A Phonotactic Analysis of the Organization of Lexicons

Shiying Yang

1 Introduction

The study of phonotactics is concerned with the phonological well-formedness of phonetic sequences. A typical definition of phonotactics refers to it as the set of language specific (or some potentially universal) rules or constraints which determines if a word may or may not appear in the lexicon of a language. This grammar can differentiate words such as /blik/ and /bnik/, where the former would be recognized as a possible non-existent word for English, while the latter would be categorized as inadmissible (Chomsky and Halle 1965). These rules are generally inducted based on existing sound patterns in the lexicon. For example, since /v/, /ð/, /z/ and /ʒ/ do not appear in any English syllable onset clusters, it can be said that any word with these sounds in an onset cluster would be impermissible in terms of English phonotactics (Giegerich 1992).

There are, nevertheless, non-categorical patterns within languages which rule-based grammars cannot easily describe. For existing words of a language, there are differences in how typical a sound sequence is in the lexicon. For instance, in English, there are more words with the sequence /ʌf/ than would be expected by the product of individual probabilities of the occurrence of /ʌ/ and /f/; while the sequence /ael/ occurs much less often than would be predicted from the probabilities of its subparts (Kessler and Treiman 1997). For “possible non-existent” words that do conform to observed patterns in a lexicon, it is also worth exploring if their non-existence is purely by chance, or if some other constraints play into the lexicon favoring some phonological sequences over others. Along this line of thought, the phonotactics of a language should also be able to evaluate how typical or representative a word is of the language, which involves the probabilistic knowledge of the sequencing of subparts of words.

Well-formedness judgments by speakers are commonly used as a proxy for the phonotactic acceptability of phonological sequences. The gradience in well-formedness judgments of non-words have been used to validate a wide range of phonotactic models over subparts of sequences focusing on syllables (e.g. Coleman and Pierrehumbert 1997), segments (e.g. Frauenfelder et al. 1993), subsegmental features (e.g. Albright 2009b) and phonological constraints (e.g. Hayes and Wilson 2008). Besides the grammaticality of the composition of words, the gradience in well-formedness judgments (or wordlikeness judgments in this context), can also be attributed to an effect in online processing due to the perceptual and subjective nature of this measurement. The more similar a non-word is to existing words in the mental lexicon, the higher its rating would be (e.g. Greenberg and Jenkins 1964).

The phonotactic and the perceptual account of the gradience in well-formedness judgments give rise to different interpretations of the psychological reality of phonotactic knowledge and the organization of lexicons. Yet effects of these factors are hard to separate due to their strong correlation with each other. The current paper aims at summarizing arguments from the two perspectives, and at providing cross-linguistic evidence and a phonotactic explanation to an observed characteristic of lexicons. In Section 2.1, I will examine variations of phonotactic models and their relations to well-formedness judgments as well as other tasks. Factors concerned with processing will be discussed in Section 2.2. Attempts at disentangling effects from the two accounts are summarized in Section 2.3. Section 2.4 follows up on the comparison of the two accounts by focusing on the issue of word clustering observed in lexicons (Dautriche et al. 2017). This observation is tested in more languages in Section 3.

2 Literature review

2.1 Phonotactic models and well-formedness

Well-formedness judgments of words, which are also referred to as acceptability judgments or wordlikeness judgments in the literature, are often seen as the manifestation of the implicit knowledge of phonotactics. In light of the gradience observed in well-formedness judgments, rule-based or constraint-based phonology (Optimality Theory; Prince and Smolensky 2004) would not be adequate in describing the issue other than associating it with performance rather than competence. Stochastic phonology, however, views phonotactics and the cognitive representation of sound structure in general as probabilistic rather than definitive (Pierrehumbert 2001). This approach makes it possible to evaluate neologism as an extension of the real lexicon. By taking into account the combinatorial properties of sounds in the lexicon, phonotactic probabilities predicted by these phonotactic models can take on similar gradient patterns as seen in well-formedness judgments. Correlation between these probabilities and well-formedness ratings are thus used to substantiate the role of various phonological units in phonotactics.

Coleman and Pierrehumbert (1997) proposed a phonotactic model where probabilities of words were defined as the product of probabilities of their syllable constituents. Log probabilities of mono- and di-syllabic non-words calculated according to this model were shown to have the “best” correlation with the number of negative responses compared to unlogged probabilities and the lowest constituent probability, despite their significant correlations with the data. Frisch, Large, and Pisoni (2000) tested the same model on ordinal well-formedness ratings and binary judgments of multi-syllabic non-words to reveal similar results which reaffirmed the relevance of syllable constituent frequencies for phonotactic evaluation. They further examined correlations of well-formedness scores with other predictors including log probabilities over segments and the log number of neighbors, and found the 3 predictors resulted in similar higher correlation coefficients than probabilities of the worst syllable or the worst segment. The two studies highlighted the importance of considering the entire word in well-formedness judgments rather than a single worst part. This served as evidence against rule-based or constraint-based grammar in phonotactic evaluations since both accounts would disqualify a word over the violation of one rule or one high-ranking constraint by a subpart.

Similarly, CVC syllables of the high rime frequency group were shown to have higher well-formedness ratings than those of the low rime frequency group by both adults and elementary school-aged children (Treiman et al. 2000). In the same study, this result was extended to blending tasks where high-frequency rimes were preserved more than low-frequency rimes. Nevertheless, probabilities of inter-syllabic clusters were also shown to be correlated with well-formedness scores of words containing these patterns (Hay, Pierrehumbert, and Beckman 2004). This finding suggests that there are actually effects of frequencies across syllable boundaries.

On the other hand, probabilities over segments or probabilities assigned by n-gram models (Jurafsky and Martin 2008) in general have often been used as baseline phonotactic models in phonological and perceptual studies with various assumptions. Bi-phone probabilities have been shown to be at least just as predictive as probabilities calculated over syllable constituents (Frisch, Large, and Pisoni 2000) or natural classes (Albright 2009b) of well-formedness scores on various sets of experimental stimuli.

