**Learning Principles of Syllabification from Word-Edge Phonotactics**

Blake Allen
University of British Columbia

## 1 Introduction

Speakers of a natural language possess the ability to judge the relative felicity of different divisions (or "syllabifications") of words in that language, as demonstrated by a variety of experimental tasks (Treiman and Danis 1988; Derwing 1992; Redford and Randall 2005). Native English speakers, for example, prefer to divide the word *alarm* as [ə / lɑɹm] rather than [əl / ɑɹm], whereas judgments about *lemon* are more evenly split among divisions like [lɛ / mən] and [lɛm / ən] (Eddington et al. 2013). These intuitions indicate that the subjects in these experiments use phonological knowledge to assess the acceptability of word divisions, and that some of this knowledge mirrors phonotactic trends observed in the lexicon (Steriade 1999; Hammond 1999; Hall 2006). This paper presents a simulation of phonotactic learning in English whose purpose is to establish which predictors of word division judgments—formalized as weighted constraints—can be said to reflect learned knowledge of phonotactic patterns in the lexicon.

In this simulation, I demonstrate that a system of accurate constraint weights can be derived from sets of responses in experimental tasks as well as from plain corpora of attested forms, and that these two sets of weights can be made to generate predictions about each other's respective data sets. More specifically, I show that constraint weights learned from a lexicon of English forms have a high correlation with the weights learned from responses on an experimental word division task. Finally, I provide evidence that whereas the role in determining word division judgments played by phonotactic restrictions on vowel and stop glottalization can be learned from patterns in the English lexicon, the role of patterns of anticipatory nasalization, stress, and ONSET/NOCODA must be due to other, perhaps phonetic, factors.

The remainder of this section provides context for the discussion of word division judgments and describes factors that previous literature has claimed affect those judgments. Section 2 explains the theoretical underpinnings and technical details of the simulation used to test the origin of knowledge of these factors. Section 3 follows by examining the results of that simulation, and section 4 concludes.

### 1.1 The Nature of Word Division Judgments

The core empirical data I consider in this paper are experimentally-elicited judgments about word division. Although such experiments can take many forms, the most typical methodology involves presenting a known word with two alternative pronunciations and asking subjects to choose which pronunciation they prefer; the pronunciation choices have been constructed by adding a brief pause at one possible syllable boundary location (Derwing 1992). See Côté and Kharlamov (2010) for a summary of the apparent effects of different varieties of word division tasks. Although these tasks tend to produce high levels of variability among responses, numerous studies have demonstrated that such judgments reflect principles which can be stated in phonological terms, such as a preference for word divisions which do not produce phonotactically impermissible sequences (Redford and Randall 2005; Eddington et al. 2013).

Due to the high degree of correspondence between such divisions and the boundary locations predicted among the theoretical units called syllables, intra-morpheme word division is typically described as *syllabification* (Itô 1986; Hayes 1989; Becker et al. 2012). Moreover, there is a strong tendency for these word division tasks to create syllable edges that mirror the phonotactics observed at word edges in that particular language. Two types of explanations for this correspondence have been proposed: 1) that word-division judgments make direct use of phonologically-encoded syllable boundaries, whose positions are learned from word-edge phonotactics via Prosodic Licensing (Itô 1986); or 2) that word division produces units that are themselves best considered words, or *sub-words*, and hence subject to the same phonotactic restrictions as "real" words (Steriade 1999; Samuels 2009). Prosodic Licensing, in this context, would dictate that because all words must be exhaustively parsed into syllables, any observed word edge will also be a syllable edge.

In this paper, I remain agnostic about which of these interpretations of word division responses is more theoretically accurate. Aside from the possible influence of language-nonspecific constraints on syllable structure, both options yield the same predictions: that the phonotactic knowledge which produces word division judgments will directly reflect the patterns observed at word edges within a particular language. Whether this knowledge is comprised of a set of constraints on allowable syllable structures or on allowable word-edges is immaterial. I remain impartial and avoid referring to word divisions as syllabifications. I describe relevant constraints as simply "edge constraints" rather than syllable- or word-edge constraints, although any of these terms may in fact be correct.

Note also that this paper investigates only patterns of word division among words with a single intervocalic consonant, e.g. *lemon* and *lemur* but not *army* or *osprey*. Also, due to the corpus of word division judgments I use, which is data from a binary forced-choice task, there will be no discussion of potential "ambisyllabic" divisions, e.g. [lɛm.mən].

### 1.2 Posited Predictors of Word Division

Previous research has shown the usefulness of numerous phonological factors in predicting patterns of word division. Because the simulation described in this paper includes a subset of these in its models of word division, in this section I give a brief description of each member of this subset.

Steriade (1999) proposed that learning from phonotactic patterns at word edges is the basis for observed patterns of word division, and that the results of this learning can be codified as phonological constraints. Steriade's analysis predicts that word divisions which produce edges that are phonotactically impermissible will be dispreferred, although not rejected outright because of conflicting constraints like ONSET. Dividing the word *lemon* as [lɛ / mən], for example, produces a right edge with an [ɛ], which is never attested in English, and so is less than optimal as a word division. The alternative [lɛm / ən], however, violates both ONSET and NOCODA. Steriade (1999) predicts, then, that neither division will be preferred categorically over the other. This prediction is borne out in the data cited from Derwing (1992) and others.

