The metrical parse is coarse-grained: phonotactic generalizations in stress assignment

Paul Olejarczuk & Vsevolod Kapatsinski

University of Oregon

Abstract

Phonotactic generalizations can be computed at different levels of granularity, from strictly categorical (*blick, dwick* > **bnick, *lbick*) to fully gradient (*blick* > *dwick* > *bnick* > *lbick*). Phonotactics that target syllable structure indirectly affect weight-sensitive stress because they influence the metrical parse. This paper investigates the sensitivity of the English metrical parse to the granularity of medial onset phonotactics. We present two experiments that feature pseudowords with medial clusters varying in phonotactic legality, probability and sonority (e.g. *vatablick, vatadwick, vatabnick, vatalbick*). The metrical parse is inferred from stress assignment in production (Exp. 1) and stress preferences in perception (Exp. 2). The results of both experiments indicate that stress is sensitive to relatively coarse-grained onset phonotactics, despite apparent lexical support for more gradient generalizations. Vocabulary simulations reveal that this level of granularity arises from the relative learnability of different generalizations, reconciling the experimental results with the lexicon.

1. INTRODUCTION

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; Frisch & Zawaydeh, 2001; Hay, Pierrehumbert & Beckman, 2003; Vitevitch, Luce, Charles-Luce & Kemmerer, 1997). In many cases, these preferences have been attributed to differences in syllable structure. A classic example calls attention 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 accidental gaps and the completely impossible (*blick, dwick* > **bnick, *lbick*), as predicted by traditional phonological theory. 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 sequence type 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, Steriade, Lennertz & Vaknin, 2007; Luce & Pisoni, 1998; Pit & 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, Lennertz, Smolensky & Vaknin-Nusbaum, 2009; Coetzee, 2009; Coleman & Pierrehumbert, 1997; Frisch, Pierrehumbert & Broe, 2004; 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 phonemes, syllables, and syllabic constituents. Nonce forms featuring highly probable sequences of such units are repeated faster, remembered better, and judged as more natural-sounding than forms composed of rare sound combinations (Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Frisch, Large & Pisoni, 2000; Hay et al., 2003; Vitevitch et al., 1997). Phoneme identification is likewise biased in the direction of phonemes that have higher transitional probabilities given neighboring segments (Pitt & McQueen, 1998). Another, related way to operationalize 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; the size of this neighborhood correlates with well-formedness ratings and production accuracy (Arnold, Conture & Ohde, 2005; Bailey & Hahn, 2001; Hammond, 2004). The average edit distance to *n*-nearest neighbors has a similar effect on wordlikeness judgments and lexical decision latencies (Ohala & Ohala, 1986; Yarkoni, Balota & Yap, 2008). 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 17 times more likely than [dw] to begin a word[1].

The second factor associated with well-formedness of a word or (more generally) a syllable is the sonority profile of its onset. Sonority is an abstract property of natural classes that roughly correlates with their relative loudness (Parker, 2002). Several sonority scales varying in granularity have been proposed in the literature; a typical, coarse scale from Clements (1990) is shown in (1), with natural classes increasing in sonority from left to right:

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

Cross-linguistically, syllables tend to rise in sonority from edge to nucleus, with steep rises preferred through onsets and gradual falls favored over 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; Bell & Hooper, 1978; Jespersen, 1904; Selkirk, 1982; Sievers, 1881). According to the SSP, rising-sonority onsets are universally preferred over falling-sonority onsets.

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 forms (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. Whatever its ontological status, 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, unattested word onsets with falling sonority profiles are more likely to be misperceived with an epenthetic schwa than novel, flat-sonority onsets ([lbɪf] → [ləbɪf] > [bdɪf] → [bədɪf]), while rising-sonority onsets tend to be perceived veridically ([bnɪf] → [bnɪf]; Berent et al., 2007). This effect appears to hold even for speakers of languages which

---

[1] Comparison made with the online Phonotactic Probability Calculator (Vitevitch & Luce, 2004).

prohibit complex onsets altogether (Berent, Lennertz, Jun, Moreno, & Smolensky, 2008). In children’s productions, cluster reduction patterns appear to be motivated by the preservation of the best sonority profile available. When presented with a novel word like [fwɪm] in a picture naming task, English-speaking toddlers are likely to reduce it to [fɪm] rather than [wɪm], presumably selecting the form with the steeper onset rise (Ohala, 1999).

In summary, people's sensitivity to sound sequences clearly goes beyond the coarse possible/impossible distinction, as demonstrated in both metalinguistic tasks and in experiments designed to tap certain online processes. In some cases, the performance is captured by a straightforward projection of lexical statistics; in others, sonority appears to be a useful cover term. Given this sensitivity to gradience, an interesting question arises regarding the interface of phonotactics with the rest of phonological knowledge. Namely, does all of phonology respond to fine-grained phonotactics, or are there phonological processes which rely on more coarse-grained phonotactic generalizations? And, if such processes exist, what is the reason for their insensitivity to finer detail?

We consider these questions by looking at the productive extension of weight-sensitive stress by English speakers. On most accounts, weight sensitivity implicitly entails sensitivity to phonotactics, because phonotactics are involved in determining syllable structure (Clements & Keyser, 1983; Hooper, 1975; Kahn, 1976; Selkirk, 1982; but see Blevins, 2003; Steriade, 1999). In other words, weight-sensitive stress assignment relies in part on the metrical parse. Conversely, the metrical parse can be inferred (by both the analyst and the learner) on the basis of stress assignment. In the present paper, we exploit this link in order to examine the 'phonotactic resolution' of weight sensitivity. To the extent that English stress is sensitive to phonotactics, what is the granularity of the relevant phonotactic generalizations? Is stress assignment guided by coarse-grained phonotactics reminiscent of classical, categorical phonology, or does it respond to the same level of phonotactic detail that underpins many other areas of language processing?

We approach the problem by investigating how speakers stress pseudowords in production, and how listeners respond to these words in perception. This method rests on the assumption that productive extension to novel forms recruits grammatical knowledge and is therefore an appropriate probe of its structure. Our focus is on those phonotactic generalizations which target syllable onsets. We embed biconsonantal clusters in the nonce probes, and treat the relative well-formedness of potential onsets as predictors of stress assignment. Following the work summarized above, we examine two potential sources of gradience: lexical support and sonority[2].

In the remainder of this section, we review the relevant facts of English stress and spell out our hypotheses about its relationship to phonotactics. The rest of the paper is organized as follows: Section 2 presents a nonce word production study and Section 3 follows with a 2-alternative, forced-choice (2AFC) judgment task using the same stimuli. Section 4 discusses the results, considering several alternative explanations before settling on an account that relates the findings to the lexicon via vocabulary simulations. Section 5 briefly concludes the paper.

1.1 English stress and the metrical parse

The role of syllable weight is widely acknowledged in formal descriptions of English word stress (Halle, 1998; Halle & Vergnaud, 1987; Hayes, 1980, 1982, 1995; Liberman & Prince, 1977; Prince, 1990). The traditional approach holds that in non-final syllables, stress assignment is sensitive to a binary weight distinction carried by the rime: light syllables end in a short vowel (C_0V) and so carry a single mora, whereas heavy syllables are made bimoraic by a long vowel, coda consonants, or both (C_0VX). In weight-sensitive systems, heavier elements preferentially attract stress, and in the case of English this is clearly illustrated by the well-known Latin Stress pattern, where the main stress in trisyllabic and longer nouns of Latinate origin tends to fall on the penult if it is heavy, else it falls on the antepenult. Under one influential version of metrical theory, Latin Stress follows from the interaction of foot type, alignment and extrametricality: bimoraic trochees are constructed right-to-left, skipping the final syllable unless superheavy (C_0VVX); main stress is then assigned to the head of the rightmost foot (Hayes, 1980, 1995; Prince, 1983). By way of example, consider the words *stamina* and *cicada*, which feature CV and CVV penults, respectively. Their metrical parses are shown below (by convention, syllable boundaries are indicated by periods, feet enclosed by parentheses and extrametrical elements contained within angle brackets).

(2) (ˈstæ.mɪ.)<nə>   sɪ.(ˈkeɪ.)<də>

---

[2] These two factors are correlated, but their influence will be teased apart in what follows.

As seen in (2), 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 assignment straightforwardly follows from the assumption that prominence is projected by foot-heads.

Because English foot construction depends on syllable weight, the computation of Latin Stress is dependent on factors that determine syllable structure[3]. Putting aside the controversial issue of ambisyllabicity, the parses of *stamina* and *cicada* are relatively straightforward: most syllable theories consider null onsets to be marked (e.g. Itô, 1989), so single intervocalic consonants parse with the following vowel as shown in (2). In contrast, consider the set of familiar, nonsense monosyllables, this time identically prepended to place the onsets in medial position: {*vatablick, vatadwick, vatabnick, vatalbick*}. What is the appropriate metrical parse of each form? This question is crucial for Latin Stress assignment — as long as the penult features a short vowel, its weight is entirely contingent on the syllabification of the cluster:

(3) (ˈva.ta.)<blɪk> vs. va.(ˈtab.)<lɪk>

One commonly accepted answer is that syllabification of intervocalic clusters is informed by the so-called 'phonotactic legality principle', a categorical restriction which relates syllable edges to word edges: if a cluster does not begin (or end) a word, it cannot begin (or end) a syllable (Treiman & Danis, 1988). According to this principle, the clusters in *vatabnick* and *vatalbick* obligatorily receive a coda-onset parse, resulting in heavy penults. All else being equal, if English speakers extend this level of phonotactic generalization to the problem of stress assignment, a wug test should yield identical rates of penult stress close to 100% in such words. In items like *vatablick* and *vatadwick,* the rates of penult stress should be lower, but it is unclear how low — both the tautosyllabic and the split parse shown in (3) are phonotactically valid, and word division studies show that speakers do not always maximize legal complex onsets (Eddington, Treiman & Elzinga, 2013). One way to test if stress assignment follows a maximal onset parsing strategy would be to compare the rate of penult stress in *vatablick* and *vatadwick* to matched items featuring embedded singletons (e.g. *vatabick*).

---

[3] Although this example is presented in derivational terms, the interdependence of stress and the metrical parse is acknowledged in constraint-based approaches as well.

An interesting alternative to the coarse-grained legality principle is that the metrical parse is probabilistic, with stress assignment reflecting the gradient well-formedness of potential onsets. This account, suggested by the findings reviewed above, predicts that [bl] should be more resistant to a split parse than [dw] due to more robust lexical support in onset position. It also predicts a cohesion asymmetry between [bn] and [lb] on the basis of sonority. If English stress assignment follows this type of parse, a wug test should reveal a gradience in penult stress rates:

(4) *vatablick* < *vatadwick* < *vatabnick* < *vatalbick*

Is there empirical evidence from English for a stochastic parser based on gradient cluster phonotactics? Most of what is known about the representation of syllable boundaries comes from metalinguistic tasks, including written word division and oral word games that require partial repetition, reduplication, permutation or infixation. The general findings from these studies appear to support the legality principle: medial clusters that form illicit word onsets are split at near-categorical rates (Eddington et al., 2013; Fallows, 1981; Redford & Randall, 2005; Treiman & Zukowski, 1990). Phonotactically legal complex onsets are more likely to be preserved, but this tendency is not categorical: in the largest English word division study to date, Eddington et al. (2013) found that about half of such onsets were split (see also Redford & Randall, 2005), with juncture intuitions cued by factors like stress, vowel quality and coarticulation.