Not only does the preference for higher segmental- and sequential-probability words occur in well-formedness judgment tasks, Jusczyk, Luce, and Charles-Luce (1994) showed that 9-month-old infants are already sensitive to segmental probabilities in the language since they looked longer at CVC non-words with higher segment and bi-phone probabilities than low-probability non-words. Moreover, Vitevitch et al. (1997) adopted the same set of CVC syllables and phonotactic models to demonstrate that not only are high-probability di-syllabic non-words rated higher by adults in terms of well-formedness, they also induce lower reaction time in an auditory repetition task (see also Vitevitch and Luce 2005). Such findings suggest that the facilitative effect of segmental probabilities is rather robust to changes in test paradigms.

Albright (2009b) proposed a phonotactic model which attributed phonotactic probabilities to both bi-phone probabilities and the probabilities of segments being analyzed in terms of natural classes. Log natural class-based bi-phone probabilities and log segment-based bi-phone probabilities were used to fit judgement data on onsets and monosyllabic non-words. Results showed that the feature-based model was predictive of well-formedness judgments both on attested and unattested sequences. It makes an independent contribution even though it does not replace the segmental model which were better predictors in terms of attested sequences. Compared to models that calculate probabilities based on segments or clusters, the advantage of a subsegmental model is the ability to differentiate between non-words whose subparts are unattested in the lexicon. The positive and independent effect of incorporating feature-based probabilities implicates the involvement of subsegmental analysis in phonotactic evaluation. This implication further justifies that the gradience in well-formedness judgments arises out of grammatical reasons that cannot be subsumed by influences in online processing.

Another way of accounting for features in gradient well-formedness judgments is using maximum entropy models which assume that the logarithm of the probability of a wordform is the linear combination of orthogonal weighted constraint violations (e.g. Jäger and Rosenbach 2006). Hayes and Wilson (2008) developed an algorithm that learns the set of constraints and assigns weights (penalties) to constraint violations along the form of a MaxEnt grammar. Phonotactic probabilities based on this model was tested on data of English speakers' judgments of onsets and demonstrated more accurate performance than probabilities calculated over syllable constituents and segments. However, the model did a better job in distinguishing between unattested forms and could not tell apart attested onsets.

2.2 Processing and well-formedness

Another line of research regarding well-formedness judgments stemmed from the idea that the more similar a non-word is to real words, the more readily can speakers accept it as a well-formed word to the lexicon. Greenberg and Jenkins (1964) used a measurement of similarity dependent on the number of valid word types that can be obtained from substituting certain numbers of segments from a non-word. They found that CCVC monosyllabic non-words that were rated higher by this similarity metric induced higher well-formedness scores. This metric is along the same lines with other measures of perceptual similarity such as neighborhood density and edit distance. Such a result would therefore imply that non-words with denser neighborhoods would be rated as more acceptable than those with sparser neighborhoods.

Moreover, well-formedness judgments were found to be significantly correlated with repetition accuracy, this phenomenon was termed as *wordlikeness effect* (Gathercole and Martin 1996). Non-words of higher phonotactic probabilities also lead to better recognition memory performance (Frisch, Large, and Pisoni 2000). Along with the implication regarding neighborhood density, these results suggest that higher neighborhood density and higher phonotactic probability should have a facilitative effect on word recognition. However, literature on spoken word recognition and lexical neighborhoods came to almost the opposite conclusion. Higher neighborhood density actually has an inhibitory effect on spoken word recognition, which was explained by competition in activation from neighbors in the mental lexicon (Goldinger, Luce, and Pisoni 1989; Luce and Pisoni 1998).

Further investigations into these phenomena revealed that for real words in the lexicon, higher neighborhood density leads to inhibition in word recognition; while for non-words, higher phonotactic probability would lead to faster recognition (Vitevitch et al. 1997; Vitevitch and Luce 1998, 1999). Vitevitch and Luce (1999) hypothesized that the reversed effect observed in words and non-words could be attributed to different dominant mechanisms in the processing of these items. Namely, real words would be processed mainly at the lexical level, while non-words would mainly be processed at the sublexical level due to their lack of representation in the mental lexicon. To test this hypothesis, they mixed words and non-words in a same-different task to elicit sublexical processing for real words assuming that the adopted strategy would depend on the most common type of words in the list of stimuli. Similarly, an auditory lexical decision task was performed with non-words to elicit lexical processing. Results showed that real words with higher neighborhood density had less inhibition on recognition and non-words with higher phonotactic probabilities were responded to more slowly rather than more quickly under these manipulations.

Taken together, these results indicate that both lexical and sublexical mechanisms are involved in the processing of real words and non-words. Sublexical information, which could be attributed to probabilistic properties captured by phonotactics (as discussed in Section 2.1), plays a more important role in speakers' well-formedness judgments of non-words, despite the prominence of neighborhood density effects. Nevertheless, well-formedness ratings in practice can be affected by the way they are elicited. This idea would be relevant again in Section 2.3.

2.3 Comparison between the phonotactic and the lexical account

Phonotactic models that focus on statistical properties within words imply that higher well-formedness ratings would be attributed to subparts that are commonly observed in the lexicon. Lexical models that focus on similarity put more emphasis on how much overlap a sequence overlap has with other words in the lexicon. As shown in previous sections, results obtained along these two approaches are highly correlated. Especially for non-words, high phonotactic probability and high neighborhood density both have positive correlations with well-formedness ratings. This is hardly surprising, since the two approaches represent two metrics of the same idea which is how representative a phonological sequence is of the lexicon of a language. Even though ways of calculating phonotactic probabilities and neighborhood densities are independent from each other, words with many neighbors that are one segment away from it would inherently have more frequent subparts.

As a result, effects of the two accounts on well-formedness ratings are hard to separate. Bailey and Hahn (2001) created stimuli of monosyllabic non-words that were generated based on a random process to investigate more generally the individual and joint predictive power of phonotactic probabilities and similarity measures on well-formedness judgments. They found that their Generalized Neighborhood Model (GNM), which involves both the number of neighbors that are 1- and 2-phoneme away and the similarity between individual corresponding phonemes, accounted for more variance than the simple count of neighbors of single edit distance and other probabilistic phonotactic models. They thus concluded that both phonotactics and lexical processing have unique effects on wordlikeness ratings, which implies distinct cognitive sources for this task, but lexical similarity is the more important factor.