Steriade's (1999) description of word division patterns in English specifically also makes use of the observation, cited from Pierrehumbert and Talkin (1992), that word-initial vowels are glottalized. Word divisions which result in unglottalized vowels at the left edge of a "syllable," then, will be dispreferred. The paper presents this contrast not as a central aspect of its argument, but rather as a single example of a type of "allophonic" distinction that can play a role in word division judgments. Although Steriade does not specify how, in her theory, this vowel glottalization is phonetically implemented, I will follow Dilley et al. (1996) and Redi and Shattuck-Hufnagel (2001) in assuming that glottalized vowels are produced with a constricted glottis during voicing, rather than the insertion of a pre-vocalic glottal stop.

Additionally, Steriade (1999) contends that the constraints ONSET and NOCODA, which do not necessarily need to be learned from word-edge phonotactics, play an equally central role in word division judgments. The strong tendency toward V/CV divisions in Derwing's (1992) and others' experimental data supports this logic. For discussion of why this imbalance between onsets and codas exists, see Tesar and Smolensky (1998) and Gordon (2005).

Eddington et al. (2013) describe a large-scale study of word division judgments in English. Their results provide strong evidence that stress patterns can affect word division judgments: a medial consonant tends to associate with whichever of its neighboring vowels is stressed. This finding was also presented by Derwing (1992), among others. In addition to adducing evidence that stress patterns affect word division judgments, Eddington et al. (2013) demonstrate using mixed effect models that many different factors, including stress position, can interact to produce word division judgments—an unsurprising finding, but one that bolsters Steriade's (1999) claim that word division patterns emerge from constraint interaction. I build on this idea in my own simulation of phonological learning. Note also that I use the data from Eddington et al. (2013) as the basis of my simulation.

Description of another potentially useful allophonic feature distinction comes from Hall (2006): word-final stops in English are non-contrastively glottalized, as in *cap* [kʰæˀp] and *log* [lɑˀg]. Again, I assume that this glottalization comes in the form of a constricted glottis simultaneous with the stop articulation rather than an inserted glottal stop before or after it.

For the purposes of this paper, I will also make use of the allophonic distinction of anticipatory vowel nasalization, as described by Cohn (1993). The generalization she presents is that vowels immediately preceding nasal consonants in English are anticipatorily nasalized, e.g. *pen* [pʰɛ̃n]. Although Cohn (1993) does not clarify what restrictions might exist on this environment for nasalization, it has been claimed that only coda nasals cause nasalization of their preceding vowel (Ladefoged and Johnson 2010). However, there is no deterministic way to establish the syllabic constituency of English words; this contrast could be considered the cause of Eddington et al.'s (2013) finding that words with a nasal medial (e.g. *lemon*) tend to be divided as VC/V more often than words with an oral stop medial, as that division would prevent a nasalized vowel from occurring at a right-edge (ɛ̃#), as is never found in real English words. To abstract away from this dilemma, I consider all vowels immediately preceding a nasal consonant to be nasalized, and carry out my simulation accordingly.

The list in (1) enumerates the factors introduced in this subsection whose role in word division judgments will be assessed by my simulation.

(1)

1. phonotactic permissibility of phonemes at word edges
2. stress position
3. word-initial vowel glottalization
4. word-final stop glottalization
5. anticipatory nasalization
6. ONSET and NOCODA

## 2 Learning Simulation

Having established a (certainly incomplete) set of phonological factors that can play a role in determining word division judgments, I now move to the central topic of this paper: simulating phonotactic learning in order to determine which of these factors can be considered to have been learned from sound patterns in the lexicon of English. To do so, I make the following assumptions:

(2)

a. The phonological grammar is comprised of weighted constraints (Legendre et al. 1990).

b. The constraint weights which constitute the phonological grammar are learned from exposure to primary language data (Hayes and Wilson 2008; Hayes 2011a).

c. All subjects in word division judgment tasks have the same phonological grammar, and so variability in responses indicates variable outputs of the grammar, not only inter-speaker variation.

d. If the weight of a constraint that determines word division judgments is learned from phonotactic patterns in the lexicon, then it is learned specifically from the patterns at word edges (see further description below).

The diagram in (3) schematizes the process of learning and then "using" a phonology from the standpoint of a language-user, as based on these assumptions.

(3) *observed forms* ⇒ *constraint weights* ⇒ *phonological judgments*

The purpose of the assumption in (2d) is only to account in a generalized way for either a) the learning effects of Prosodic Licensing if word divisions are taken to represent syllabifications (i.e. licit syllable edges are learned from word edges) or b) for the claim that if word division can be considered to produce multiple words, then standard word-edge phonotactics will also apply to these. While it is possible that phonotactic constraints not directly related to material at word edges—for example, word minimality effects—could be serving some role in determining word division judgments, I do not take these yet-undescribed factors into account. This restriction is also partly due to the impracticality of running phonotactic simulations of domains larger than word edges. As described later in this paper, even constraints in this restricted category are able to accurately predict word division judgments.

With these assumptions in place, it becomes possible to create a quantitative description (or grammar) in terms of constraint weights both of the English lexicon in general and of a set of attested word division judgments. Under the expectation that word division judgments are driven by the phonological grammar, and that this grammar is itself constructed on the basis of patterns in the lexicon, then in principle these two sets of constraint weights—one based on the lexicon and the other based on word division judgments—should be identical or at least equivalent. This methodology can be described as shown in (4).