Despite these results, the conclusion that stress assignment must also attend to a coarse-grained parse might be premature. One reason for skepticism is that metalinguistic tasks may tap into decision strategies informed by factors besides syllable representations, such as word-level morphological processes or knowledge of prescriptive rules of written word division (Goslin & Floccia, 2007; Morais & Kolinsky, 1997; Pierrehumbert & Nair, 1995; Smith & Pitt, 1999; Titone & Connine, 1997; Treiman, Bowey & Bourassa, 2002). It is possible that such strategies may be less sensitive to gradient phonotactics. Results across the different tasks also correlate poorly with each other, at least in languages other than English (Bertinetto et al., 1994, 2007; Côté & Kharlamov, 2011), raising questions about validity. A second reason is that 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. In Redford (2008), 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 word division 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.

The detection of stochastic parsing strategies may thus require the use of a sensitive online task, or else a training period. There is good reason to hypothesize that stress assignment could follow such a strategy, because productive extension of weight-driven stress has been shown to be sensitive to structures beyond the predictions of standard metrical theory. For example, the results of wug tests suggest that both onset and rime complexity have gradient, cumulative effects on stress (Kelly, 2004; Ryan, 2011), challenging the traditional assumptions that English weight is binary and exclusive to the rime. The productivity of Latin Stress is also modulated by the structure of the final syllable (Domahs, Plag & Carroll, 2014) the identity of the final vowel (Moore-Cantwell, 2015), and word length (Ernestus & Neijt, 2008). In this paper, we extend the line of inquiry into the productivity of weight-sensitive stress, focusing on the influence of onset phonotactics on the metrical parse.

1.2 The present study

Guion, Clark, Harada, & Wayland (2003) presented English-speaking adults with pairs of isolated, stressed monosyllables varying in structure, and asked the participants to concatenate them into pseudowords. The elicited productions revealed that initial CVV syllables attracted stress more often than initial CV syllables, and the same asymmetry was observed in final CVVC vs. CVC structures. The production results were mirrored in a subsequent 2AFC preference task. The experiments described in the present paper rest on the assumption that follows from these findings: stress patterns elicited in nonce words can, under the right circumstances, reveal the metrical parse applied by the speaker. To be clear, we do not assume that syllable structure is the only (or even the most important) influence on pseudoword stress assignment. Several studies have revealed sensitivity to a variety of other influences, including lexical class, morphological structure, and analogy to existing words (Baker & Smith, 1976; Baptista, 1984; Guion et al., 2003). The present goal is not to adjudicate the relative strength of these factors, and we do not seek to offer a comprehensive model of stress assignment. Instead, we control for other influences and focus on the granularity of the metrical parse: to the extent that phonotactic knowledge affects stress, what is the nature and source of this knowledge?

This paper presents the results of two experiments. Experiment 1 elicited productions of trisyllabic pseudowords of the types exemplified by the set {*vatabick, vatablick, vatadwick, vatabnick, vatalbick*}. That is, the forms consisted of controlled context frames with different inserts. These inserts were either singletons, or else clusters of varying phonotactic probabilities and sonority profiles[4]. All of the items featured zero lexical neighbors, and average edit distances to real words were controlled as described below. We investigated the nature of the phonotactic generalizations involved in stress assignment with respect to the four independent hypotheses presented below (examples of predicted asymmetries in penult stress rates are shown in parentheses):

(5) *The hypotheses:*
H1: Stress is sensitive to the legality principle.
(*vatablick, vatadwick* < *vatabnick, vatalbick*)
H2: Stress is sensitive to phonotactic probabilities of attested onsets.
(*vatablick* < *vatadwick*)
H3: Stress in sensitive to sonority profiles of unattested onsets.
(*vatabnick* < *vatalbick*)
H4: Stress follows onset maximization.
(*vatabick* = *vatablick, vatadwick*)

Experiment 2 tested the extent to which the stress patterns observed in production align with perceived well-formedness. The subjects performed a 2AFC preference task where the trials

---

[4] The productions were elicited using orthographic prompts. English orthography is phonologically opaque, which is potentially problematic, since a penult vowel realized as tense would attract stress independent of cluster phonotactics. Guion et al. (2003) solved this problem through auditory presentation of monosyllables. In the present study, it was crucial to present the entire pseudowords unparsed, since the focus of the investigation was the parse itself. Because it is difficult to avoid perceptual cues to stress in an auditory presentation of a trisyllable, we chose to employ orthography and discard any problematic responses; our exclusion criteria are detailed below.

featured aurally presented, minimal stress pairs created from a subset of the items used in Experiment 1 (e.g. ˈ*vatablick* ~ *va*ˈ*tablick*). The same four hypotheses shown in (5) were considered; the aim was to investigate whether the phonotactic generalizations employed in production and perception are equivalent in granularity.

To anticipate the major findings, both experiments provide support for *H1*: stress assignment in production and perception was affected by coarse-grained onset phonotactics. In Section 4 we focus on the production results, which held two surprises in light of previous research. First, contrary to categorical treatment in word division studies, illegal clusters elicited relatively low rates of penult stress. Second, the speakers ignored a statistically significant dependency between cluster sonority and stress in the English lexicon. To investigate both discrepancies, we conducted vocabulary simulations inspired by Pierrehumbert (2001). The results support a link between granularity and learnability and argue for a frequency-matching account of Latin Stress.

2. EXPERIMENT 1

2.1 Method

2.1.1 Participants

Thirty-six INSTITUTION undergraduates took part in the experiment. All participants self-reported as monolingual, native speakers of American English with corrected-to-normal vision and no hearing impairments. All were enrolled in introductory psychology and linguistics courses and received course credit for participating. Data from six participants was 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 subjects were analyzed.

2.1.2 Stimuli

Target CC clusters and singletons were embedded in CVCV___VC context frames to create trisyllabic pseudowords for orthographic presentation. The clusters were divided into three types based on word-initial legality and sonority profile: *legal, illegal rise* and *illegal fall*. All of the *legal* clusters were composed of obstruents followed by sonorants and thus featured rising sonority. Obstruents also preceded sonorants in the *illegal rise* clusters; for *illegal fall* items, this order was reversed. Each of the cluster types featured 19 unique clusters; the *singleton* category featured 12 different obstruents. Table 1 lists all of the inserts.

Table 1. C(C) inserts used in the Experiment

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

Within types, each insert was featured in two unique frames. These frames were held constant across types, providing identical context. For example, *daka___uth* and *shepi___oph* took the same set of inserts, producing the following pseudowords: *dakadwuth, shepidwoph* (*legal*), *dakadmuth, shepidmoph* (*illegal rise*), *dakamduth, shepimdoph* (*illegal fall*), and *dakaduth, shepidoph* (*singleton*). This arrangement yielded 38 pseudowords per type, for a total of 152 target items. All of the stimuli are listed in Appendix 1.

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[5]. Because spoken word recognition may involve activation of competing embedded forms (McQueen, 2004), there was a potential for such forms to influence parsing and stress placement strategies in orthographically presented pseudowords. To examine whether stress placement cued by embedded words correlated with cluster type, a linear regression model was fit to the data. Comparison with a null model revealed that stress placement favored by embedded words was distributed evenly across the cluster types ($F(3,148) = .80, p = .49$).

In addition to the target items, 524 pseudoword fillers were created. Eighteen of these had the same CV structure as the target items but featured medial clusters with flat sonority profiles[6] (*sp,*

---

[5] 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).

[6] These were treated as fillers rather than additional cluster types because their frames were not shared by any other items.

st, sk, zb, zd, zg). The remaining 506 were randomly generated with Wuggy software (Keuleers & Brysbaert, 2010). These were either 1, 2, 4 or 5 syllables in length, created by concatenating legal English syllables of various structures.

2.1.3 Procedure

The experiment was administered in E-Prime 2.0 (Schneider, Eschman & Zuccolotto, 2002). Participants were seated alone in a quiet room in front of a computer screen. The stimuli were presented in black, lower-case font on a white background, randomly paired with images representing unique alien creatures. The subjects were told that the words represented the creature names. These instructions contextualized the stimuli as nouns, in an effort to control for the effect of interpreted lexical class on stress assignment (Baker & Smith, 1976; Guion et al., 2003). Trial order was pseudo-random, with each target item separated by four fillers of varying length in order to minimize sequence effects between trisyllabic metrical frames. The slides advanced automatically after a time interval of 5 seconds for the targets and 3-5 seconds for the fillers, depending on length. 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. A headset microphone was used to record responses for offline coding of stress placement and acoustic analysis.

2.1.4 Predictors

The influence of phonotactics on stress assignment was measured by a combination of categorical, ordinal and continuous variables. The categorical measure was *cluster type*, which featured 4 levels: *singleton, legal, illegal rise* and *illegal fall*. This predictor was meant to simultaneously evaluate the effects of onset maximization, the legality principle, and coarse sonority profile. The other variables were intended to provide additional measures of gradience within legal and/or illegal items: *sonority slope, word-initial phonotactic probability* and *word-average phonotactic probability*. These predictors are described below (see also Appendix 2).

Sonority slope captured both the direction and magnitude of each insert's sonority profile in more detail than *cluster type*. The measure was based on Jespersen's (1904) fine-grained sonority hierarchy, recapitulated in Table 2:

Table 2. Sonority values of natural classes

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 cluster 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 (see also Gouskova, 2004, and McGowan, 2009 for similar implementations).

For every *legal* and *singleton* item, *word-initial phonotactic probability* was calculated using the online Phonotactic Probability Calculator (Vitevitch & Luce, 2004). The calculator derived the values by first checking the frequency counts in Kučera & Francis (1967) for all words containing a given C(C) sequence in initial position, then summing the log-values of these frequencies, and finally dividing the result by the summed log-frequencies of all words that contained at least two (or one) phonemes. The raw values ranged from 0.0003 (*dw*) to 0.1024 (singleton *s*); these were log-transformed prior to the analysis.

Because word-initial probability cannot differentiate among initially unattested onsets, *word-average phonotactic probability* was also calculated for each cluster. This measure captured position-independent segment co-occurrence; the values were obtained from the Irvine Phonotactic Online Dictionary (IPhOD; Vaden, Halpin & Hickok, 2009), which is based on counts in the SUBTLEXus corpus (Brysbaert & New, 2009).

In addition to the predictors of interest, we also examined *analogical bias*, a nuisance predictor meant to measure similarity to real words. Analogy to lexical neighbors has been shown to outperform syntactic and semantic factors in predicting the distribution of stress in English noun-noun compounds (Arndt-Lappe, 2011). It has also been shown to influence stress assignment in nonce forms (Baker & Smith, 1976; Guion et al., 2003). The analogy measure used in the present study was based on Yarkoni, et al. (2008), where it was shown that average edit distance to the nearest 20 lexical neighbors is a good predictor of lexical decision latencies and pronunciation accuracy. The variant used in our study limited the number of neighbors to ten[7]. It was calculated on a database of trisyllabic word forms retrieved from the English Lexicon Project (Balota et al., 2007). The database was split into two lists: words stressed on the antepenultimate syllable (*n*=13,667), and those stressed on the penult (*n*=7,601). Each pseudoword in the stimulus set was then separately compared to each list using the *ald()* function from the *vwr* package (Keuleers, 2013) in R (R Development Team, 2014), which was set to return the average edit distance from the 10 nearest neighbors. Each item was thus assigned a separate score for similarity to antepenult- and penult-stressed words. Subtracting the latter from the former yielded a single value, a measure of *analogical bias* favoring penult stress. Because similarity to known words was largely controlled in the design of the stimuli by fully crossing frames with cluster types, the *analogical bias* measure likely reflected position-specific frequencies of the inserts.

2.1.5 Coding and Analysis

Stress was coded offline by the first author, who relied on loudness, duration, pitch movement and vowel centralization, all of which are known to serve as perceptual cues to English lexical stress (see Cutler, 2005 for a review). In the event of multiple productions within the 5 second response window, only the final production was considered. Responses were coded into five categories: antepenult stress, penult stress, final stress, ambiguous stress, and production error. A total of 4,560 response trials were recorded (30 participants x 152 items). Of these, 874 (19.2%) were coded as errors and excluded from the main analysis (these are analyzed separately in Section 2.2.2).