In light of this result, Shademan (2006) proposed that the observed difference in contributions of the two types of factors in Bailey and Hahn (2001) could have stemmed from the design of experimental stimuli where fillers of real words were added. In this study, in the experiment where real words were not a part of the stimuli, only phonotactic probabilities over syllable constituents had a significant main effect. Only in the experiment where real words were a part of the stimuli did lexical similarity show a significant main effect along with phonotactic probabilities. From these results, Shademan (2006) concluded that the contribution of lexical similarity to well-formedness ratings can vary depending on experimental design, which is in accordance with the theory of differences in adopted lexical or sublexical mechanisms in the presence of different stimuli (Vitevitch and Luce 1999); while the effect of phonotactic probabilities was invariant across the experiments.

GNM was also fitted to different datasets in Albright (2009b, see also 2007) in addition to phonotactic models over segments and natural classes as discussed in Section 2.1.

Similar to Bailey and Hahn (2001), this study also showed unique effects of lexical similarity and phonotactic probabilities on well-formedness ratings. However, phonotactic probabilities, especially log bi-phone probabilities, rather than lexical similarity were found to account for the bulk of variance in well-formedness ratings. Besides the influence of real words pointed out in Shademan (2006), the discrepancy was also attributed to the lack of non-words on end-points of the well-formedness scale in the original stimuli since all test items in Bailey and Hahn (2001) were either one- or two-phoneme away from real words. Additionally, it was pointed out that lexical models would fail at distinguishing unusual words with few neighbors from implausible sequences.

To summarize, lexical similarity and phonotactic probabilities both play significant roles in the task of well-formedness judgments, and it is still hard to delimit lexical effect from effects driven by sublexical statistical information. Therefore, both factors should be taken into account in theorizing about higher level issues. In addition, it should be reliable to see well-formedness as a measure of phonotactic acceptability since their correlation is less susceptible to changes in experimental design.

2.4 A phonotactic explanation of the clustering of lexicons

The two accounts for well-formedness judgments and their implications on phonological theory and processing can be used to shed light on broader issues such as the emergence or composition of lexicons. Constraint-based models have the concept "Richness of the Base" which assumes that for any language, all inputs are possible (Prince and Smolensky 2004). So the actual lexicon of a language can be thought of as an output set that includes inputs that pass the language-specific ranking of constraints. In line with this conception of lexicons, a more general way of thinking about available wordforms in lexicons is that they are selected from a pool of candidates that conform to the phonotactics of the language.

Based on this assumption, Dautriche et al. (2017) set out to explore whether lexicons across languages would be sparser or more clustered due to competing functional pressures for distinctiveness and regularity. They adopted 5-phone phonotactic models over segments for each language and created baseline lexicons assuming that wordforms would be randomly selected from the pool of candidates approved by the 5-phone model according to their phonotactic probabilities. In order to test which functional pressure has the more prominent effect on lexicons, a range of metrics on overall wordform similarity including the number of minimal pairs, average string edit distance and network measures of phonological neighborhoods were utilized to compare the real lexicon to baseline lexicons. Results from these comparisons suggest that the real lexicon is more clustered than would be expected by general phonotactic concerns. This preference for regularity observed in Dutch, English, French and German lexicons was attributed to processing advantages for words with dense neighborhoods in retrieval, memory and acquisition (e.g. Vitevitch 2002), despite their disadvantage in word access. Potentially these processing preferences would shape a phonotactically-filtered base to become a lexicon that is more clustered than sparse.

Rather than relying entirely on a processing account for explaining the clustering in lexicons, a more refined system of phonotactics can also contribute to such observed discrepancy between the real lexicon and the phonotactically-controlled baseline.

Different ways of quantifying phonotactic probabilities discussed in Section 2.1 so far, despite their distinct levels of analyses, recognize that the evaluation of a word is contingent on the acceptability of all subparts. Also, the ways these acceptability metrics are defined are all based on the assumption that these subparts are independent from each other, and that their probabilities or constraint violations can be linearly combined to render the overall probability or score of a sequence.

One important implication of considering phonotactics as stochastic is the need to distinguish between accidental or systematic gaps in the lexicon. Pierrehumbert (1994) used the product of probabilities of word-initial and word-final consonants and clusters to predict the occurrence of word-medial clusters in the English lexicon and showed that clusters with lower predicted probabilities are less accepted by speakers and indeed do not show up in the lexicon. However, Frisch (1996) pointed out that at least a certain subset of low-probability clusters as categorized by Pierrehumbert (1994) actually did not have joint probabilities so low that would completely prevent any of them from surfacing. Seeing this seemingly systematic absence of lower-probability clusters, Frisch (1996) speculated that combinations with higher expected frequencies have more exemplars in the lexicon and would be modeled after more in the event of selecting a new word, which drives a “rich get richer” and “poor get poorer” effect.

Albright (2009a) looked into similar patterns in Lakhota where clusters, fricatives, aspirates and ejectives do not tend to co-occur in the initial and medial onset positions. Analysis on Lakhota mono- and di-syllabic roots confirmed that almost any combination of these structures occur less often than expected, with the degree of under-attestation rising as the frequency of individual structure decreases. These findings are in accordance with the “poor getting poorer” effect proposed by Frisch (1996). However, it was also found that another rare structure in Lakhota, nasals, do not conform to the same pattern as other rare structures. Combinations involving nasals all have frequencies around expected values. Albright (2009a) suggested that nasals are not in themselves dispreferred in Lakhota, they only occurred less frequently because there are fewer of them. On the other hand, clusters, fricatives, aspirates and ejectives are dispreferred by the grammar, and this dispreferrence is correlated with their frequencies in the lexicon. In other words, only grammatically dispreferred structures would induce such "superadditive" effect of penalties, which leads to the under-attestation of their co-occurrences. The evidence regarding nasals is used to establish this effect as being prompted by the grammatical status of structures, rather than frequencies themselves. Going back to the emergence of lexicons, the language-specific superadditive penalties on patterns would be imposed on inputs on top of the previously defined language-specific stochastic grammar to filter candidate wordforms for the lexicon.

Similar patterns of superadditivity of constraints were also attested respectively in Colloquial Bambara and Dioula d'Odienné by Green and Davis (2014) and Shih (2017). In particular, Shih (2017) adopted a model that encapsulates superadditivity along the lines of a MaxEnt model as described in Section 2.1. The only difference was that weights were also assigned to interaction terms (conjunctions) of constraints rather than only to individual constraints. Results showed that a model with some constraint conjunctions improves the explanatory power of the grammar without driving up complexity. These instances together point to the possibility that superadditivity of penalties on certain structures can be prevalent in languages.