(4)
$$\begin{array}{lllll} \textit{observed forms} & \Rightarrow & \textit{constraint weights}_{\textit{lexicon}} & & \\ & & \approx & & \\ & & \textit{constraint weights}_{\textit{judgments}} & \Leftarrow & \textit{phonological judgments} \end{array}$$

It will come as no surprise, however, that the constraint weights that optimally describe these two sets of data (as calculated using a Maximum Entropy classification algorithm, described below) are not identical, although they are unquestionably similar in many respects. For example, the constraint penalizing [ɪ] at a right edge, *ɪ#, has a high weight in both. Constraint weights will also vary depending on the types of constraints used. For the purposes of this simulation, I encode constraints referring to the presence at a word edge of any phoneme (e.g. *ɛ#), as well as constraints referring to word-edge vowel glottalization (e.g. *ˀV#), stop glottalization (*#ˀC), vowel nasalization (*Ṽ#), stress (*V́#, *V́C#), and also ONSET and NOCODA. Each of these types of constraints is present in both a word-initial and word-final version, and for both polarities (+ and -) of the relevant feature. This includes the constraints ANTI-ONSET and ANTI-NOCODA. A full list of the constraints used in this simulation are provided in section 2.2.2.

Constraint weights are calculated using the MaxEnt Grammar Tool (Wilson and George 2009), an implementation of the maximum entropy classification algorithm. This algorithm has been mathematically demonstrated to converge on the optimal weights to describe any provided set of data (see Hayes and Wilson (2008) for details), and so can be relied upon to produce the correct grammar, or set of weights, given that the attested forms, the available constraints, and each form's violations of those constraints are known. The following subsections describe the specifics of the theoretical underpinnings of phonological grammars as sets of weighted constraints (2.1) and describe the particulars of the implementation of that theory which I use here (2.2).

## 2.1 Specifics of the Theory

In order to make use of maximum entropy classification in determining the optimal constraint weights to describe one of the two target corpora in this study, it is necessary to assume a model of grammar in which knowledge is parametrized as a set of constraints which are weighted relative to each other, and in which grammar outputs can be probabilistic rather than deterministic. This assumption corresponds to the approach taken in Maximum Entropy Grammar (Goldwater and Johnson 2003; Hayes and Wilson 2008; Hayes 2011a). As in Harmonic Grammar (Legendre et al. 1990; Coetzee and Pater 2008) and Optimality Theory (Prince and Smolensky 2004), output forms can either incur constraint violations for particular sequences or forms in an output (Markedness) or for mismatches between elements in the input and elements in the output (Faithfulness). Because all the constraints relevant to word division judgments which I discuss here are Markedness constraints, I do not include Faithfulness constraints in my simulation or this paper.

In weighted constraint models of grammar, the wellformedness of a particular output—that is, its probability of occurring in the language—is the weighted sum of its constraint violations. As an example, consider a language in which nearly all words begin with a consonant and end with a vowel. The grammar that describes (this aspect of) this language may have weights similar to those in (5).

(5)

| *constraint* | *weight* |
|---|---|
| *C# | 5 |
| *#V | 5 |
| *V# | 0 |

Higher weights indicate that a particular constraint is an accurate descriptor of patterns in the language. Because this language has virtually no words ending with a consonant, the constraint *C#, which incurs violations for forms with word-final consonants and can be thought of as a statement that "this language disprefers word-final consonants," has a high weight. Conversely, because all words in this language end in a vowel, the constraint *V# ("this language disprefers word-final vowels") is highly inaccurate, and thus receives a low weight.

The tableau in (6) demonstrates the wellformedness assessment of three novel words based on the grammar (constraint weights) of the language described above. The top row shows the three relevant constraints and their respective weights. In each cell is a statement of how many violations the form in that row incurs of the constraint in that column, multiplied by the weight of that constraint. Unlike in Optimality Theory, the products of violation count times constraint weight are summed for each form, resulting in a harmony score. Higher harmony scores indicate more violations of higher weighted constraints, and thus a worse fit to the sound patterns of the language described by the provided weights.

(6)

| | *C#: 5 | *#V: 5 | *V#: 0 | *harmony score* | *probability* |
|---|---|---|---|---|---|
| la | 0*5 | + 0*5 | + 1*0 | **= 0** | $\approx$ **0.99** |
| bap | 1*5 | + 0*5 | + 0*0 | **= 5** | $\approx$ **0.007** |
| at | 1*5 | + 1*5 | + 0*0 | **= 10** | $<$ **0.0001** |

One crucial aspect of harmony scores is that they can be converted into probabilities. To do so, harmony scores are negated and *e* raised to that power, producing a number in the range [0, 1]. This process is performed for every possible form in the language—usually in practice by the sum of these numbers across all *listed* potential outputs—and the [0, 1] value for a given form is divided by that sum. (See Hayes and Wilson (2008) for further discussion.) In the example tableau, the probability 0.99 of the form [la] indicates that this form is a highly acceptable word in this language, much more so than [bap] and [at]. In reality no single linguistic form could have a probability close to 0.99, since the probabilities of all forms within an entire language must sum to 1, and so the values given here are purely illustrative.

This method of conversion between harmony scores and probabilities renders it possible to calculate constraint weights *from* the observed probabilities of a set of forms. To again use the example in (6): given a list of 10,000 words in which 70 are of shape CVC, 1 is of shape VC, and all the rest are of shape CV, the relative probabilities of each of these word shapes can be used to calculate the weights of these three constraints which most accurately describe this list of words. Those weights will be nearly identical to those listed in (5). This is the process that underlies maximum entropy classification, which is used in the simulation described here.