Of the 3,686 error-free responses, 159 (4.3%) 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, 168 items (4.8%) received final stress and 325 productions (9.2%) elicited ‘ambiguous’ judgments. These items were included in the reliability assessment presented in the following section; however, the main analysis was restricted to those productions where stress was clearly placed on either the antepenult or the penult. These amounted to 3,034 tokens, about 86% of the error-free productions.

---

[7] The stimuli were designed to avoid obvious similarities to known words and thus had sparse neighborhoods. We felt that any neighbors beyond the nearest 10 would be unrecognizable as such and unlikely to affect processing.

All analyses were performed in R, using mixed-effects regression models constructed with the *lme4* package (Bates et al., 2014). Categorical variables were modeled with logistic regressions fit by the glmer() function, which uses the Laplace approximation and derives *p*-values from the normal distribution. Continuous variables were centered and modeled with linear regressions fit by the lmer() function, which uses restricted maximum likelihood estimation. The *p*-values for these models were estimated by the *lmerTest* package (Kuznetsova, Brockhoff & Christensen, 2014), which relies on *t*-distributions with degrees of freedom derived by the Satterthwaite approximation. All mixed models featured maximal random effects (Barr, Levy, Scheepers, & Tily, 2013); unless otherwise specified, this meant random intercepts for subject and frame, and random by-subject and by-frame slopes for all predictors. Hierarchical model comparisons were performed with the anova() function, which relies on likelihood ratios and returns the χ^2 statistic. All planned comparisons featured Bonferroni-adjusted alphas. Additional details about individual model specifications are presented below.

2.1.6 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 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 by the first author, who made the final decision.

In addition to being subjected to inter-rater reliability, the coding was checked against two acoustic correlates: duration, and intensity[8]. 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 by the first author, who used Praat (Boersma, 2001) to visually inspect the waveforms and spectrograms. Segmentation followed criteria standard in the field (e.g. Klatt, 1976), with vowels defined by the presence of formant structure, fricatives by sustained, aperiodic energy, stops by closure, release and VOT, nasals by the presence of anti-resonances, and liquids by upper formant movements and changes in amplitude relative to neighboring vowels. For the vast majority of the items, the visual

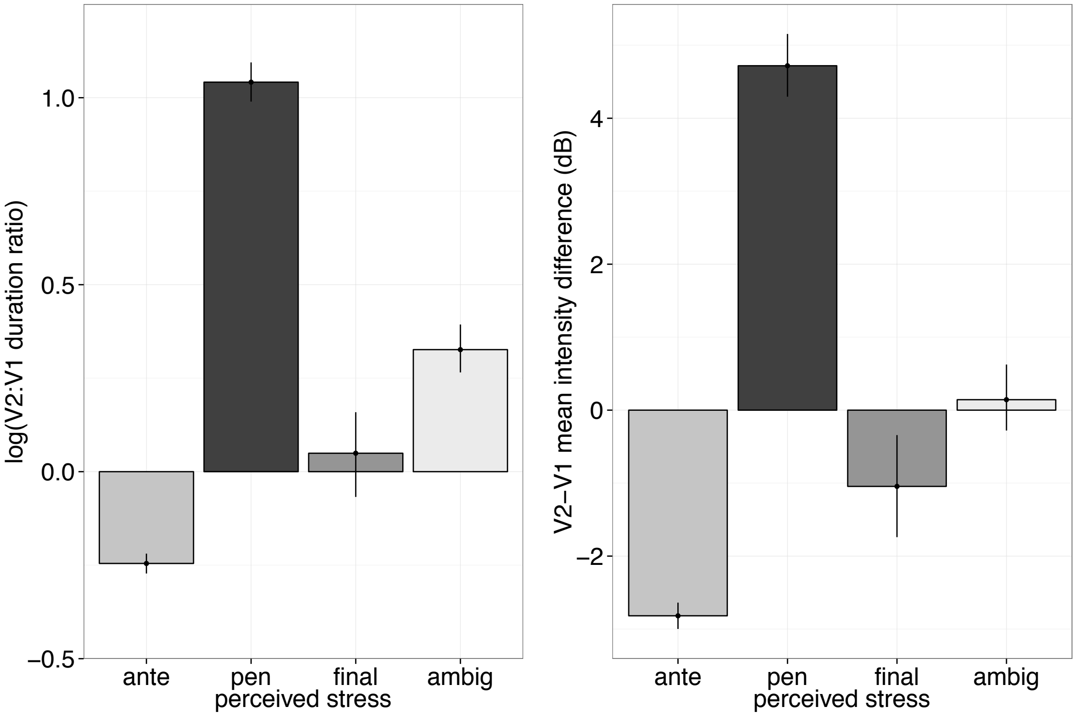
---

[8] Pitch was not used because a large proportion of the productions featured creaky phonation, resulting in unreliable tracking of F0.

information 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 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 *thanarbis* (stressed on the antepenult) would be transcribed in two ways: as [θænəɹbɪs] and [θænə˞bɪs]. Since the acoustic correlate measures relied on vocalic intervals, we 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, we arbitrarily report the measures derived from the segmentations that split coarticulated vowels and liquids at the midpoint of the sequence.

Figure 1 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, we 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. These duration ratios were then log-transformed, resulting in a normal distribution of values. As the panel shows, items coded as having penultimate stress featured longer penultimate vowels, whereas words perceived with initial stress had longer initial vowels. Note also that the ambiguous cases were intermediate on the measure.

Figure 1. Acoustic correlates by coded stress. Error bars are 95% CI.



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 (final stress was not of interest and was excluded from the model). The model significantly improved fit over a null model that featured only the random effects ($\chi^2(2) = 81.33, p < .0001$). 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 1 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[9]). 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

---

[9] Taking maximum as opposed to mean intensity values produced the same pattern of results, with even stronger effect sizes.

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: $\beta = 2.57$, $S.E. = .33$, $t(22.80) = 7.77$, $p < .0001$).

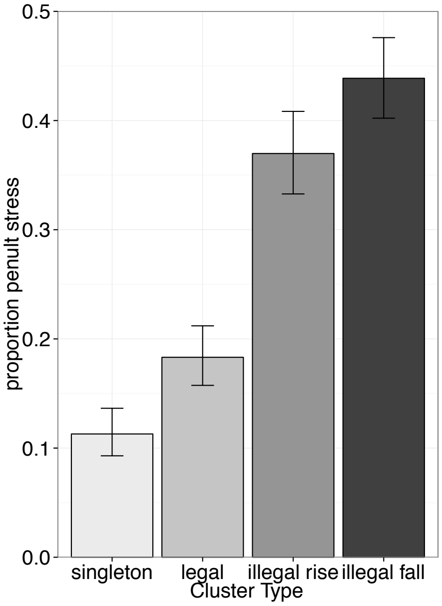
Taken together, the results of the reliability analysis indicate that the coders were consistent with each other in relying on duration and intensity, two of the acoustic correlates implicated in the realization and perception of English lexical stress. We now turn to the main results of the experiment.

2.2 Results

2.2.1 Stress assignment

Overall, 800 of the 3,034 valid responses (26.4%) were stressed on the second syllable. The distribution of penult vs. antepenult stress was modulated by *cluster type* as shown in Figure 2. For each type, the proportion of penult stress was as follows — *singleton*: 0.11; *legal*: 0.18; *illegal rise*: 0.37; *illegal fall*: 0.44.

Figure 2. Proportion penult stress by cluster type. Error bars are 95% CI.



To test for the significance of *cluster type*, a mixed effects logistic regression was fit to the data. This model significantly improved fit over a null, random-effects-only model ($\chi^2(3) = 45.54, p < .0001$). Table 3 provides the model output along with the R code used to construct it. Every *cluster type* received significantly more penult stress than the *singleton* reference level, indicating that the subjects were sensitive to syllable weight in assigning stress.

Table 3. Initial model output, penult stress by cluster type

	β	*S.E.*	*z*-value	*p*-value
Reference: Singleton	-3.03	.40	-7.53	< .0001 ***
Legal	.63	.31	2.04	< .05 *
Illegal Rise	2.27	.34	6.65	< .0001 ***
Illegal Fall	2.68	.33	8.14	< .0001 ***

glmer(penult_stress~ClusterType+(1+ClusterType|subject)+(1+ClusterType|Frame), family="binomial")

To test whether edit distance to known words contributed to stress assignment beyond *cluster type*, the model was expanded to include *analogical bias*. The addition of this predictor did not significantly improve fit ($\chi^2(14) = 18.75, p = .17$). Neither the main effect of *analogical bias* nor its interaction with *cluster type* were statistically significant[10] (all *p*s > .05). The expanded model was therefore abandoned in favor of the initial model without *analogical bias*.

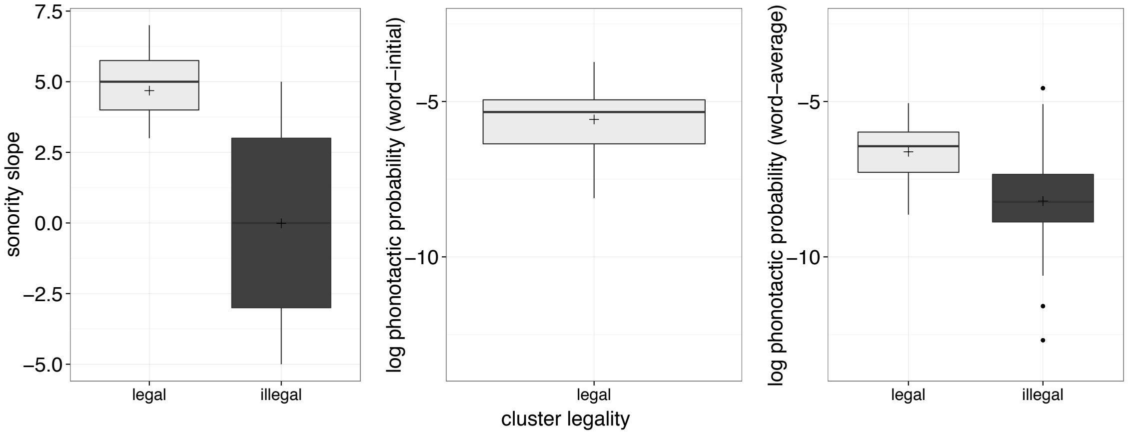
To test the extent to which the legality principle and coarse sonority influenced stress placement, pairwise comparisons were performed between the three non-singleton *cluster types*, again using mixed logistic regressions. The results indicated that *legal* items elicited significantly lower rates of penult stress than both *illegal rise* ($\beta = 1.57$, $S.E. = .27$, $z = 5.81$, $p < .0001$) and *illegal fall* items ($\beta = 2.03$, $S.E. = .33$, $z = 6.08$, $p < .0001$). The difference between *illegal rise* and *illegal fall* items was not significant ($\beta = .42$, $S.E. = .28$, $z = 1.50$, $p = .133$).

Turning to the continuous measures of *sonority slope, word-initial phonotactic probability* and *word-average phonotactic probability*, Figure 3 shows their distribution across cluster legality (*illegal rise* and *illegal fall* items are collapsed in the figure).

---

[10] A version of the expanded model without the interactions also did not improve fit over the initial model.

Figure 3. Distribution of continuous measures within legal and illegal cluster types



As seen in the figure, these measures correlated with legal status, which was already found to significantly predict stress. For this reason, the items were subset by legality and the continuous predictors were used to model responses within each subset. This way, it was possible to determine whether *sonority slope* and the two probability measures contributed additional gradience beyond cluster legality. None of the continuous measures significantly predicted stress placement within either data subset (all *p*s > .1).

