008; Hayes et al., 2009) represent the phonological grammar as a series of probability distributions over possible outcomes. A MaxEnt representation of a preference for antepenultimate stress would predict antepenultimate stress on a given output a majority of the time, but not 100% of the time. Whether that majority is 60%, 80%, or 99.9% is determined by the weights on the constraints involved. Such a system is capable of representing the probability matching behavior found in many wug-tests of probabilistic phonological patterns (e.g. Becker et al., 2011; Hayes et al., 2009; Ernestus and Baayen, 2003; Zuraw, 2000, 2010).

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

In the experiment, participants produced stress patterns on novel words, but they also gave similar real words for each novel word. If the process sketched in section 2 were responsible for participants' matching of the lexical trends, then the similar real words ('analogical bases') given by participants for each nonword item should match the stress pattern chosen for that item. No relationship was found between the stress pattern of the analogical bases and the stress patterns chosen in production, although both followed the lexical trends.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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