### 2.2 Specifics of the Implementation

Using the approach outlined above, I calculate the optimal constraint weights for describing a) word-edge phonotactics in the English lexicon and b) experimentally-elicited word division judgments, in order to determine which constraint weights used to predict syllabification judgments can be said to have been learned from the general word-edge phonotactics. This subsection lays out the details of my methodology: the corpora I use to represent both data sets, their coding, and the implementation I use of the relevant constraints.

With these corpora and constraints, I use the MaxEnt Grammar Tool created by Wilson and George (2009) to determine the optimal constraint weights for each corpus based on the principle of maximum entropy—essentially "spreading out" the responsibility for attested patterns among as many constraints as possible while still maximizing predictive power, as shown to be phonologically justifiable by Hayes (2011a) and Martin (2011)—and the theory of weighted constraints in (2.1).

#### 2.2.1 Coding of Corpora

To represent the lexicon of English that a learner of the language might encounter, I use the Carnegie Mellon University Pronouncing Dictionary (Weide 2005), specifically the corrected subset of entries used in the software BLICK (Hayes 2011a). This lexicon includes 18,033 words from the CMU Pronouncing Dictionary, chosen for having a CELEX (Burnage et al. 1990) frequency greater than zero, not including a productive affix, not being an acronym, etc.

It is crucial for learning accurate English phonotactics that the MaxEnt Grammar Tool is provided not only with the forms in the CMU Pronouncing Dictionary, but also with a list of unobserved—zero-probability—forms. Running the MaxEnt Grammar Tool with only the CMU Pronouncing Dictionary's forms provided yields weights that plainly cannot represent the grammar of English: the constraint *#ŋ, for example, ends up with a *lower* weight than constraints that should have low weights, including *#tʃ and *#θ (2.112, compared to 2.997 and 3.292, respectively). Only when zero-probability [ŋ]-initial forms are provided to the Tool will *#ŋ emerge with an appropriately high weight.

The other difficulty in calculating constraint weights from the CMU Pronouncing Dictionary is its sheer size. 18,033 words is a large number of forms for the MaxEnt Grammar Tool to process, and this would require immense computational capacity and/or time. Adding in all of the (potentially infinite) zero-frequency forms of English would compound the problem further—if it is even possible to represent all possible but unattested pseudo-English words.

For this paper, I use a coding technique which solves both the problem of including non-observed forms and the problem of lexicon size. Because the domain under examination is division of words with only a single intervocalic consonant with no option for ambisyllabicity, and in which only a single vowel flanks this consonant on either side, the only possible responses are *...V.CV...* or *...VC.V...*. Accordingly, the only lexical phonotactics which will be relevant to these judgments are those pertaining to the first segment or last segment of a word, in addition to the stress level of the first or last vowel. The compression I performed, exemplified in the difference between (7) and (8), significantly reduces the number of individual cases for the MaxEnt Grammar Tool to evaluate, and therefore speeds up its run time significantly.

| | | |
|---|---|---|
| | kǽt | 1 |
| | kɛ́ŋ | 1 |
| (7) | moʊ́t | 1 |
| | ... | |
| | (length = 18,033) | |

| | | |
|---|---|---|
| | k V́ X X X | 2 |
| | m V́ X X X | 1 |
| (8) | X X X V́ t | 2 |
| | X X X V́ N | 1 |
| | ... | |
| | (length = 141) | |

The other advantage to this method of coding the lexicon is that it allows inclusion of the unobserved forms in a straightforward and numerically conservative way. In addition to the sequences as in (8) which *are* observed in the CMU Pronouncing Dictionary, I added all sequences which are *not* observed there, giving them a frequency of zero, in order to produce the master lexicon to be given to the MaxEnt Grammar Tool, resulting in a manageable 432 entries. This process is parallel to, if more manual than, the inner workings of the UCLA Phonotactic Learner (Hayes and Wilson 2008).

| | | |
|---|---|---|
| | k V́ X X X | 2 |
| | m V́ X X X | 1 |
| | X X X V́ t | 2 |
| | X X X V́ N | 1 |
| (9) | ŋ V́ X X X | 0 |
| | X X X V́ h | 0 |
| | ... | |
| | (length = 432) | |

This master list of forms was provided to the MaxEnt Grammar Tool as a single “tableau" of forms. Although the different lexical items of English are not in any sense competing outputs of a single input form, providing them this way to the Tool allows it to assess the phonotactic principles underlying the patterns evident in the lexicon. As the following section will explain, this methodology is also empirically validated by the highly accurate weights it produces.

Aside from using the CMU Pronouncing Dictionary to represent the lexicon of English, I adopted the results from Eddington et al.’s (2013) large-scale study of English syllabification as my corpus of attested word division judgments. Parallel to the CMU corpus, I refer to this data set as the BYU corpus, named after the location at which it was elicited. This study collected word division judgments from 842 native speakers of English for a total of 125 words (from a total of five thousand test words) per subject, resulting a grand total of over ten thousand responses. Their methodology was a variant of the “pause-break task” (Derwing 1992), in which subjects seated at a computer screen were visually presented with two possible divisions of each word written in ARPABET and asked to choose one or the other. For example, on the item *lemon*, subjects were asked to indicate a preference for either *L EH / M AH N* or *L EH M / AH N*. Responses were aggregated across all subjects.