2.2.2 Error analysis

Because production accuracy of word-initial consonant sequences may be related to their phonotactics, it is reasonable to hypothesize a similar relationship in medial clusters. To the extent that phonotactic well-formedness is implicated in both fluency and syllabification, the distribution of speech errors should resemble that of stress assignment. In other words, words with clusters that are more likely to be split should also be more prone to mispronunciation. This section presents the results of analyses that investigated this possibility.

There were 874 total production errors, including various cluster repairs, disfluencies and null responses. Table 4 provides a breakdown by error type.

Table 4. Typology of production errors

Error Type	Example	Count (%)
deletion (cluster C)	*tamapmish* → "tamapish"	89 (10.2)
deletion (V)	*tamapish* → "tampish"	5 (0.6)
deletion (other)	*lidigmeph* → "ligmeph"	1 (0.1)
epenthesis (cluster C)	*sipalbesh* → "sipalblesh"	61 (7.0)

epenthesis (V)	*sipalbesh* → "sipaləbesh"	99 (11.3)
epenthesis (other)	*sanankep* → "sansankep"	31 (3.5)
metathesis (cluster CC)	*sipalbesh* → "sipablesh"	43 (4.9)
metathesis (other)	*nepantep* → "neptanep"	47 (5.4)
substitution (cluster C)	*zepazriss* → "zepadriss"	54 (6.2)
pause/disfluency	*zepazriss* → "zepaz...driliss"	413 (47.3)
null response	*zepazriss* → "..."	31 (3.5)
TOTAL		874 (100)

As seen in the table, some of the errors were local to the clusters, while others involved larger parts of the words in addition to the clusters. Moreover, the former error type sometimes (but not always) resulted in repairing an illegal cluster. One could reasonably consider either that illegality of a cluster could increase the likelihood of all kinds of errors or that it would particularly increase the likelihood of cluster repairs. Therefore, two analyses were performed: one modeled the overall error rate while the other modeled structure-improving errors (repairs) only. Both analyses used the same set of predictors featured in modeling stress placement.

Beginning with total errors, the left panel in Figure 4 plots their proportion by *cluster type*. The values were as follows — *singleton*: 0.10; *legal*: 0.13; *illegal rise*: 0.30; *illegal fall*: 0.25. *Cluster type* significantly predicted total errors over a null model ($\chi^2(3) = 51.93$, $p < .0001$). Table 5 lists the model specification and output. All cluster items featured significantly more total errors than singletons, indicating that the longer words were more difficult to pronounce.

Figure 4. Production errors by cluster type, as a proportion of total trials. Left panel shows all production errors; right panel shows cluster repairs only. Error bars are 95% CI

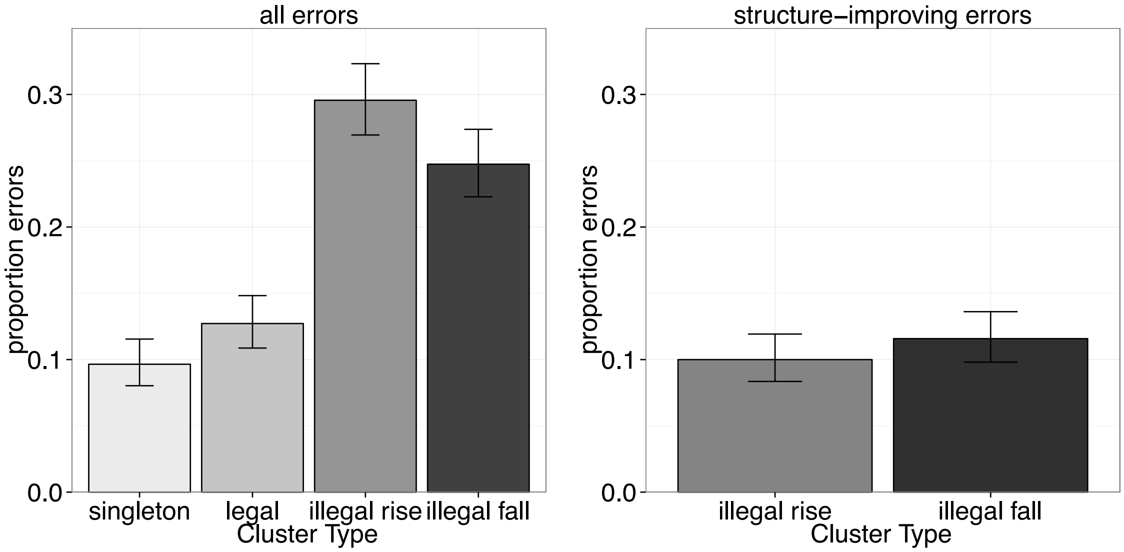


Table 5. Model output, total production errors by cluster type

	β	*S.E.*	*z*-value	*p*-value
Reference: Singleton	-3.04	.33	-9.32	$< .0001^{***}$
Legal	.52	.25	2.09	$< .05^{*}$
Illegal Rise	2.00	.27	7.47	$< .0001^{***}$
Illegal Fall	1.69	.26	6.54	$< .0001^{***}$

glmer(error~ClusterType+(1+ClusterType|subject)+(1+ClusterType|Frame), family = "binomial")

In order to test for the effects of legality and sonority, additional logistic regressions were used to perform planned comparisons between all non-singleton cluster types. The results indicated that *legal* items were significantly less likely to be mispronounced than either *illegal rise* ($\beta = 1.42$, *S.E.* = .19, $z = 7.53$, $p < .0001$) or *illegal fall* items ($\beta = 1.11$, *S.E.* = .18, $z = 6.18$, $p < .0001$). There was no significant difference between the two *illegal* cluster types ($\beta = -.31$, *S.E.* = .19, z = -=1.64, $p = .10$), and the numeric trend was in the opposite direction than that predicted by sonority sequencing.

Moving on to the continuous predictors, neither *sonority slope* nor *word-average phonotactic probability* explained additional variance in total production errors within either the *legal* or *illegal* word set (all *p*s > .05). *Word-initial probability* did reach significance, with more likely word onsets eliciting fewer production errors when embedded in medial position ($\beta = -.42$, *S.E.* = .19, $z = -2.16$, $p < .05$). With the exception of this predictor, total production accuracy appeared sensitive to the same phonotactic influences as stress assignment.

As for cluster repairs, these consisted of 245 out of 874 total errors. The right panel in Figure 4 above shows their distribution across the illegal items. The repair proportions were 0.10 for *illegal rise* and 0.12 for *illegal fall*; these did not differ significantly ($\beta = .06$, *S.E.* = .29, $z = .19$, $p = .85$). *Sonority slope* did not significantly predict repairs ($\beta = -.02$, *S.E.* = .04, $z = -.39$, $p = .70$). *Word-average cluster probability* showed a trend but failed to reach significance ($\beta = .15$, *S.E.* = .08, $z = 1.74$, $p = .08$).

2.3 Discussion

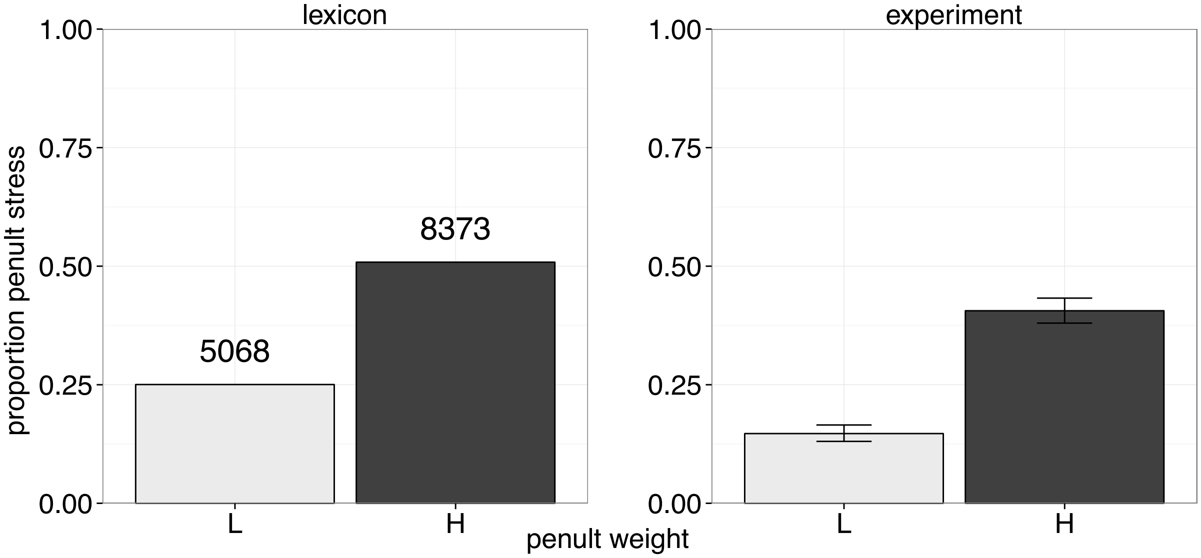
Of the four hypotheses introduced in Section 1.2, only *H1* was supported by the results of Experiment 1. The metrical parse governing stress assignment appeared to be guided by the legality principle. On the whole, pseudowords with embedded illegal clusters elicited higher rates of penult stress than did items with legal sequences. Neither sonority nor quantitative measures of lexical support accounted for additional variance in the data. There was a small but significant difference between *singleton* and *legal* items, suggesting that stress assignment did not always follow a maximal onset parse.

Qualitatively, these results agree with the findings from metalinguistic syllabification tasks reviewed in Section 1.1. However, the quantitative patterns are less consistent. Whereas in syllabification studies, illegal clusters are split at rates over 90% (Eddington et al., 2013; Fallows, 1981; Redford & Randall, 2005; Treiman & Zukowski, 1990), items containing such clusters received penult stress less than half of the time. This reluctance to stress the penult is also surprising given the dictionary counts reported in Moore-Cantwell (2015), where nearly all monomorphemic words of three or more syllables are stressed on the penult if it is heavy. One way to reconcile the present results with the lexicon is to posit that the speakers did not restrict their lexical search to monomorphemes, but included compounds and inflected forms when computing the stress generalization. We investigated this possibility by examining the stress patterns in all trisyllabic word-forms from the CMU Pronouncing dictionary (Weide, 1994, syllabified by Bartlett, Kondrak, & Cherry, 2009), filtered to exclude items unattested in the SUBTLEXus corpus. We ignored the marginal number of words that received final stress, focusing on initial and penult-stressed words to match the productions analyzed in Experiment 1. The left panel in Figure 5 presents the results of the search, with 13,441 total items. It appears that once polymorphemic words are included in the search, heavy penults are unstressed quite often. The right panel in the figure recapitulates the results of Experiment 1 for comparison, lumping all *illegal* items under the "H" (for heavy) penult category, and combining *legal* with *singleton* items under "L" (for light) to match the dictionary. The patterns are strikingly similar across the two panels in the figure[11], suggesting that stress statistics were projected from the lexicon and computed over all trisyllables.

---

[11] The penult stress rates elicited in the experiment are somewhat lower than the dictionary counts; this will be discussed in Section 4.

Figure 5. Penult stress rates in trisyllabic word forms in CMU dictionary (left; numbers indicate counts) and Experiment 1 (right; error bars = 95% CI).



The distribution of production errors in Experiment 1 closely resembled that of stress assignment. *Singleton* items were produced with the most accuracy, followed by stimuli containing legal clusters, which in turn elicited fewer errors than illegal items. There were no effects of sonority or *word-average phonotactic probability* on the rate of errors, although *word-initial probability* affected accuracy of legal items in the expected direction. The overall consistency across these results suggests that phonotactic generalizations of similar granularity underlie both the metrical parse and production accuracy -- bad clusters were either split or mispronounced. That said, the error results are only partly consistent with prior production studies, which found no sonority or statistical effects on error rates in novel word onsets (Davidson, Jusczyk & Smolensky, 2004; Davidson, 2006). The discrepancy may be due to the particular measures of lexical support: Davidson and colleagues investigated type and token frequencies of the clusters, whereas the present study used phonotactic probabilities.