The next section explores the clustering of lexicons from the perspective of superadditivity in phonotactics. If there are superadditivity effects in a language, lexicons generated based on probabilities from additive models would include expected occurrences of sequences that are supposed to be under-attested in the real lexicon. Since superadditivity effects are usually associated with structures that are of low frequencies to begin with, the surfacing of under-attested patterns in generated lexicons would lead to more words with lower probabilities than in the real lexicon. Whereas if there is no superadditivity effect in a language, current additive models should be sufficient in capturing the phonotactics of the language. Lexicons generated from these models should have the same distribution of structures and frequencies as the real lexicon. This will serve as the null hypothesis for the study in Section 3.

3 The distribution of phonotactic probabilities in lexicons

3.1 Methods

3.1.1 Materials

Six languages from different language families were used in the current study. Lexicons of different languages were studied based on their core vocabularies in order to eliminate potential influence from atypical infrequent words or borrowings. Function words and proper names were filtered out since closed class words tend to have high frequencies and unique phonological properties that are not representative of the rest of the lexicon. Since I only used the phonological representations of words, homophones were counted once in the lexicon with all frequencies summed up. For each language, wordforms whose frequencies were greater than or equal to the $5000th$ highest frequency in the spoken corpus were selected to constitute the training dataset.

Frequencies and phonological representations of American English words were taken from the CMU Dictionary (Weide 2008). I used the LDC lexicons with corresponding frequency data for Egyptian Arabic and Japanese (Gadalla et al. 1998; Kobayashi et al. 1997). The Korean Telephone Conversations Lexicon (Han 2003) was used for Korean, but word frequencies were taken from the OpenSubtitle Corpus (Lison and Tiedemann

Table 1: Information on filtered lexicons

	Size of lexicons	% of word types in lexicons	% of word tokens covered in lexicons	Cutoff frequency
English	5030	15.76 %	96.47 %	94
German	5004	5.64 %	84.60 %	279
Spanish	5001	3.16 %	82.05 %	12496
Arabic	6003	33.26 %	88.66 %	3
Japanese	5153	100.00 %	100.00 %	1
Korean	5003	27.06 %	94.38 %	78

[a] All Japanese words were used due to a lack of frequencies in the Callhome transcripts

2016) due to the small amount of frequencies in the Korean Telephone Conversation Transcripts. The OpenSubtitle Corpus was also used for German and Spanish for word frequency counts, while their phonological representations were transcribed using the text-to-speech software eSpeak (Duddington 2014). Table 1 shows that filtered lexicons used for the study only include a small amount of word types appearing the each corpus, but despite the exclusion of function words, they still represent the majority of word tokens occurring in the corpora.

3.1.2 The language model

After function words, proper names and low-frequency words were filtered out, a phonotactic model was trained for each language in the form of a n-gram language model over segments.[1] In a preliminary study conducted on monosyllabic words in American English and Mandarin (Yang, Sanker, and Cohen Priva 2018), tri-phone models were used. I used stricter 4-phone models for the current study due to the inclusion of

[1] The choice for a phonotactic model that only considers frequencies, but not specific grammatical structures is due to both the practical efficiency of working with multiple languages, and the fact that the change in focus would not alter predictions made according to the hypothesis. For structures that are assigned low probabilities due to their accidental lack of attestation in the lexicon, their presence in baseline lexicons would be the same as their occurrences in the real lexicon.

multisyllabic words and potential long-distance dependencies such as vowel harmony across syllables. The extension of the conditioning environment further constrained the lexical space of candidates for each language.

The probability of a word is, therefore, defined as the product of conditional probabilities of each segment in the phonological representation of the word given the previous 3 segments, which is represented by the number of occurrences of the 4-segment sequence divided by that of the preceding 3-segment sequence in the lexicon (as common practice, the beginning and end of words were also counted as segments in the calculation). The number of occurrences of sequences were counted based on word types in the lexicons, not on word tokens in the corpora since multiple studies agreed on the positive correlation between word probabilities calculated from type frequencies and well-formedness judgments (Greenberg and Jenkins 1964; Treiman et al. 2000; Hay, Pierrehumbert, and Beckman 2004); while the incorporation of token frequencies in the model often does not drastically improve the predictive power of the model (Bailey and Hahn 2001; Albright 2009b). In order to take stress into consideration, with the exception of Japanese and Korean, vowels with primary stresses are counted separately from their unstressed counterparts.

Phonotactic probabilities were computed in log format where log probabilities of subparts were combined additively. In later comparisons between phonotactic probability distributions, log probabilities were also used. Besides the avoidance of numerical underflow, this choice was made in accordance to previous findings, where logarithmic scales for probabilities have been found to correlate stronger with gradient well-formedness ratings (e.g. Coleman and Pierrehumbert 1997; Frisch, Large, and Pisoni 2000) and account for more variance than unlogged probabilities (e.g. Bailey and Hahn 2001).

3.1.3 The sampling procedure

The 4-phone model of a language using transitional probabilities permits all 4-phone sequences that are present in the real lexicon. Based on these context-segment sequences that are found in the lexicons, phonotactically-acceptable words were enumerated up to 8 segments with their probabilities assigned by the 4-phone model so that all phonotactically-plausible words were listed out. Longer words were not taken into consideration for computational convenience. These enumerated words together make up the pool of words for the lexicon to choose from.

In the real lexicons to be later compared to phonotactically-controlled sample lexicons, there are 4164 English words that were enumerated, 3921 words in German, 3402 words in Spanish, 4668 words in Arabic, 4179 words in Japanese and 4001 words in Korean.

The sampling process of baseline lexicons followed certain rules. Each lexicon has an identical composition as the real lexicon in terms of the number of words with specific numbers of segments and syllables. Within each subgroup of a specific segment and syllable combination, words were randomly selected from the pool of enumerated words without replacement according to probabilities attributed by the 4-phone model. Under such circumstances, each sample lexicon would be identical to the real lexicon with respect to their sizes and the distribution of word and syllable lengths.