For this study, I used only a subset of the words tested by Eddington et al. (2013). Specifically, I removed:

(10)

- words with orthographically double consonant or cluster, e.g. ‘kipper’, ‘badges’

- words in which either side of any syllabification is an independent morpheme, e.g. ‘lackey’
- recent or obscure loanwords like ‘burro’, ‘couchant’
- words with multiple common pronunciations, e.g. ‘dilute’
- words with more than one intervocalic consonant, e.g. ‘escape’

The BYU corpus reflects data from a binary forced-choice task. Accordingly, I provided them to the MaxEnt Grammar Tool as a set of two-output tableaux. Each output form was reduced to only its crucial ...VCV... sub-part, and the the order of two halves of each output word division were reversed so as to allow each one’s constraint violations to be assessed using regular expressions. See (11) for an example; the numbers 1 and 0 indicate stress levels, and “xx xx" a contentless buffer string between the two sides of each division.

| | | | |
|---|---|---|---|
| (11) | yogi | G IY0 xx xx OW1 | 21 |
| | | IY0 xx xx OW1 G | 1 |
| | wicket | K AH0 xx xx IH1 | 10 |
| | | AH0 xx xx IH1 K | 10 |

Both the CMU and BYU corpora were provided coded in ARPABET using symbols that correspond to what are sometimes considered the “phonemes” of English, i.e. 24 consonants and 15 vowels, including three diphthongs. Because I am testing for the effects of constraints on allophonic distinctions at word edges, such as glottalization and nasalization, I also coded these lexicons to contain these allophones in the positions suggested by the references listed in section 1.2. For example, all word-initial vowels were marked for glottalization, and all vowels before nasal consonants were marked for anticipatory nasalization.

### 2.2.2 Coding of Constraints

In the absence of any particular theoretical motivation for limiting the number or type of constraints to provide to the weighting algorithm, one approach might be to provide constraints based on all possible featural combinations. The MaxEnt Grammar Tool requires only that all constraints be stated negatively (penalizing forms). Given the input data, the relevant constraints would achieve a high weight and the irrelevant ones would not. However, as with coding the lexicons, there are practical reasons to be prudent in choosing which constraints to allow. Above all is the need to keep the number of constraints to a number that can be assigned weights in a finite and reasonably short period of time.

To this end, rather than use constraints penalizing all possible featural combinations at either edge of a phonological domain, I include two different types of constraints: those on the presence of particular “phonemic” feature sets, and those on the presence of “allophonic” features, both at either the left or right edge[1]. Using both types of constraints is crucial because it allows multiple constraints’ violations to contribute to a form’s (lack of) wellformedness—so-called “gang effects."

The chart in (12) illustrates the phonemic constraints included in the simulation, which are listed as segments but correspond to the appropriate feature values. Allophonic constraints are listed in (13), where only the [+F] versions are listed for economy of space; note that [-F] versions of all constraints are also used in the simulation. Example strings are not all real or even phonotactically permissible words in English.

[1] I do not intend to endorse a model of phonology in which there is a strict contrast between phonemes and allophones. By “phonemic” feature sets, I simply refer to the features of a segment which remain constant across all of its variants I have coded into the CMU and BYU corpora, such as glottalization and nasalization. See Hall (2009) for a discussion of predictability in the relationships among segments.

(12)

| *constraint name* | *description* | *example strings* |
|---|---|---|
| *#$a_{\text{phonemic}}$ | pick out all allophones of /a/ at a left edge | [a], [ˀaso] |
| *$a_{\text{phonemic}}$# | pick out all allophones of /a/ at a right edge | [a], [tã] |
| *#$ɔ_{\text{phonemic}}$ | picks out all allophones of /ɔ/ at a left edge | [ɔ], [ˀɔs] |
| *$ɔ_{\text{phonemic}}$# | picks out all allophones of /ɔ/ at a right edge | [ɔ], [rɔ̃] |
| *#$t_{\text{phonemic}}$ | picks out all allophones of /t/ at a left edge | [t], [ti] |
| *$t_{\text{phonemic}}$# | picks out all allophones of /t/ at a right edge | [t], [eˀt] |
| (other "phonemic" constraints elided) | | |

(13)

| *constraint name* | *description* | *example strings* |
|---|---|---|
| *#[+stress] | picks out all left edges with a stressed vowel | [ˀál] |
| *[+stress]# | picks out all left edges with a stressed vowel | [ní] |
| *#C[+stress] | picks out all left edges with a consonant and whose first vowel is stressed | [tá] |
| *[+stress]C# | picks out all right edges with a consonant and whose last vowel is stressed | [ˀáp] |
| *#[+stopglottalized] | picks out all glottalized stops at a left edge | [ˀpoɪ] |
| *[+stopglottalized]# | picks out all glottalized stops at a right edge | [teɪˀk] |
| *#[+vglottalized] | picks out all glottalized vowels at a left edge | [ˀan] |
| *[+vglottalized]# | picks out all glottalized vowels at a right edge | [sˀeɪ] |
| *#[+anticnasal] | picks out all nasalized vowels at a left edge | [ũm] |
| *[+anticnasal]# | picks out all nasalized vowels at a left edge | [sɛ̃] |
| (and all [-F] versions of these constraints) | | |

Following Steriade (1999), I also include Onset and NoCoda constraints. As their "featural" complements, the constraints Anti-Onset and Anti-NoCoda have been implemented as well.