3. EXPERIMENT 2

The aim of Experiment 2 was to assess whether the same relationship between phonotactics and stress that emerged in the spontaneous production task would also guide listener judgments of novel forms. Would items featuring ill-formed clusters sound better when stressed on the penult, indicating a coda-onset parse of these clusters? Would gradient onset phonotactics make a difference in perception? To this end, we administered a 2AFC task where participants heard pairs of pseudowords differing only in stress placement (ˈ*vatablick* ~ *va*ˈ*tablick*) and indicated their preference for one of the pair members. Prior work has shown that stressed syllables in known words attract codas in metalinguistic tasks). We therefore took the stress preferences to reflect implicit evaluations of the competing metrical parses.

The 2AFC task was similar to that employed in Guion et al. (2003) and Daland et al. (2011). There were two reasons why it was chosen instead of a Likert scale rating. First, we 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 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).

3.1 Method

3.1.1 Participants

Fifty-two INSTITUTION undergraduates were recruited to participate in the study in exchange for course credit. Seven individuals were excluded from the analysis: six due to fluency in another language, and one due to self-reported dyslexia. Data from the remaining 45 participants were analyzed. These subjects were all monolingual, native speakers of American English with normal hearing and normal-to-corrected vision.

3.1.2 Stimuli

Experiment 2 used half of the pseudowords from Experiment 1. All of the same inserts were represented, but only 19 of the 38 frames from Experiment 1 were retained (each insert thus appeared in a single frame instead of two). See Appendix 1 for the complete list of target items.

The stimuli were presented both orthographically and aurally. In the visual presentation, the items appeared exactly as in Experiment 1. The auditory stimuli were constructed as follows. The pseudowords were read in isolation by a phonetically trained native speaker of American

English, who pronounced each item in two ways: with either antepenultimate or penultimate stress. The mapping between orthography and pronunciation was kept constant across the stimuli, with all stressed vowels pronounced as lax and all unstressed vowels reduced to either [ə] or [ɪ] as appropriate. The speaker provided three productions of each minimal stress pair.

The pronunciations were digitally recorded in a quiet, sound-treated room using a condenser microphone. The middle production of each recording was excised and saved to a separate audio file, and the files were batch normalized in Praat to the same amplitude. Visual inspection of the waveforms confirmed the presence of F0, amplitude and duration cues to stress. A total of 76 pseudoword pairs were generated in this manner (19 frames x 4 cluster types).

3.1.3 Procedure

The experiment was administered using the same software and room setup as Experiment 1. The participants were presented with the pseudoword pairs over headphones at a comfortable listening level, with trial order randomized for each subject and the within-pair order of stimuli counterbalanced across listeners. Pair members were separated by 500 milliseconds of silence. Auditory presentation was accompanied by the appropriate orthographic form, which appeared 500ms after the offset of audio and stayed on the screen until the subject made a response. Trials were separated by 500ms. Each pair was presented once to each listener.

Participants were instructed to listen to each pair, consider the written form, and decide which pronunciation would be better if the word were to be introduced into the English language as a new noun. The subjects entered their choice by pushing a button on a serial response box.

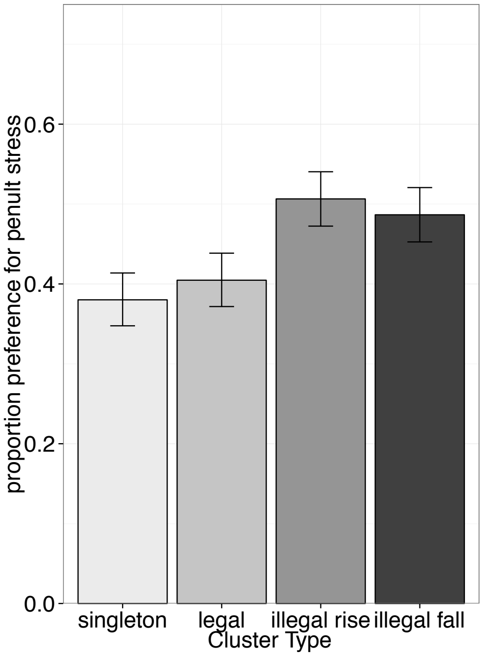
3.1.4 Analysis

The dependent variable in Experiment 2 was preference for penult-stressed items. Since this preference was binary, it was modeled with the same mixed-effects logistic regressions used in Experiment 1. The predictor set was also unchanged.

3.2 Results

The data consisted of 3,420 observations (45 subjects x 76 responses). Overall, participants preferred penult-stressed items 44.4% of the time. The stress preferences were modulated by *cluster type* as seen in Figure 6. For each *cluster type*, the proportion of the times the penult-stressed version was preferred was as follows — *singleton*: 0.38; *legal*: 0.40; *illegal rise*: 0.51; *illegal fall*: 0.49.

Figure 6. Penult-stress preference by cluster type. Error bars are 95% CI



A mixed-effects logistic regression evaluating *cluster type* as a predictor of penult preference significantly improved fit over a null model that featured only the random effects ($\chi^2(3) = 17.03$, $p < .001$). The output of the model is presented in Table 6. Both *illegal rise* and *illegal fall* items elicited significantly more preferences for penult-stressed versions than did *singletons*; there was no significant difference between *singletons* and *legals*.

Table 6. Initial model output, stress preference by cluster type

	β	*S.E.*	*z*-value	*p*-value
Reference: Singleton	-.55	.17	-3.32	< .001 ***
Legal	.15	.16	.92	= .36
Illegal Rise	.58	.17	3.50	< .0005 ***
Illegal Fall	.49	.19	2.65	< .01 **

glmer(penult_stress~ClusterType+(1+ClusterType|subject)+(1+ClusterType|Frame), family="binomial")

To test whether edit distance to lexical neighbors contributed to predicting stress preference beyond cluster phonotactics, the model was expanded to include *analogical bias*. The expanded model did not significantly improve fit ($\chi^2(16) = 11.72$, $p = .76$), and neither the main effect of *analogical bias* nor its interaction with *cluster type* emerged as significant[12] (all *p*s > .15). The expanded model was therefore dropped in favor of the original.

To test the extent to which legality and coarse sonority influenced stress preferences, planned pairwise comparisons were performed between the three non-singleton cluster types. The results reveal that *legal* items were significantly less likely than either *illegal rise* (β = .43, $S.E.$ = .14, z = 3.01, p < .005) or *illegal fall* items (β = .34, $S.E.$ = .10, z = 3.33, p < .001) to elicit preferences for penult-stress. There was no significant difference between the two *illegal* cluster types (β = -.08, $S.E.$ = .18, z = -.47, p < .64).

Moving on to *sonority slope* and the two phonotactic probability measures, none of these predictors explained additional variance in judgments within either *legal* or *illegal* items (all *p*s > .25).

3.3 Discussion

Experiment 2 sought to evaluate the extent to which perceptual preferences mirror production with respect to phonotactic influences on the syllabification of medial clusters. The overall pattern of results was similar to Experiment 1. When the frames contained embedded clusters that were initially attested, subjects were less likely to prefer the penult-stressed version than when the items contained illegal clusters. As in Experiment 1, there was no effect of sonority, *word-initial phonotactic probability*, or *word-average phonotactic probability*, indicating that as in production, preferences were based on coarse phonotactics that distinguished the clusters along the lines of the legality principle.

In spite of the overall similarities, the results diverged from Experiment 1 in two related ways. First, whereas Experiment 1 found that *legal* items attracted more penult stress than *singleton* items (contra onset maximization), no significant difference between these item type*s* emerged in Experiment 2. Second, while penult-stressed forms were volunteered only 26.4% of the time in the production task, they were chosen as better 44.4% of the time in Experiment 2. A comparison of Figure 6 with Figure 2 reveals that this was mostly driven by *singletons* and *legals*, whose penult-stressed versions were accepted at much higher rates than they were offered.

---

[12] Removing the interactions from the expanded model also did not improve fit over the original model.

One possible explanation for this difference is perceptual noise — subjects may have had difficulty perceiving the difference between the penult- and antepenult-stressed productions they were asked to compare. This was not the case, however — immediately following the judgment task, the subjects participated in a learning study (Anonymous, forthcoming), 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 cannot therefore be attributed to misperception. One possible alternative lies in the nature of the 2AFC task, which provides the subject with a closed set of alternatives to choose from. Closed-set tasks have been argued to reduce listener sensitivity to phonetic variability and lexical neighborhood effects during spoken word recognition (Sommers, Kirk, & Pisoni, 1997). It is possible that providing the illicit forms essentially primed them, boosting their acceptability. Such effects have been reported in syntactic acceptability judgments (Luka & Barsalou, 2005; Snyder, 2000). In addition, some of the difference may have been due to conflicting parse cues from stress and coarticulation. Specifically, illegal fall items that contained liquids sometimes featured velarized [l], regardless of stress pattern. This phonetic realization cued coda assignment, which came into conflict with the parse assigned by antepenultimate stress. However, there was no significant difference in penult preference between items with liquid-initial and nasal-initial clusters ($\beta = -.06$ *S.E.* $= .18$ $z = -.32$, $p = .75$), indicating that *illegal fall* items behaved as a group. Any difference due to conflicting cues was therefore unlikely to have a meaningful effect on the interpretation of the results.

4. GENERAL DISCUSSION

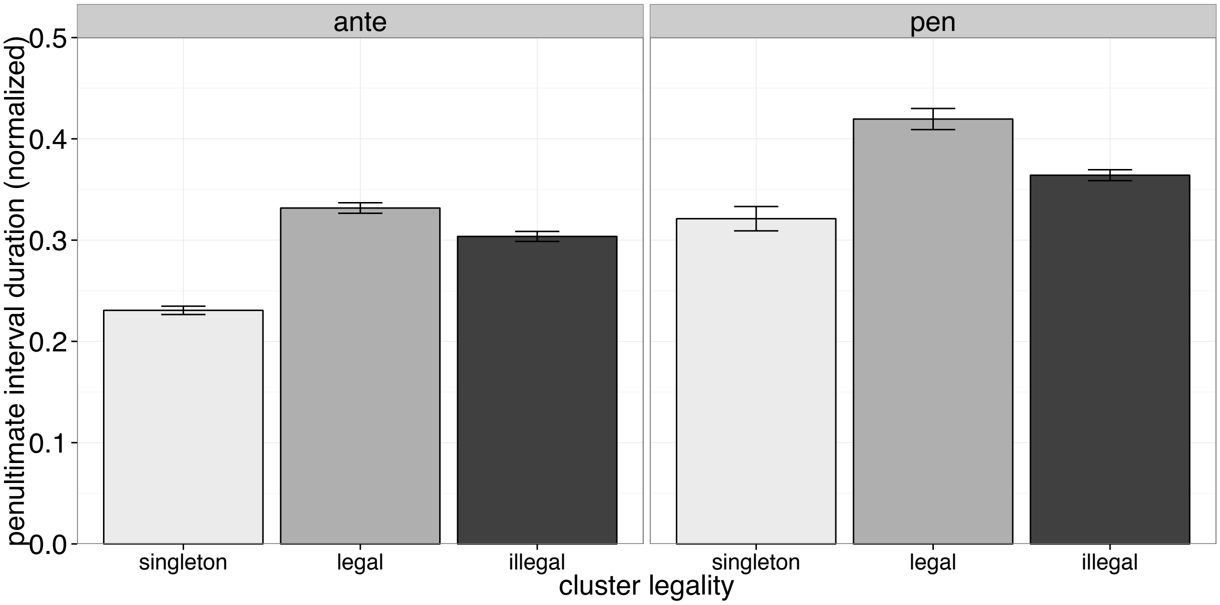
The process of assigning stress to unfamiliar forms involves consulting one's lexicon, either directly or via the grammar projected from it. Many sources of information potentially compete for the solution. One could in principle choose the overall most common stress pattern in the language, or else restrict the search in a number of ways — by lexical class, morphological structure, number of syllables, *n*-nearest neighbors, and so on. The design of the present study encouraged the subjects to employ those generalizations that make reference to the sublexical, structural description of a word — its division into light and heavy syllables. Our aim was to investigate the granularity of the phonotactic generalizations that govern the parse relevant to stress. Specifically, we were interested in the phonotactics of complex onsets, and the degree to which their word-initial well-formedness is paralleled by their word-internal cohesion. The four hypotheses outlined in Section 1.2 roughly map onto two distinct types of parsing model. One type is deterministic, with stress assignment categorically following the coarse-grained, phonotactic legality parse. This parsing model has been influential in phonological theory ever since Kahn (1976), and it appears to be employed in metalinguistic syllabification tasks (Eddington et al., 2013; Fallows, 1981; Redford & Randall, 2005; Treiman & Zukowski, 1990). The alternative model predicted stochastic stress assignment informed by intersegmental cohesion (Bertinetto, 2004; Bertinetto et al., 1994, 2007; Dziubalska-Kołaczyk, 2002, 2009). Support for the general notion of stochastic parsing has been suggested on the basis of well-formedness judgments and short-term memory tasks (Lee, 2006; Lee & Goldrick, 2008). Specific to complex onsets, the stochastic parser appears to play a role in word segmentation and phonotactic learning (Ettlinger et al., 2011, Redford, 2008). Based on prior work on word onsets, we asked whether the cohesion of intervocalic clusters was a reflex of their sonority profile and phonotactic probability.