Sample baseline lexicons of a language were generated under the assumption (as discussed in Section 2.4) that the lexicon of a language is formed by randomly selecting from all phonotactically-plausible candidates based on their probabilities. As discussed in Section 2.4, If the language model which evaluates the probability of a word by sequentially and independently incorporating its subparts captures the nature of phonotactics, sample baseline lexicons would be no different from the real lexicon. Since the distribution of log probabilities are of concern in the current study, the distribution of log probabilities in sample lexicons and that of a real lexicon should be from the same population. Statistical analysis was conducted to test this null hypothesis.

3.2 Results

3.2.1 General Trends

Since the question of concern is if log probabilities of words in the real lexicon is sampled from the same distribution as those in the sample lexicons, hypothesis testing for distributions was adopted rather than hypothesis testing for parameters. As will be shown later, the distributions of log probabilities of real lexicons are highly skewed. Therefore, in order to see how the real lexicon might differ from the baseline, test statistics which are functions of the variables including the median and other quartiles were chosen. The *p*-values in these cases would depend on where the observed values of these parameters (given the real lexicon) lie in the cumulative distributions of these parameters which are in themselves random variables.

No assumptions can be made about these distributions. Their cumulative distribution functions were therefore estimated using empirical distribution functions of these parameters given a large amount of sample lexicons. The reason for relying on empirical distribution functions is that they are guaranteed to asymptotically converge to the cumulative distribution functions (based on the strong law of large numbers).

As far as the quartiles are concerned, Table 2 shows that the distributions of log probabilities in every real lexicon are not the same as those of any baseline lexicons in terms of at least 1 parameter. For English and German, given $10,000$ simulated values of each quartile in the baseline distribution, quartiles of the real lexicon fall

Table 2: Quartiles of the distributions of log probabilities and estimated 95% intervals (based on $10,000$ sample lexicons)

language	parameter	lower bound (2.5%)	upper bound (97.5%)	real lexicon	percentile	
English	Q1	-10.246	-10.133	-10.015	100.00 %	*
	median	-9.210	-9.111	-8.993	100.00 %	*
	Q3	-8.523	-8.523	-8.523	100.00 %	*
German	Q1	-10.117	-10.022	-9.904	100.00 %	*
	median	-9.211	-9.211	-9.147	100.00 %	*
	Q3	-8.613	-8.565	-8.518	100.00 %	*
Spanish	Q1	-10.845	-10.715	-10.855	1.32 %	*
	median	-9.775	-9.692	-9.707	89.10 %	
	Q3	-8.923	-8.923	-8.923	99.77 %	*
Arabic	Q1	-10.867	-10.779	-10.859	5.73 %	
	median	-9.903	-9.818	-9.799	100.00 %	*
	Q3	-9.142	-9.105	-9.057	100.00 %	*
Japanese	Q1	-11.958	-11.821	-11.842	91.15 %	
	median	-10.285	-10.179	-10.234	55.98 %	
	Q3	-8.968	-8.953	-8.953	99.70 %	*
Korean	Q1	-10.869	-10.746	-10.721	99.47 %	*
	median	-9.585	-9.516	-9.552	31.18 %	
	Q3	-8.643	-8.613	-8.623	63.13 %	

[a] Real lexicon parameters outside of the estimated 95% intervals are marked with "*";

[b] Real lexicon parameters greater than all values of the corresponding parameter in sample lexicons are marked as 100% in the "percentile" column;

[c] The English Q3, the Spanish Q3 and the German median are the same for all sample lexicons due to the discrete nature of log probabilities generated from the language model and their clustering around certain values. The same goes for the Japanese Q3 which only falls on very few values