(14)

| *constraint name* | *description* | *example strings* |
|---|---|---|
| Onset | picks out initial vowels | [ˀa], [ẽɪ] |
| NoCoda | picks out final consonants | [iˀp], [oʊl] |
| Anti-Onset | picks out initial consonants | [lˀa], [kẽɪ] |
| Anti-NoCoda | picks out final vowels | [i], [toʊ] |

Using this set of constraints allows the simulation to accurately assess all of the phonological material which can be relevant to word division judgments (according to the assumptions of this paper) while also keeping the number of constraints down to a mere 102: 78 "phonemic" constraints (one for each phoneme, for each edge), 20 "allophonic" constraints (one for each allophonic feature, for each feature value, for each edge), and ONSET, NOCODA, ANTI-ONSET, and ANTI-NOCODA.

The MaxEnt Grammar Tool requires that users provide input files consisting of tableaux with the violations that each form incurs for each constraint already marked. In order to generate these constraint violation matrices, I implemented each constraint as a regular expression, a standard method of string-matching, and used these to automatically produce these matrices.

The set of phonological features used by this simulation was taken from Hayes (2011b). Additions to this feature set were made for each of the above allophones and allophonic features such as glottalization.

## 3 Results

In this section I give an overview of the results of the simulation described above. 3.1 describes the optimal constraint weights calculated for the experimental word division judgment data set ("the BYU data"), 3.2 provides the weights calculated for the English lexicon ("the CMU data") and describes the changes that need to be made these weights in order to make them comparable to those from the BYU data, and 3.3 gives the CMU-to-BYU fit data that constitute tests for the central question of this paper: which factors that influence word division judgments can be said to have been learned from patterns in the English lexicon.

### 3.1 Corroboration of Factor Relevance

The weights calculated from the responses given in Eddington et al.'s (2013) study at BYU indicate that all of the factors described in 1.2, when expressed as phonological constraints, are active as expected in contributing to those word division judgments, with the possible exception of the stress-based constraints. In the absence of any known method of testing significance levels of the difference between two weights, I compare weights in an absolute fashion, i.e. based only on whether a weight is higher or lower than another. The table in (15) shows a subset of the resulting weights; a full discussion follows.

(15)

| | | | |
|---|---|---|---|
| *#i | 0.5320174414 | *i# | 0.6891315062 |
| *#æ | 0.5768239627 | *æ# | 1.4975683889 |
| *#C[+stress] | 0.3032418691 | *[+stress]C# | 0.3914900691 |
| *#C[-stress] | 0.3071900827 | *[-stress]C# | 0.7640361593 |
| *#[+stress] | 0.5741923413 | *[+stress]# | 0.4859441412 |
| *#[-stress] | 0.5702441277 | *[-stress]# | 0.1244878106 |
| *#[+stopglottal] | 0.4387171052 | *[+stopglottal]# | 0.4387171052 |
| *#[-stopglottal] | 0.2324364706 | *[-stopglottal]# | 0.6449977398 |
| Onset | 0.7057193637 | NoCoda | 0.7168091232 |
| Anti-Onset | 0.1717148468 | Anti-NoCoda | 0.1717148468 |

Constraints related to legality of particular segments (phonemes) at word edges exhibited constraint weights in line with expectations. For example, constraint weights for segments observed at right edges of words had an average weight of approximately 0.5, whereas constraint weights for segments not observed at right edges were much higher, e.g. *ɛ# had a weight of 1.62. Constraints like *ɛ# had among the highest weights in the set. Note that this does not hold true for a small handful of segments like [h]; the constraint *h# had a weight of only 0.439. This is transparently due, however, to the fact that there is only a single word in the included subset of the BYU data with a medial [h].

The highest weight for the [+/-stopglottalized] series is *[-stopglottalized]# (0.645, as compared to 0.439, 0.232, and 0.438 for the other constraints), as expected on the basis of the lexical generalization that word-final stops are always glottalized. Note that *#[+stopglottalized], which would also be expected to have a high weight, remains low because there are no [+stopglottalized] consonants among the intervocalic consonants in the BYU corpus.

In the [+/-vglottalized] series, the expected bearer of the highest weight, *#[-vglottalized], did emerge with a higher weight (0.706) than other constraints in the series (0.439, 0.301, and 0.310). This result corresponds to the generalization from Pierrehumbert and Talkin (1992) that word-final vowels are glottalized.

The weight of *[+anticnasal]# is higher (0.459) than that of *[-anticnasal]# (0.159), as predicted by the generalization in section 1.2. The *#[+/-anticnasal] constraints, while both of higher weight than *[+anticnasal]#, are not of interest in determining the utility of these constraints, as the nasality of a post-division vowel (e.g. ɔ/gĩ) depends only on the consonant that follows it and is not expected to affect word division judgments.

Unsurprisingly, the constraints ONSET and NOCODA had weights (0.706 and 0.717, respectively) much higher than those of ANTI-ONSET and ANTI-NOCODA (both 0.172).