The results of both experiments described in this paper clearly argue for the relevance of coarse-grained onset phonotactics to stress assignment. The legality principle was found to influence both the placement of stress in pseudowords and the well-formedness evaluation of these forms. No effect of cluster sonority or phonotactic probability emerged in either experiment. Of the two tasks, production was more sensitive, revealing that, for the purposes of stress assignment, there was a weak tendency against maximizing complex onsets.

Before discussing the implications of these findings, we address two alternative interpretations. The first is the possibility that stress assignment did not reflect a variable metrical parse, but rather gradient weight distinctions motivated by other factors. On this interpretation, all clusters received the same parse (resulting in, say CVC penults), with stress dependent on some property of the rime. A version of this argument is presented for Spanish 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 conclusion that CVG syllables are heavier than CGV syllables in Spanish. For this explanation to apply to our results, any weight distinctions would have to be derived from properties of the penult codas, since their onsets and nuclei were controlled across cluster types. However, an examination of Table 1 and Appendix 1 reveals that both *legal* and *illegal rise* items contained (usually identical) obstruents in C3 position, and yet the latter attracted significantly more penult stress than the former. Conversely, *illegal fall* items featured C3 sonorants, which are treated as heavier than obstruents in at least some languages (Gordon, Jany, Nash, & Takara, 2008; Zec, 1995), and yet no difference in stress between *illegal fall* and *illegal rise* items was observed. Given the stimuli used in the present study, the 'invariant parse - variable weight' explanation does not seem to hold.

A second alternative is presented by Interval Theory, which assumes that the domain of weight computation is not the syllable rime, but rather the total vowel-to-vowel interval (Steriade, 2008 *ms*; see also Hirsch, 2014). Intervals run from the onset of a vowel to the onset of the following vowel; an interval parse of our stimuli thus invariantly yields VCC penults (VC in the case of *singleton* items). Under this proposal, weight is computed by considering the acoustic duration of each interval. To account for the asymmetries in penult stress rates observed in Experiment 1, Interval Theory predicts that penult interval durations should fall along the scale *singleton* < *legal* < *illegal*. To test this prediction, we measured the duration of each penult interval, normalizing it by word duration. Figure 7 plots the results for the entire set of valid responses (3,034 productions). In order to remove the influence of stress on duration, the ratios are plotted separately for antepenult- and penult- stressed items.

Figure 7. Normalized penultimate interval durations by cluster legality and stress



As seen in the figure, *singleton* items featured the shortest penult intervals. However, *legal* and *illegal* words patterned in the opposite direction than predicted by Interval Theory. Mixed models with random intercepts for subjects and frames revealed that the VCC penult intervals were longer in *legal* than *illegal* items for both stress locations (both *p*s < .0001). Since interval durations cannot account for the observed results, we conclude that stress assignment likely reflected variability in cluster parsing.

Although the metrical parse appeared to be coarse-grained, stress assignment was markedly less categorical than the parsing behavior observed in overt syllabification tasks. Here, the difference can be ascribed to lexical statistics: a cursory comparison of the production results with dictionary counts in Section 2.3 showed evidence of probability matching of Latin Stress. Similar behavior has been reported in prior studies, where categorical syllable boundaries were assumed a priori (Domahs et al., 2014; Ernestus & Neijt, 2008; Kelly, 2004; Ryan, 2011). Interestingly, when assembling the set of words from which to generalize Latin Stress, the subjects appeared to consider all word forms — morphologically simple as well as inflected, derived and compound. This may have been a consequence of the study design, since no manipulation attempted to restrict the lexical search to monomorphemic words. On the other hand, it may be the case that the search is broad by default. This possibility is supported by an overall tendency to undergeneralize penult stress from trisyllabic words (note that both bars in the right panel of Figure 5 are lower than those in the left panel).

One way to account for this under-generalization is to allow for some influence of shorter words, which are overwhelmingly stressed on the initial syllable (Cutler & Carter, 1987). In other words, reluctance to stress the penult may have been the result of competition from initial stress. Similar competition between stress patterns was reported by Turk et al. (1995), where 9-month-old infants showed preference for both strong initial syllables and strong heavy syllables. The difference between that study and the present results was in the outcome of the competition: whereas the infants studied by Turk and colleagues showed a strong initial bias with some weight sensitivity, the adults in Experiment 1 showed good projection of Latin Stress with some influence of initial bias. This difference may be related to the relative learnability of the two patterns: whereas initial stress is a simple, first-order generalization that maps prominence onto syllable position, Latin Stress is a more complex, second-order pattern where stress is contingent on a structural description of a word. Second-order phonotactic generalizations have been shown to be more difficult to learn in the lab (Warker & Dell, 2006) and in computer simulations (Pierrehumbert, 2001); it is possible that robust learning of weight-sensitive stress requires several years of exposure. Given this evidence, the right panel of Figure 5 could be interpreted as an aggregate outcome of stochastic competition between stress patterns in the adult lexicon.

If stress was indeed indicative of a metrical parse, the question remains why stress assignment should adhere to the coarse-grained and not the fine-grained parser. In other words, why did performance in the stress assignment task resemble performance in word division rather than speech segmentation? Here, our two hypothesized influences — lexical statistics and sonority — warrant separate discussion. With respect to the former, we caution that our results speak only to the phonotactics of potential onsets, leaving open the possibility that the parser could be stochastically guided by other measures of lexical support. A good candidate for such a measure is rime cohesion. It is well known that the strength of nucleus-coda associations varies continuously across VC combinations (Kessler & Treiman, 1997), and that English speakers are sensitive to this strength when recalling CVC pseudowords and judging their acceptability (Lee, 2006; Lee & Goldrick, 2008). These findings invite the hypothesis that stronger rimes should resist a heterosyllabic parse, attracting penult stress. Nevertheless, we chose to focus exclusively on onset phonotactics for two reasons. First, much of the work on gradient well-formedness has focused on word onsets under the implicit assumption that the findings generalize to internal syllables (e.g. Berent et al., 2007; see Treiman, et al., 1995 for a critique of this assumption). The explicit sonority and lexical support predictions we set out to test follow from this work. The second reason is methodological: because phonetic vowel quality often depends on stress, predicting stress assignment from VC statistics can be circular[13]. That said, we acknowledge that rime statistics may play a gradient role in the metrical parse and leave the question open for future investigation with more appropriate methods. What *can* be concluded here is that cluster

---

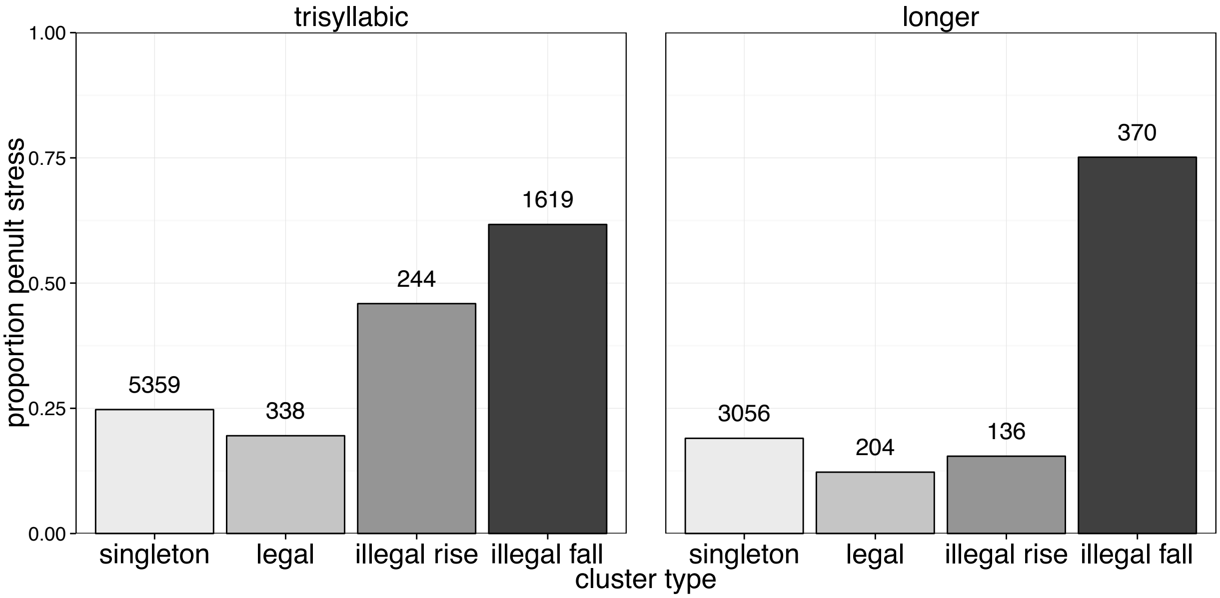
[13] Imagine a speaker who, when presented with the orthographic prompt *madaplazz*, produces [ˈmædəplæz]. Did the stress skip the penult because its rime /əp/ is statistically underrepresented (leading to the maximization of /.pl/), or did the penult vowel surface as [ə] because it was skipped by stress? Given that most of the unstressed vowels in produced Experiment 1 were phonetically centralized, this problem affects a large portion of our results. Focusing on the phonotactics of CC sequences, also independently motivated, allowed us to sidestep the issue. For what it's worth, we calculated various association measures for *orthographic* rimes, including transition probability and ΔP (both backward and forward), and Pearson's *r* (see Perruchet and Peerman, 2004, for discussion of these measures). Following Lee (2006), we based the calculations on the entire set of monosyllabic words in Kessler & Treiman (1997); none predicted the results of either experiment.

probability alone is insufficient to drive a stochastic metrical parse (see Kharlamov, 2009 for similar conclusions from Russian well-formedness judgments).