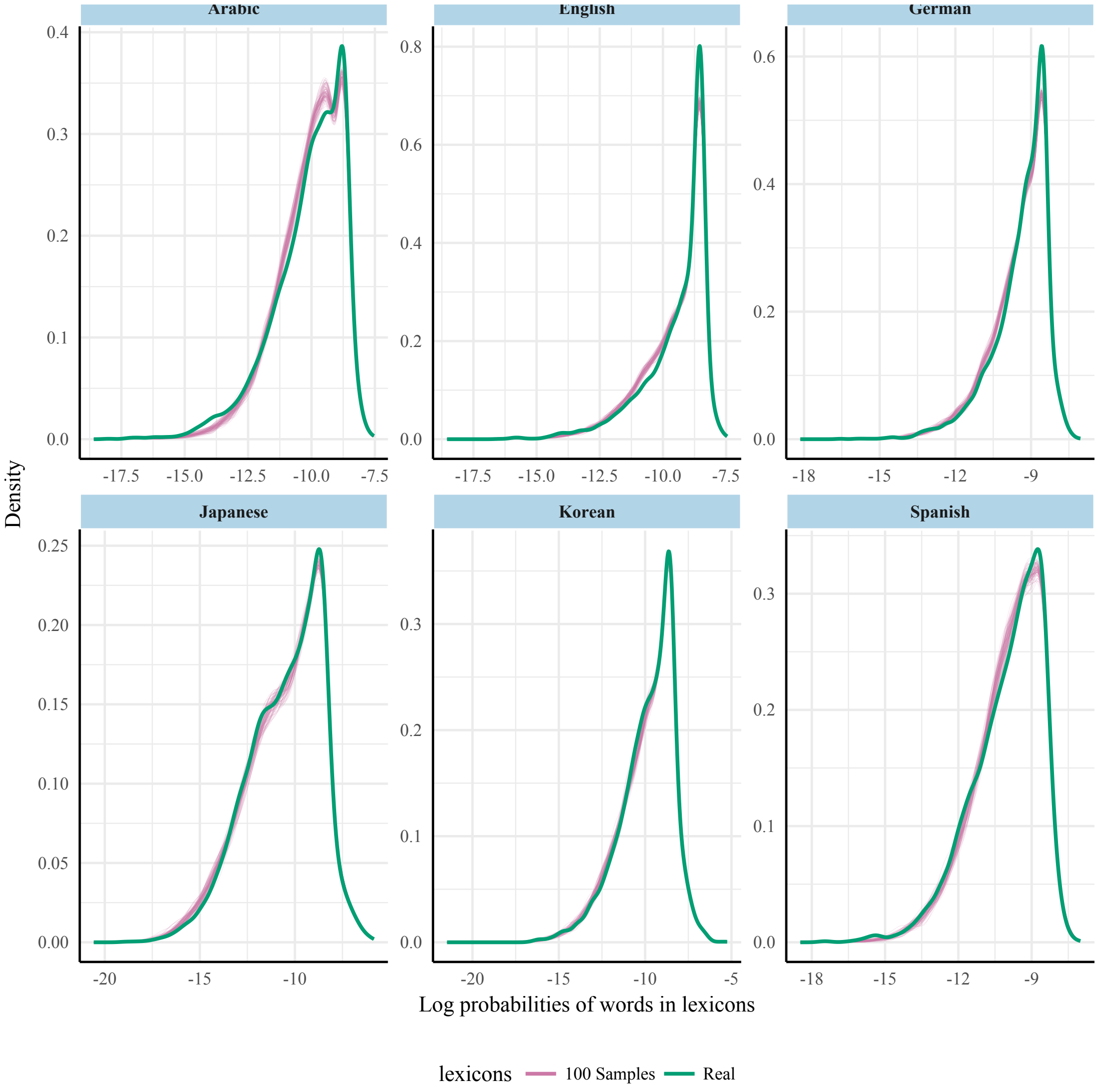

Figure 1: Probability density distributions of log probabilities from the real lexicon and 100 simulated sample baseline lexicons

exceed the upper bound for each of these parameters. This indicates that densities of log probabilities of the real lexicons for these 2 languages are overall more shifted to higher values than their baseline counterparts. Such a trend is captured in Figure 1 where density distributions of 100 sample lexicons for each language are represented. The peaks in the probability density distributions of real English and German lexicon which lie in the range of high log probabilities stand out from the generated baseline.

Similar to the results of English and German, Arabic median and Q3 are greater than those of any sample lexicons, which also suggests that the body of the density distribution of the real lexicon is more clustered around higher values than that of the simulated baseline distributions. The Arabic Q1, however, is lower than the majority of simulated Q1s (at 5.73%), which indicates that there are also more lower-probability words in the real lexicon. This corresponds to the fatter tail of the real lexicon in Figure 1.

The Spanish Q1 of the real lexicon is significantly lower than expectation for a significance level of 0.05, which also corresponds to a fatter tail in Figure 1. The Spanish median is greater than the majority of simulated medians (at 89.1%), while the Q3 is the same as those of the sample lexicon distributions. Given that the Q1 is especially low, the higher median and the comparable Q3 still indicates that the body of the real lexicon distribution shifts more to the right than expected.

For Japanese, however, there is not sufficient evidence to say that the real lexicon has a distribution significantly different from the baseline distribution. For Korean, the only quartile of the real lexicon that stands out from the sample lexicons is Q1, which is significantly higher. As shown in Figure 1, the distributions of the real lexicons of both these languages align almost perfectly with distributions of sample lexicons.

3.3 Discussion

Analysis of 6 languages shows that more than half of languages tested in the current study have lexicons with more higher-probability words than expected by 4-phone phonotactic model over segments. This general shift towards higher phonotactic probabilities displayed in English, German, Spanish and Arabic lexicons is compatible with predictions of the superadditivity account: A 4-phone phonotactic model is not adequate in modeling the phonotactics of these languages; there are under-attestation of words in real lexicons, and their phonotactic probabilities according to this model are on the lower end of the distribution. The significance of this finding is that even if real lexicons behave differently than baseline lexicons, there is no inherent motivation for them to shift to the same side of higher probabilities, as observed in 4 languages in the results. The set of languages studied here is still too small to draw any definite conclusions. However, the potential cross-linguistic existence of such a trend does provide partial evidence for the prevalence of superadditivity in languages.

There are limitations to using a simple baseline n-gram model for the estimation of phonotactic probabilities. Under common practice, the probability of a sequence is defined as the product of transitional probabilities of its subparts. This means that the model would assign zero probability to any sequence with any subparts that may accidentally not be present in the lexicon. Given a fixed size of the lexicon, the higher the n, the more likely such accidents would occur.

For the current study with $n = 4$, therefore, the conservative language model would theoretically only generate a subset of words that can potentially appear in the lexicon of a language. Moreover, this subset contains only words that are most representative of existing words, which means it would overfit the original lexicon. However, real lexicons of languages are still shown to be more or less different from baseline lexicons.

Additionally, over half of the investigated lexicons are significantly more shifted to higher probabilities, which can imply some systematic predisposition that calls for the exploration of more languages. With such restrictions in the degrees of freedom, this finding further demonstrates the inadequacy of modeling subparts of phonological sequences as independent or equal components to the phonotactic probability or the well-formedness of a word. Nevertheless, in light of discoveries made with the current model, further research can employ more refined models that respectively incorporate different levels of phonological information to draw more concrete conclusions about what clusters, constraints or subsegemental features are under- or over-attested in the real lexicon.

Practically speaking, the 4-phone model used here was unable to capture the fine-grained differences for words with less than 4 segments due to the calculation of conditional probabilities, which renders a certain amount of overlap between the real lexicon and generated lexicons in higher probabilities. This can be improved by incorporating back-offs into the model.

The significant deviance of several lexicons from their presumed phonotactic baseline implies that structures which induce superadditivity effects in these languages can be found within the range of words whose probabilities under the current model were under-attested. Therefore, more detailed analysis of what words and structures fall under this range should provide more insight into superadditivity in each language.

The Japanese results might be incomparable to results of other languages in the current study due to the lack of representativeness of words given the small number of frequencies in the corpus (as shown in Table 1). Yet Korean also does not provide sufficient evidence to reject the null hypothesis. This may call for more detailed analysis in words of different lengths in Korean. However, if this result is taken as an indication of the adequate explanatory power of an additive phonotactic model for Korean, then it would mean that there are no superadditivity or very few superadditivity effect to be found in the Korean lexicon. It would be important to see which unique characteristics of Korean can result in such conformity to probabilities.

Another interesting observation in the distributions is that towards the very low end of modeled phonotactic probabilities, there is an obvious over-attestation of words in the real lexicons of Arabic, English and Spanish (shown more clearly in Figure 2 in the appendix). In other words, there are words that are deemed of very low probabilities by the 4-phone model that are over-attested in these languages. Further analysis can look into structures of real words across this probability range to see how they differ from other low-probability wordforms, and how such differences drive their surfacing in lexicons.