The stress-based constraints alone are ambiguous in their accordance with their respective empirical finding, i.e. that a V/CV division is more likely when the second vowel is stressed and less likely when the first is stressed. Weights of the right-edge constraints *[-stress]C# and *[+stress]# (0.764 and 0.486) are, as predicted, higher than those of *[-stress]C# and *[+stress]# (0.391 and 0.124). However, the weights across the other two pairs are nearly identical: *#C[+stress] and *#C[-stress] are at 0.303 and 0.307, respectively, and *#[+stress] and *#[-stress] are at 0.574 and 0.570. I interpret these results as indicating that the majority of explanatory power for these stress-based patterns has—for whatever reason—fallen primarily to the right-edge constraints, leaving the left-edge constraints as placeholders whose weights do not greatly affect output choice.

### 3.2 Standardizing the Weight Sets

Weights generated from the CMU Pronouncing Dictionary fit rather well to the weights generated from the BYU word division judgment data: the Pearson's product-moment correlation coefficient of the two sets of weights is 0.726 ($p < 1 \times 10^{-132}$). However, the BYU data weights diverge from the CMU data weights primarily in their magnitude, and therefore in their predicted degrees of variability. Whereas weights for the BYU data range from zero to approximately 2.3, weights for the CMU data range from zero to approximately 16. It is not clear whether this is simply because of the high levels of variability in the BYU compared to CMU, or because of the higher token counts in CMU—or both. In any case, these high weights for the CMU data set mean that there is extremely little variation in the word division judgments that they predict: a larger range of weights means that differences between harmony scores will be greater, and hence their differences as log-odds quickly exceed .999 to .001, as exemplified in (16).

(16) *Observed probabilities:*

| | | |
|---|---|---|
| holy | [...oʊ / li...] | 0.81 |
| | [...oʊl / i...] | 0.19 |

*Predicted probabilities with raw CMU weights:*

| | | |
|---|---|---|
| holy | [...oʊ / li...] | 0.999999574671 |
| | [...oʊl / i...] | $4.2532863268 \times 10^{-07}$ |

One intuitive solution to this issue would be increasing the effect of the Gaussian prior that forms part of the way each constraint's weight is calculated by the MaxEnt Grammar Tool. Essentially, the lower the sigma value used for the Gaussian prior of a constraint, the more strongly the weight of that constraint will be pulled to some mean value (0 in this case), meaning that a lower sigma will force the range of constraint weights to be smaller. Indeed, decreasing the sigma value for all constraints does decrease the large difference between predicted outputs, but by bringing it low enough ($sigma^2 = 0.001$) to ensure that the ratios of probabilities among output pairs are all within observed ranges, the constraint weights cease to have relative rankings that can be considered representative of English phonotactics. For example, despite the fact that English (and the CMU Pronouncing Dictionary) includes no words that end in [h], drastically reducing the gaussian prior yields a constraint *h# whose weight, 0.15, is *smaller* than that of constraints penalizing frequently observed word-final sequences, such as *ð# (0.261) and *f# (0.214).

My solution is to standardize the CMU constraint weights so that they have the same mean and standard deviation as the BYU weights. Doing so brings the range of predicted variability in word division judgments based on the (standardized) CMU weights to the expected levels and increases the correlation coefficient to 0.814. Compare the example in (17) with the un-standardized weight predictions in (16). One alternative, of setting the CMU weights to have the same standard deviation as the BYU weights but a minimum weight of 0, produces negligibly different weights from this solution.

(17) *Observed probabilities:*

| | | |
|---|---|---|
| holy | [...oʊ / li...] | 0.81 |
| | [...oʊl / i...] | 0.19 |

*Predicted probabilities with standardized CMU weights:*

| | | |
|---|---|---|
| holy | [...oʊ / li...] | 0.788179064606 |
| | [...oʊl / i...] | 0.211820935394 |

### 3.3 Factor Effects on CMU-BYU Correlation

The central research question of this paper is which of the allophonic distinctions described in 1.2 are active in informing word division judgments because it (or rather, its weights) was learned from phonotactic patterns in the English lexicon. I test these hypotheses by investigating the effect that removing each of these categories of constraints has on the correlation between word divisions predicted from CMU weights (the English lexicon) to those based on the BYU weights (attested word division judgments). The motivating logic of this methodology is that if constraint weights learned from the lexicon do not help predict judgments, then adding those constraints to the set used to predict word divisions will have either no effect or a *negative* effect on the the correlation of those predictions to the observed word divisions, a difference which can be recovered by removing those constraints and checking the change in correlation.

Using the standardized CMU weights, I calculated the Pearson's product-moment correlation coefficient ($r$) between word divisions predicted by each sets of weights in (18) to those predicted by the optimized BYU weights. Specifically, I correlated the predicted probabilities for the first output of each input form in the BYU corpus; since the sum of the probabilities of each pair of outputs will always be 1, comparing lists containing the predicted probabilities of both outputs for each input form would artificially inflate the correlation percentage. The resulting correlations are as follows:

(18)

| *excluded constraints* | *Pearson's* r |
|---|---|
| (none) | 0.7969 |
| [stress] type | **0.8156** |
| [stopglottalized] type | 0.7428 |
| [vglottalized] type | 0.2210 |
| [anticnasal] type | **0.8639** |
| ONSET, NOCODA, etc. | **0.8088** |

Excluding the [stopglottalized] (final stop glottalization) and [vglottalized] (initial vowel glottalization) sets of constraints lowered the correlation between the two sets of predicted word divisions. These results are therefore consistent with (albeit not conclusive verification of) the hypothesis that the effect of sensitivity to glottalized vowels and glottalized stops on word division judgments could be learned from patterns in the English lexicon. These points are only evidence that it *possible* that the the grammatical weight of the glottalization contrasts could be learned from exposure to the lexicon of English rather than arising independently from some phonetic or other factor. An alternative explanation—that an external cause is responsible both for responses in the experimental task and for the corresponding patterns within the English lexicon—would predict similar results.