With respect to sonority, one recent proposal argues that its influence is dependent on the nature of the representations accessed by the experimental task. In a study investigating the perception of word onsets, Berent, Lennertz & Balaban (2012) found that sonority effects emerged in syllable counting ("how many syllables in *mdiff?*" yielded many "2" responses) but not in phoneme monitoring ("does *mdiff* contain *e*?" yielded more "no" responses). The authors argued for a 'processing levels' explanation, which hinged on the assumption that syllable counting involves phonological processing while phoneme detection taps phonetic encoding. The greater sensitivity of the former task to sonority profile was then taken as evidence that sonority is part of phonological knowledge. This kind of explanation is not compatible with our results — if both stress assignment and sonority-based generalizations are the domain of phonology, the wug test used in Experiment 1 is exactly the kind of task that should uncover a potential relationship between them.

Given that sonority-based stress is apparently not part of English speakers' knowledge, an interesting question is whether it is also absent from the lexicon. In other words, does the input offer a potential generalization that is being ignored by speakers? To investigate this question, we looked at Latin Stress in trisyllabic and longer word forms found in the CMU dictionary (filtered by SUBTLEXus frequency as described in Section 2.3). To match the relevant characteristics of the responses analyzed in Experiment 1, we constrained the search to words with (a) singletons and CC clusters between V_{penult} and V_{final} that matched the 4 insert types investigated in our study, (b) stress on either the penult or the antepenult, and (c) no stress on long vowels. Figure 8 shows the distribution of penult stress in the resultant 11,326 entries, divided into trisyllabic and longer words.

Figure 8. Distribution of Latin Stress in a subset of the CMU dictionary, by cluster type and word length (numbers indicate counts)



Among the longer words, there appears to be a clear sonority effect, with *illegal fall* items exhibiting a much higher rate of penult stress than *illegal rise* items. The later appear to pattern with *legal* words, which also have rising sonority profiles. Among the trisyllabic forms, the sonority effect is weaker, but still statistically significant: a mixed logistic regression model with random intercepts for word revealed that *illegal fall* items featured significantly more penult stress than *illegal rise* words ($\beta = 1.01$, $SE = .25$, $z = 4.22$, $p < .001$).

The CMU dictionary counts thus suggest that English speakers ignore a statistical pattern present in the lexicon[14]. Missed generalizations have been reported elsewhere in the phonological literature. For example, Becker, Ketrez & Nevins (2011) showed that Turkish speakers do not internalize a statistical dependency between stem-final laryngeal alternations and the quality of the preceding vowel. The authors argue that such a dependency is phonologically unmotivated because the grammatical architecture (in that case, Optimality Theory) does not encode the interaction of vowel and laryngeal features in a straightforward way. They conclude that a set of analytical biases shaped by this architecture (i.e. universal grammar) acts as a hard filter on learnability. It is unclear whether such an explanation can be extended to the present

---

[14] The difference between *singleton* and *legal* items shown in the figure is not significant, but it trends in the opposite direction from the results of Experiment 1. We do not at this time have an explanation for why our subjects did not maximize legal onsets.

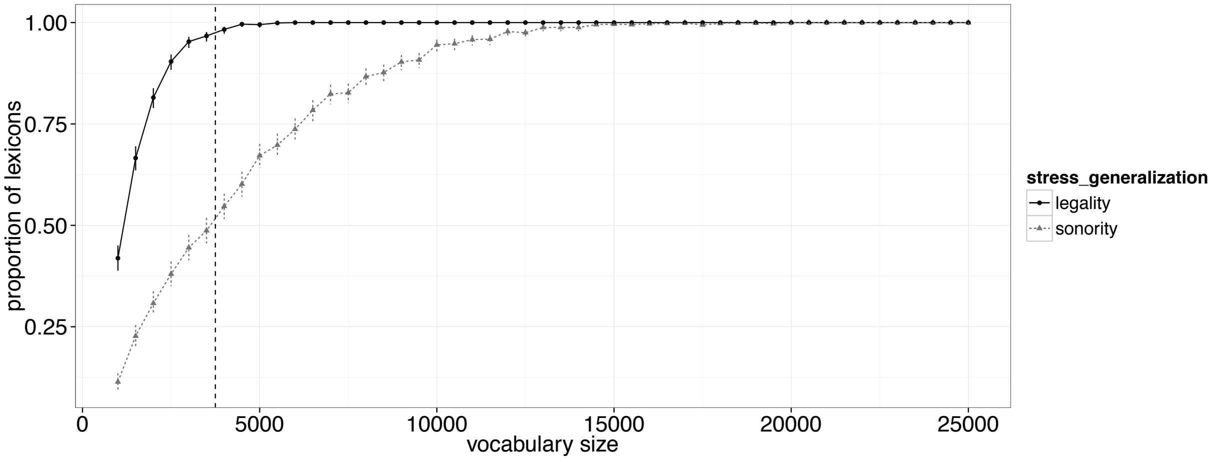
results, since sonority and metrical phenomena are often formally linked via syllable structure (e.g. Selkirk, 1982), and stress based on vowel sonority has received formal treatment (e.g. deLacy, 2004).

An alternative explanation is suggested in Pierrehumbert (2001), who argues that phonological constraints must be somewhat coarse-grained in order to be robustly transmittable across individual lexicons. Using a series of learning simulations where the training data consisted of randomly sampled vocabularies of various sizes, Pierrehumbert showed that formally simple phonological regularities were acquired relatively easily because they were supported by even the smallest lexicons. By contrast, second-order generalizations based on fine-grained phonotactics were statistically unstable, requiring greater overlap in the vocabularies of the learning agents. Specifically, the simulated learners internalized a first-order metrical pattern (initial stress on trisyllables) perfectly, even from a vocabulary of 400 words. They were also able to learn the relative well-formedness of medial nasal-obstruent clusters based on frequency. However, learning of a second-order regularity that paired the stress pattern with cluster identity was relatively poor.

This kind of mechanism appears to provide a plausible explanation for at least part of the present results. Note that the counts displayed in Figure 8 reveal that *illegal rise* clusters have relatively low type frequency. Type frequency has been argued to drive productivity; phonological patterns exemplified in few items do not spread easily, even if those items themselves are common (Bybee, 2001). If the metrical parse is to be inferred by learners from the behavior of weight-sensitive stress, then the sonority-based parse may be difficult to learn until one has acquired a considerably large lexicon. By comparison, a legality-based parse, being a superordinate generalization, should by definition have better lexical support.

To gain quantitative insight into the relative learnability of these two generalizations, we conducted a set of simulations similar to Pierrehumbert (2001). Vocabularies ranging in size from 1,000 to 25,000 items were sampled from the filtered CMU dictionary of about 53,000 total word forms. The sampling was weighted by SUBTLEXus counts to simulate the fact that frequent words tend to be learned early. Vocabulary size was incremented in 500-word steps, and 1,000 lexicons were sampled at each step. Following the sampling, each lexicon was restricted by the same criteria as the items that compose Figure 8, collapsing across the length dimension. Two separate logistic regression models were then fit to each restricted lexicon. The legality model tested whether words with *illegal* clusters (collapsed across sonority) featured higher rates of penult stress than words with *legal* clusters, and the sonority model tested the same relationship between *illegal fall* and *illegal rise* items only. Figure 9 presents the proportion of lexicons that acquired each generalization across vocabulary size.

Figure 9. Proportion of simulated lexicons of various sizes that acquired the legality-based and sonority-based stress generalizations (error bars = 95% CI). The dashed, vertical line represents where the individual trends observed in Experiment 1 fall on the two curves.



As seen in the figure, perfect learning of the legality-based generalization was achieved at 6,000 word forms, whereas sonority-based stress required a 20,000-word vocabulary to reach ceiling. In other words, the sonority-based generalization demanded over three times the data in order to completely spread through the community.

To compare the simulation outcome to the results observed in Experiment 1, we looked at the numerical trends in individual performance. Under the simplifying assumption that college undergraduates have vocabularies of roughly equal size *s*, the observed proportions of subjects who acquired each generalization can be predicted from Figure 9 by checking where each curve intersects a vertical line at x = *s*. Out of 30 subjects, 29 showed numerically higher penult stress on *legal* vs. *illegal* items (0.97 proportion). By contrast, only 16/30 subjects (0.53) showed sensitivity to sonority. These proportions correspond to about a 4,000 word vocabulary in Figure 9 (see the dashed, vertical line). The value of *s* should not be interpreted in absolute terms; estimating actual vocabulary size is notoriously difficult, and our filtered CMU dictionary is only a sample of the total word-forms in the English lexicon. What is important is the suggestion that the coarse-grained nature of stress assignment is related to the relative learnability of second-order phonotactic generalizations of different type frequencies. By the time one acquires a vocabulary large enough to reliably support the sonority-based generalization, years of practice with coarsely-conditioned stress may have biased one against the hypothesis that sonority may at some point become relevant.

5. CONCLUSION

Evidence for the view that phonotactic knowledge is gradient is by now overwhelming. What is needed next is an effort aimed at understanding how this knowledge interacts with the rest of phonology. The results of the present study show that the metrical parse applied during stress assignment does not make use of all of the information at its disposal. This alone argues for a flexible model of phonotactic knowledge, where different phonological processes can recruit phonotactic generalizations at different levels of specificity. Following prior work, we suggest that learnability differences driven by differences in type frequency constitute an important factor in the emergence of the level of generalization relevant to stress assignment. Other potential factors remain open to future investigation.

Appendix 1: List of Stimuli

All stimuli were used in Experiment 1; items in shaded rows were used in Experiment 2.

Frame	Singleton	Legal	Illegal Rise	Illegal Fall
daka__uth	*dakaduth*	*dakadwuth*	*dakadmuth*	*dakamduth*
deba__ab	*debapab*	*debaprab*	*debapmab*	*debampab*
depa__ish	*depasish*	*depasnish*	*depavrish*	*depansish*
faza__ish	*fazabish*	*fazablish*	*fazabnish*	*fazanbish*
fiba__ath	*fibagath*	*fibagrath*	*fibagnath*	*fibangath*
gidi__op	*gidizop*	*gidikwop*	*gidizrop*	*gidirzop*
kapa__iss	*kapathiss*	*kapathriss*	*kapathliss*	*kapalthiss*
kena__ozz	*kenadozz*	*kenadrozz*	*kenadlozz*	*kenalbozz*
kini__em	*kinitem*	*kinitrem*	*kinitlem*	*kiniltem*
lapa__up	*lapashup*	*lapashrup*	*lapashnup*	*lapanshup*
leka__op	*lekagop*	*lekagrop*	*lekagnop*	*lekangop*
lepa__azz	*lepabazz*	*lepablazz*	*lepabnazz*	*lepanbazz*
lidi__eph	*lidigeph*	*lidigleph*	*lidigmeph*	*lidimgeph*
mada__azz	*madapazz*	*madaplazz*	*madapnazz*	*madalpazz*
mene__uss	*menesuss*	*menesluss*	*menesruss*	*menelsuss*
nara__ish	*naragish*	*naraglish*	*naragmish*	*naramgish*
nepa__ep	*nepatep*	*nepatwep*	*nepatnep*	*nepantep*
nibi__im	*nibifim*	*nibifrim*	*nibifnim*	*nibimfim*
pima__ib	*pimavib*	*pimasmib*	*pimavlib*	*pimalvib*
pimi__oth	*pimitoth*	*pimitwoth*	*pimitnoth*	*pimintoth*
reda__osh	*redathosh*	*redathrosh*	*redathlosh*	*redalthosh*
saka__ud	*sakasud*	*sakasnud*	*sakavrud*	*sakansud*
sana__ep	*sanakep*	*sanakrep*	*sanaknep*	*sanankep*
sebi__aph	*sebishaph*	*sebishraph*	*sebishnap*	*sebinshaph*
shepi__oph	*shepidoph*	*shepidwoph*	*shepidmoph*	*shepimdoph*
shiga__eff	*shigapeff*	*shigapleff*	*shigapneff*	*shigalpeff*
shima__eph	*shimabeph*	*shimabreph*	*shimabweph*	*shimarbeph*
sipa__esh	*sipadesh*	*sipadresh*	*sipadlesh*	*sipalbesh*
taba__ub	*tabavub*	*tabasmub*	*tabavlub*	*tabalvub*
tama__ish	*tamapish*	*tamaprish*	*tamapmish*	*tamampish*
thana__iss	*thanabiss*	*thanabriss*	*thanabwiss*	*thanarbiss*
thibi__ar	*thibifar*	*thibiflar*	*thibizlar*	*thibilfar*
vata__iss	*vatafiss*	*vatafliss*	*vatazliss*	*vatalfiss*
vemi__oph	*vemikoph*	*vemikroph*	*vemiknoph*	*veminkoph*
waba__iss	*wabatiss*	*wabatriss*	*wabatliss*	*wabaltiss*
wibi__eph	*wibiseph*	*wibisleph*	*wibisreph*	*wibilseph*
zeda__up	*zedafup*	*zedafrup*	*zedafmup*	*zedamfup*
zepa__iss	*zepaziss*	*zepakwiss*	*zepazriss*	*zeparziss*