4 General Discussion

Cross-linguistic analysis in the previous section show that lexicons of English, German, Spanish and Arabic have more high-probability words and less words of a lower-probability range than expectations of additive phonotactic models. This result is in line with predictions of superadditivity in phonotactics, which argues that the conjunction of dispreferred (low-frequency) structures would receive higher penalties that lead to their under-attestation in lexicons.

Results from Section 3 and results from Dautriche et al. (2017) are in practice very similar to each other. The two studies both used n-gram models as baseline phonotactic models to generate sample lexicons. The current study used 4-phone models and compared density distributions of phonotactic probabilities, while Dautriche et al. (2017) used 5-phone models and compared measures of similarity. Since similarity measures such as neighborhood density and phonotactic probabilities are highly correlated, the observation of a clustered real lexicon is more or less the same as the observation of a real lexicon where more words are of higher phonotactic probabilities than expected. Nevertheless, just as the study of well-formedness judgments, differences in metrics lead to different perspectives in the interpretations of these results. Therefore, the clustering observed in real lexicons can both be interpreted in terms of a functional pressure for lexicons to have more easily producible words, and in terms of superadditive penalties on dispreferred structures that prevent them from surfacing.

The study in Section 3 does not find the same clustering across all examined languages. For English and German which were also studied in Dautriche et al. (2017), discrepancies between real lexicons and samples lexicons were the most prominent. Yet the current null results, especially with Korean, imply that the clustering of lexicons regarding similarity measures would not be obvious in these lexicons either. In short, the overall positive results in Dautriche et al. (2017) can have something to do with their choice of closely-related languages which happen to display significant patterns. Therefore, both the superadditivity explanation and the processing explanation need to account for cross-linguistic variation in the degree to which clustering happens in lexicons.

Despite being separate explanations, a phonotactic account of clustering in lexicons is by no means a rejection of attributing such a phenomenon to functional predispositions such as efficiency in word retrieval and word learnability. On the contrary, the proposal that grammars would penalize combinations of constraints cross-linguistically would suggest that there are some phonetic or perceptual motivation.

Appendix

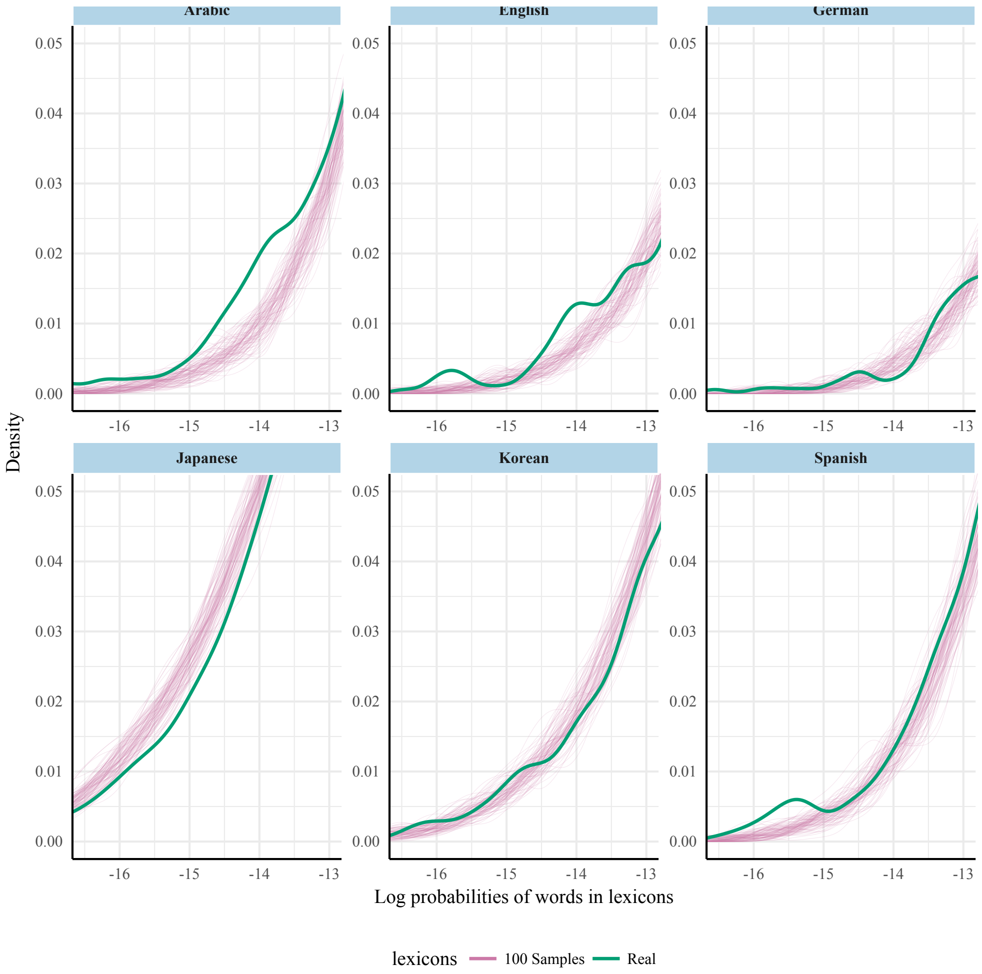

Figure 2: Probability density distributions of log probabilities from the real lexicon and 100 simulated sample baseline lexicons (zoomed)

References

Albright, A. 2007. “Gradient Phonological Acceptability as a Grammatical Effect.” Unpublished manuscript.

———. 2009a. “Cumulative Violations and Complexity Thresholds.” Unpublished manuscript.

———. 2009b. “Feature-Based Generalisation as a Source of Gradient Acceptability.” *Phonology* 26 (1). Cambridge University Press: 9–41.

Bailey, T. M., and U. Hahn. 2001. “Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods?” *Journal of Memory and Language* 44 (4). Elsevier: 568–91.

Chomsky, N., and M. Halle. 1965. “Some Controversial Questions in Phonological Theory.” *Journal of Linguistics* 1 (2): 97–138.

Coleman, J., and J. Pierrehumbert. 1997. “Stochastic Phonological Grammars and Acceptability.” In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*, edited by John Coleman, 49–56. Somerset, NJ: Association for Computational Linguistics.