This conclusion contrasts with that for the [anticnasal] (anticipatory nasalization), [stress], and ONSET/NOCODA/etc. sets.. Removing each of these sets of constraints increases the correlation between the CMU-based probabilities and the BYU-based probabilities rather than diminishing it, and so their weights appear to introduce unhelpful "noise" into the calculation of word division probabilities. I conclude that these biases arise from phonetic or other non-phonological factors. This position is consistent with Cohn's (1993) conclusion that anticipatory nasalization in English is a purely phonetic, not phonological, process. I am also not aware of any arguments that stress-based influences on word division are learned correspond to lexical patterns, unlike the glottalization cases which are transparently tied to lexical phonotactics. Therefore my results confirm the implicit assumption that stress-related factors in word division arise from some phonotactics-external source. With the case of vowel nasalization, an additional possibility is that the way I have coded the CMU and BYU lexicons is phonetically inaccurate.

## 4 Conclusion

This paper has presented a simulation of the learning of principles of word division ("syllabification") in English. Its primary methodological innovation is the direct comparison of constraint weights calculated from a lexicon of forms to constraint weights calculated from responses in an experimental task for the purpose of establishing which aspects of the experimental results can be attributed to patterns in the lexicon. The results of this simulation indicate that among the tested factors that have been argued to affect word division judgments, the role of allophonic vowel and stop glottalization patterns constraints could have been learned from lexical patterns, whereas the role of anticipatory nasalization, stress position, and ONSET/NOCODA must arise from other factors.

As more studies of phonological learning are carried out, methods of assessing whether certain knowledge is a generalization based on observed phonological patterns or an independent bias will be crucial in testing the hypotheses that will emerge. In this paper I have given a preliminary description of one such method, using comparisons between the predictions of different sets of weights learned as MaxEnt Grammars.

## References


Becker, M., Nevins, A., and Levine, J. (2012). Asymmetries in generalizing alternations to and from initial syllables. *Language*, 88(2):231–268.

Burnage, G., Baayen, R. H., Piepenbrock, R., and van Rijn, H. (1990). *CELEX: A guide for users*.

Coetzee, A. W. and Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory*, 26(2):289–337.

Cohn, A. (1993). Nasalisation in english: phonology or phonetics. *Phonology*, 10(1):43–81.

Côté, M.-H. and Kharlamov, V. (2010). The impact of experimental task on syllabification judgments: a case study of russian. *Handbook of the syllable*, pages 273–94.

Derwing, B. (1992). A 'pause-break' task for eliciting syllable boundary judgments from literate and illiterate speakers: preliminary results for five diverse languages. *Language and Speech*, 35(1-2):219–235.

Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24(4):423–444.

Eddington, D., Treiman, R., and Elzinga, D. (2013). The syllabification of American English: Evidence from a large-scale experiment part i. *Journal of Quantitative Linguistics*, 20.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, pages 111–120.

Gordon, M. (2005). A perceptually-driven account of onset-sensitive stress. *Natural Language & Linguistic Theory*, 23(3):595–653.

Hall, K. (2009). *A probabilistic model of phonological relationships from contrast to allophony*. PhD thesis, The Ohio State University.

Hall, T. (2006). English syllabification as the interaction of markedness constraints. *Studia Linguistica*, 60(1):1–33.

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

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

Hayes, B. (2011a). Interpreting sonority-projection experiments: the role of phonotactic modeling. In *Proceedings of the 17th International Congress of Phonetic Sciences*.

Hayes, B. (2011b). *Introductory phonology*, volume 36. Wiley-Blackwell.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Itô, J. (1986). *Syllable theory in prosodic phonology*. PhD thesis, University of Massachusetts at Amherst.

Ladefoged, P. and Johnson, K. (2010). *A course in phonetics*. Wadsworth Publishing Company.

Legendre, G., Miyata, Y., and Smolensky, P. (1990). *Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations*. Citeseer.

Martin, A. (2011). Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language*, 87(4):751–770.

Pierrehumbert, J. and Talkin, D. (1992). Lenition of /h/and glottal stop. *Papers in laboratory phonology II: Gesture, segment, prosody*, pages 90–117.

Prince, A. and Smolensky, P. (2004). *Optimality Theory: constraint interaction in generative grammar*. Wiley Online Library.

Redford, M. and Randall, P. (2005). The role of juncture cues and phonological knowledge in English syllabification judgments. *Journal of Phonetics*, 33(1):27–46.

Redi, L. and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, 29(4):407–429.

Samuels, B. D. (2009). *The Structure of Phonological Theory*. PhD thesis, Cambridge, MA.

Steriade, D. (1999). *Alternatives to syllable-based accounts of consonantal phonotactics*. UCLA.

Tesar, B. and Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, 29(2):229–268.

Treiman, R. and Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language*, 27(1):87–104.
Weide, R. (2005). The Carnegie Mellon pronouncing dictionary [cmudict. 0.6].
Wilson, C. and George, B. (2009). Maxent grammar tool (computer software).