Appendix 2: Continuous predictor values by insert

Type	Insert	Son. slope	log(P): initial	log(P): wd-avg.	Type	Insert	Son. slope	log(P): initial	log(P): wd-avg.
legal	bl	3	-2.30	-2.71	illegal rise	bn	2	NA	-4.47
	br	4	-2.13	-2.68		bw	5	NA	-4.61
	dr	4	-2.32	-2.83		dl	3	NA	-3.12
	dw	5	-3.52	-3.53		dm	2	NA	-3.54
	fl	4	-2.20	-2.78		fm	3	NA	-5.03
	fr	5	-2.25	-2.82		gm	2	NA	-3.74
	gl	3	-2.51	-2.93		gn	2	NA	-3.29
	gr	4	-2.09	-2.55		kn	4	NA	-3.59
	kr	6	-2.03	-2.48		pm	4	NA	-4.17
	kw	7	-2.32	-2.70		pn	4	NA	-4.13
	pl	5	-2.22	-2.57		shn	3	NA	-3.74
	pr	6	-1.62	-2.27		sr	5	NA	-3.85
	shr	5	-3.00	-3.75		thl	4	NA	-4.02
	sl	4	-2.39	-2.80		tl	5	NA	-2.92
	sm	3	-2.77	-3.07		tn	4	NA	-3.49
	sn	3	-2.82	-3.19		vl	3	NA	-3.47
	thr	5	-2.74	-3.39		vr	3	NA	-3.74
	tr	6	-1.91	-2.19		zl	2	NA	-3.81
	tw	7	-2.89	-3.37		zr	3	NA	-4.43
illegal fall	lb	-3	NA	-3.56	singleton	b	6	-1.29	-2.74
	lf	-4	NA	-3.40		d	6	-1.29	-2.36
	lp	-5	NA	-3.58		f	7	-1.33	-2.71
	ls	-4	NA	-3.37		g	6	-1.59	-3.27
	lt	-5	NA	-2.86		k	8	-1.03	-3.20
	lth	-4	NA	-4.23		p	8	-1.07	-2.87
	lv	-2	NA	-3.39		s	7	-0.99	-2.26
	md	-2	NA	-3.34		sh	7	-2.01	-2.12
	mf	-3	NA	-3.66		t	8	-1.35	-2.59
	mg	-2	NA	-5.51		th	7	-2.17	-3.62
	mp	-4	NA	-2.49		v	5	-1.65	-2.80
	nb	-2	NA	-3.63		z	5	-2.59	-3.79
	ng	-2	NA	-3.68					
	nk	-4	NA	-2.98					
	ns	-3	NA	-2.21					
	nsh	-3	NA	-3.19					
	nt	-4	NA	-1.99					
	rb	-4	NA	-2.94					
	rz	-3	NA	-2.23					

References

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

Arndt-Lappe, S. (2011). Towards an exemplar-based model of stress in English noun–noun compounds. *JL*, *47*(03), 549-585.

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

Bailey, Todd M. & Ulrike Hahn. 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.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L.m Simpson, G.B. & Treiman, R. (2007). The English Lexicon Project, *39*(3), 445–459.

Baptista, B. O. (1984). English stress rules and native speakers. *Language and Speech, 27*, 217–233.

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.

Bartlett, S., Kondrak, G., & Cherry, C. (2009). On the Syllabification of Phonemes (pp. 308–316). *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL.*

Bates D, Maechler M, Bolker B and Walker S (2014). _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.7, http://CRAN.R-project.org/package=lme4.

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.

Bell, A., & Hooper, J. B. (1978). Issues and evidence in syllabic phonology. In A. Bell, & J. B. Hooper (Eds.), *Syllables and segments* (pp. 3–22). Amsterdam: North-Holland Publishing.

Berent, I., Lennertz, T., & Balaban, E. (2012). Language universals and misidentification: a two way street. *Language and Speech, 55*(3), 311-330.

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., 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.

Bertinetto, P. M. (2004). On the undecidable syllabification of /sC/ clusters in Italian: Converging experimental evidence. *Italian Journal of Linguistics, 16*, 349-372.

Bertinetto, P.M., Caboara, M., Gaeta, L. & Agonigi, M. (1994). Syllabic division and intersegmental cohesion in Italian. In Wolfgang U. Dressler, Martin Prinzhorn & John R. Renisson (eds.), *Phonologica. Proceedings of the 7th International Phonology Meeting*, Rosenberg & Sellier, Torino. 19-33.

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*, Saarbrücken. 1953-1956.

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

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.

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.

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.

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., McQueen, J. M., Jansonius, M., & Bayerl, S. (2002). The lexical statistics of competitor activation in spoken-word recognition (pp. 40–45). *Proceedings of the 9th Australian International Conference on Speech Science Technology*, Melbourne.

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

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

Davidson, L., Jusczyk, P., & Smolensky, P. (2004). The initial and final states: Theoretical implications and experimental explorations of richness of the base. In R. Kager, W. Zonneveld, & J. Pater (Eds.), *Fixing priorities: Constraints in phonological acquisition.* Cambridge: Cambridge University Press.

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.

Dziubalska-Kołaczyk, K. (2002). *Beats-and-Binding Phonology*. Frankfurt/Main: Peter Lang.

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

Eddington, D., Treiman, R., & Elzinga, D. (2013). 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. *JL*, *17*(2), 309–317.

Frisch, S., 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, S. A, Pierrehumbert, J.B., and Broe, M. (2004). Similarity Avoidance and the OCP, *NLLT, 22*, 179-228.

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

Gordon, M., Jany, C., Nash, C., & Takara, N. (2008). Vowel and Consonant Sonority and Coda Weight: A Cross-Linguistic Study. In C. B. Chang & H. J. Haynie, (pp. 208–216). *WCCFL 26*, Sommerville, MA.

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

Gouskova, M. (2004). Relational Hierarchies in Optimality Theory: The Case of Syllable Contact. *Phonology*, *21*(2), 201–250.

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.

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

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

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

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

Hayes, B. (1980). *A metrical theory of stress rules*. PhD. Dissertation, MIT.

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

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

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI,* 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*.

Hooper, J. B. 1975. The archi-segment in natural generative phonology. *Lg,* 51, 536-560.

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.

Kahn, D. (1976). *Syllable Based Generalizations in English Phonology*. PhD. dissertation, MIT, Cambridge, MA.

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. http://CRAN.R-project.org/package=vwr

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator, *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.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *JASA*, *59*(5), 1208–1221.

Kučera, H., & Francis, W.N. (1967). Computational analysis of present- day American English. Providence, RI: Brown University Press.

Kuznetsova, A., Brockhoff, P.B., and Bojesen Christensen, R.H. (2014). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-20. http://CRAN.R-project.org/package=lmerTest.

de Lacy, Paul (2004). Markedness conflation in Optimality Theory. *Phonology, 21*(2),145-199.

Lee, Y. (2006). *Sub-syllabic constituency in Korean and English*. PhD. Dissertation, Northwestern University.

Lee, Y., & Goldrick, M. (2008). The emergence of sub-syllabic representations. *Journal of Memory and Language*, *59*, 155-168.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *LI, 8*, 249–336.

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.

McGowan, K. B. (2009). Gradient Lexical Reflexes of the Syllable Contact Law (Vol. 45, p. 445). *Presented at the Cognitive Society Meeting*.

McQueen, J. M. (2004). Speech perception. In K. Lamberts & R. Goldstone (Eds.), *The handbook of cognition* (pp. 255-275). London: Sage.

Moore-Cantwell, C. (2015). The phonological grammar is probabilistic: New evidence pitting abstract representations against analogy. *Paper presented at the Annual Meeting on Phonology*, Vancouver.

Morais, J., & Kolinsky, R. (1997). Levels of Processing in the Phonological Segmentation of Speech. *Language and Cognitive Processes*, *12*(5-6), 871–876.

Morrill, T. (2012). Acoustic correlates of stress in English adjective–noun compounds. *Language and Speech*, *55*(2), 167–201.

Ohala, D. K. (1999). The influence of sonority on children's cluster reductions, *Journal of Communication Disorders*, *32*(6), 397–422.

Ohala, J. J. & Ohala, M. (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In John J. Ohala & Jeri Jaeger (eds.) *Experimental phonology*, 239-252. Orlando: Academic Press.

Parker, Stephen G. (2002). Quantifying the sonority hierarchy. PhD dissertation, University of Massachusetts, Amherst.

Perruchet, P., and Peereman, R. (2004). The Exploitation of Distributional Information in Syllable Processing. *Journal of Neurolinguistics*, *17*, 97-119.

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.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of*

Memory and Language, *39*, 347–370.

Prince, A. (1983). Relating to the Grid, *LI, 14*, 19-100.

Prince, A. (1990) Quantitative consequences of rhythmic organization. *CLS*, 26.2, 355–98.

Redford, M. A. (2008). Production constraints on learning novel onset phonotactics. *Cognition*, *107*, 785–816.

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, Kevin M. (2011). *Gradient weight in phonology.* PhD. dissertation, UCLA.

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. (1982). The syllable. In H. van der Hulst & N. Smith (eds.), *The structure of phonological representations* (pp. 337–383). Dordrecht: Foris.

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.

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.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *LI*, *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. (I999). Alternatives in syllable-based accounts of consonantal phonotactics. In Fujimura, O., Joseph, B., & Palek, B. (eds.), *Proceedings of LP I998*, *R*, 205-246. Prague: Charles University and Karolinum Press.

Steriade, D. (2008). Resyllabification in the quantitative meters of Ancient Greek: Evidence for an Interval Theory of Weight. *ms*, MIT.

Titone, D., & Connine, C. M. (1997). Syllabification strategies in spoken word processing: Evidence from phonological priming. *Psychological Research*, *60*, 251–263.

Treiman, R., Bowey, J. A., & Bourassa, D. (2002). Segmentation of spoken words into syllables by English-speaking children as compared to adults. *Journal of Experimental Child Psychology*, *83*(3), 213–238.

Treiman, R. & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language*, *27*, 87-104.

Treiman, R., Fowler, C. A., Gross, J., Berch, D., & Weatherston, S. (1995). Syllable structure or word structure? Evidence for onset and rime units with disyllabic and trisyllabic stimuli. *Journal of Memory and Language, 34*(1), 132-155.

Treiman, R., & Zukowski, A. (1990). Toward an understanding of English syllabification. *Journal of Memory and Language*, *29*(1), 66–85.

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.

Vaden, K.I., Halpin, H.R., Hickok, G.S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. Available from http://www.iphod.com.

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments Computers*, *36*(3), 481–487.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech*, *40*(1), 47–62.

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.

Weide, Robert L. 1994. CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

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.

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.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, *12*(1), 85–129.