Dautriche, I., K. Mahowald, E. Gibson, A. Christophe, and S. T. Piantadosi. 2017. “Words Cluster Phonetically Beyond Phonotactic Regularities.” *Cognition* 163: 128–45.

Duddington, J. 2014. *eSpeak* (version 1.48.03). http://espeak.sourceforge.net/.

Frauenfelder, U. H., R. H. Baayen, F. M. Hellwig, and R. Schreuder. 1993. “Neighborhood Density and Frequency Across Languages and Modalities.” *Journal of Memory and Language* 32 (6): 781–804.

Frisch, S. A. 1996. “Similarity and Frequency in Phonology.” PhD thesis, Northwestern University.

Frisch, S. A., N. R. Large, and D. B. Pisoni. 2000. “Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords.” *Journal of Memory and Language* 42 (4). Elsevier: 481–96.

Gadalla, H., H. Kilany, H. Arram, A. Yacoub, A. El-Habashi, A. Shalaby, K. Karins, et al. 1998. *LDC Callhome Egyptian Colloquial Arabic Lexicon.* Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Gathercole, S. E, and A. J Martin. 1996. “Interactive Processes in Phonological Memory.” In *Models of Short-Term Memory*, edited by S. E Gathercole, 73–100.

London: Psychology Press.

Giegerich, H. J. 1992. *English Phonology.* Cambridge Textbooks in Linguistics. Cambridge University Press.

Goldinger, S. D, P. A Luce, and D. B Pisoni. 1989. “Priming Lexical Neighbors of Spoken Words: Effects of Competition and Inhibition.” *Journal of Memory and Language* 28 (5). NIH Public Access: 501.

Green, C. R, and S Davis. 2014. “Superadditivity and Limitations on Syllable Complexity in Bambara Words.” *Perspectives on Phonological Theory and Development, in Honor of Daniel A. Dinnsen*, 223–47.

Greenberg, J. H, and J. J. Jenkins. 1964. “Studies in the Psychological Correlates of the Sound System of American English.” *Word* 20 (2). Taylor & Francis: 157–77.

Han, N. 2003. *Korean Telephone Conversations Lexicon.* Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Hay, J., J. Pierrehumbert, and M. E. Beckman. 2004. “Speech Perception, Well-Formedness and the Statistics of the Lexicon.” In *Phonetic Interpretation : Papers in Laboratory Phonology VI*, edited by J. Local, R. Ogden, R. Temple, M. E. Beckman, and J. Kingston, 58–74. Cambridge University Press.

Hayes, B., and C. Wilson. 2008. “A Maximum Entropy Model of Phonotactics and Phonotactic Learning.” *Linguistic Inquiry* 39 (3): 379–440.

Jäger, G., and A. Rosenbach. 2006. “The Winner Takes It All—Almost: Cumulativity in Grammatical Variation.” *Linguistics* 44 (5). Walter de Gruyter: 937–71.

Jurafsky, D., and J. H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* 2nd ed. Prentice Hall. Pearson Education, Inc.

Jusczyk, Peter W, Paul A Luce, and Jan Charles-Luce. 1994. “Infants’ Sensitivity to Phonotactic Patterns in the Native Language.” *Journal of Memory and Language* 33 (5). Academic Press: 630.

Kessler, B., and R. Treiman. 1997. “Syllable Structure and the Distribution of Phonemes in English Syllables.” *Journal of Memory and Language* 37 (3). Elsevier: 295–311.

Kobayashi, M., S. Crist, M. Kaneko, and C. McLemore. 1997. *LDC Japanese Lexicon.* Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Lison, P., and J. Tiedemann. 2016. “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.” In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).* http://www.

opensubtitles.org/.

Luce, P. A, and D. B Pisoni. 1998. “Recognizing Spoken Words: The Neighborhood Activation Model.” *Ear and Hearing* 19 (1). NIH Public Access: 1.

Pierrehumbert, J. 1994. “Syllable Structure and Word Structure: A Study of Triconsonantal Clusters in English.” In *Phonological Structure and Phonetic Form*, edited by P. A. Keating, 168–88. Papers in Laboratory Phonology. Cambridge University Press. https://doi.org/10.1017/CBO9780511659461.011.

———. 2001. “Stochastic Phonology.” *Glot International* 5 (6). Citeseer: 195–207.

Prince, A., and P. Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Publishing.

Shademan, S. 2006. “Is Phonotactic Knowledge Grammatical Knowledge.” In *Proceedings of the 25th West Coast Conference on Formal Linguistics*. Vol. 371379. Citeseer.

Shih, S. S. 2017. “Constraint Conjunction in Weighted Probabilistic Grammar.” *Phonology* 34 (2). Cambridge University Press: 243–68.

Treiman, R., B. Kessler, S. Knewasser, R. Tincoff, and M. Bowman. 2000. “English Speakers’ Sensitivity to Phonotactic Patterns.” In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, edited by M. B. Broe and J. B. Pierrehumbert, 269–82. Cambridge, England: Cambridge University Press.

Vitevitch, M. S. 2002. “The Influence of Phonological Similarity Neighborhoods on Speech Production.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28 (4). American Psychological Association: 735.

Vitevitch, M. S, and P. A Luce. 1998. “When Words Compete: Levels of Processing in Perception of Spoken Words.” *Psychological Science* 9 (4). SAGE Publications Sage CA: Los Angeles, CA: 325–29.

———. 1999. “Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition.” *Journal of Memory and Language* 40 (3). Elsevier: 374–408.

———. 2005. “Increases in Phonotactic Probability Facilitate Spoken Nonword Repetition.” *Journal of Memory and Language* 52 (2). Elsevier: 193–204.

Vitevitch, M. S, P. A Luce, J. Charles-Luce, and D. Kemmerer. 1997. “Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words.” *Language and Speech* 40 (1). SAGE Publications Sage UK: London, England: 47–62.

Weide, R. L. 2008. *The CMU Pronunciation Dictionary*. 0.7a ed. Carnegie Mellon University.

Yang, S., C. Sanker, and U. Cohen Priva. 2018. "The Organization of Lexicons: A Cross-Linguistic Analysis of Monosyllabic Words." In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, 164–73.