A Phonotactic Analysis of the Organization of Lexicons

Shiying Yang

1 Introduction

Different languages have different sets of wordforms to cover roughly the same range of word meanings. What determines which patterns can appear in these lexicons? Phonotactics of a language is concerned with whether a sound sequence can be in the lexicon or not. For example, for two words /blik/ and /bnik/, neither of them are real words in English. However, the former can be recognized as a possible but unattested word for the language, whereas the latter would be deemed as inadmissible by English phonotactics (Chomsky and Halle 1965, 101).

Whether a word can or cannot appear in a lexicon has traditionally been thought of as a binary question. However, the dichotomy between acceptable and unacceptable words cannot account for people's perception of varying degrees of acceptability (e.g. *skell* > *scoil* > *smum*, Albright (2009b)). A probabilistic conception of phonotactics based on segments and subparts of wordforms would be better suited for describing chances of a wordform being in the lexicon.

The use of probabilities of individual parts in accounting for the phonotactic probability of an entire sequence assumes a lexicon's choices of subparts are independent from each other. This is consistent with the conventional assumption of locality in phonological

evaluation that each phonological process is applied independently within its local domain. However, probabilities based on individual subparts are not enough to justify discrepancies between expected frequencies of patterns and their actual attestation in the lexicon. For instance, in English, the sequence /æl/ occurs much less often than would be predicted from product of probabilities of its subparts /æ/ and /l/ (Kessler and Treiman 1997).

Moreover, two recent studies provide evidence that contradicts the assumption that words in lexicons are randomly sampled according to cumulative phonotactic probabilities of their subparts. Dautriche et al. (2017) showed that words in real lexicons are more similar to each other than would be predicted by a phonotactic model that chooses subparts of a word one after another, while Yang, Sanker, and Cohen Priva (2018) found that Mandarin and English lexicons have fewer phonotactically complex words than random baselines generated from a similar phonotactic model. Therefore, contrary to the conventional conception of locality, it is possible that the phonotactic probability of one subpart in a word can affect the probability of another distant part's emergence, and that the overall acceptability of a word is determined not only by the cumulative acceptability of its subparts, but also by how likely subparts of varying acceptabilities combines together.

2 Literature review

2.1 Phonotactics and probabilistic phonotactic models

The study of phonotactics is interested in the phonological acceptability of phonetic sequences. It is typically thought of as a set of language specific (and some potentially

universal) rules or constraints which determines if a word may or may not appear in the lexicon of a language. For example, as mentioned in Section 1, despite their shared un-attestation, the phonotactics of English would differentiate between the acceptable /blik/ and the unacceptable /bnik/. In this case, /bl/ is a prevalent onset in English while the onset /bn/ is both unattested and often cannot be elicited naturally from English speakers (e.g. Wilson and Davidson 2013). On closer inspection, laterals (/l/) are more sonorant than nasals (/n/), and it has been theorized that English requires onsets with a larger rise in sonority (see Clements 1990; Berent et al. 2007).[1]

Phonotactics can be thought of as a accumulation of such principles that are induced based on existing sound patterns in a lexicon. However, models differ in whether acceptability (or well-formedness) is categorical or probabilistic. Under rule-based or constraint-based phonology (Optimality Theory; Prince and Smolensky 2004), whether a sequence is licit or not is a binary question with respect to the adopted grammar.

There are, nevertheless, non-categorical patterns within languages which non-probabilistic grammars cannot easily describe. Well-formedness (also referred to as acceptability and wordlikeness) ratings of non-words consistently show gradience across test paradigms. If speakers' judgments are any indication of their implicit knowledge about phonotactics, this result should imply that there is gradience in phonotactics as well. Along this line of thought, it is not enough for the phonotactics of a language to judge whether a sequence is licit or not. The grammar should also be able to evaluate how typical or representative a word is of the language.

In light of the gradience observed in well-formedness judgments, rule-based or constraint-based phonology would not be adequate in describing the issue other

[1]This generalization is not accurate when borrowed proper names such as Vladimir and zloty are considered (e.g. Duanmu 2002). The fact that these clusters can be naturally elicited from English speakers suggest that /vl/ and /zl/ are also acceptable onsets for English and that the requirement of larger sonority rise is not absolute.

than associating it with *performance* rather than *competence* (Chomsky 1965, 4). The fine-grained gradience calls for phonotactic models to incorporate probabilistic knowledge of the sequencing of word subparts in the lexicon. The term "stochastic phonology" was proposed to put emphasis on the probabilistic nature of sound structures (Pierrehumbert 2001). This line of theories recognize phonotactics and the cognitive representation of sound structure in general as probabilistic rather than definitive. This approach makes it possible to evaluate neologisms as extensions 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 various models with emphasis on different phonological units in phonotactics from syllables and segments to subsegmental features and phonological constraints.

In terms of syllables, 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 untransformed 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 these 3 predictors resulted in similar higher correlation coefficients than probabilities of the worst syllable or the worst segment. These two studies highlighted the importance of considering the en-

tire word in well-formedness judgments rather than a single worst part. This finding 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.

The role of syllable constituent frequencies in phonotactic probabilities is also highlighted in Treiman et al. (2000). This study showed that CVC syllables of the high rime frequency group are rated as more well-formed than those of the low rime frequency group by both adults and elementary school-aged children. In this study, groups are divided based on type frequencies of rimes from dictionaries. The high frequency group was selected from the top half while rimes in the low frequency group all come from the bottom half of frequency counts. In the same study, this result was extended to blending tasks where high-frequency rimes were preserved more than low-frequency rimes. These findings are consistent with predictions of a phonotactic model based on syllable constituents.

Other than phonotactic probabilities over syllable constituents, probabilities over segments or probabilities assigned by n-gram models in general have often been used as baseline phonotactic models in phonological and perceptual studies. An N-gram language model is a common way of assigning probabilities to the last word of an N-gram (an n-word sequence) by computing conditional probabilities of a word given $n-1$ words preceding it (see Jurafsky and Martin 2008, chap. 4). In turn, the probability of a complete sequence is the product of conditional probabilities of all words in it. For the purpose of assigning probabilities to phonological sequences, the unit of consideration is often segments. In this case, N-phone models are adopted, where the probability of each segment is conditioned on the previous $n-1$ sounds (see Section 3.2.2 for an example of the implementation of an N-phone model). Bi-phone probabilities from

bi-phone language models have been shown to be at least 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 at low-probability non-words. Moreover, Vitevitch et al. (1997) adopted the same set of CVC syllables and phonotactic models to demonstrate that besides being rated higher by adults in terms of well-formedness, high-probability di-syllabic non-words 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 in perception and perceived well-formedness 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.

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.

In brief, despite that classical generative phonology and OT treat phonotactics as a set of categorical rules or constraints, there is evidence that probabilistic distributions of levels of phonological representations can be reflected in speakers' gradient intuition about the acceptability of sequences. Therefore, phonotactic models can benefit from incorporating probabilistic knowledge of the sequencing of word subparts in the lexicon.

2.2 Accidental gaps

According to rule-based and constraint-based phonology, any non-existent pattern that is deemed illicit (e.g. /bnik/) can be considered as a structural or systematic gap in the lexicon, whereas unattested patterns that are permitted by the grammar (e.g. /blik/) would be considered accidental gaps. For this line of theories, there is a clear-cut boundary between structural and accidental gaps given the imposed rules or constraints. This distinction between structural and accidental gaps is blurred when taking into consideration the probabilistic nature of phonotactics.

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 word-medial triconsonantal clusters with lower predicted probabilities are less accepted by speakers and indeed do not show up in the lexicon. In this study, clusters were divided into groups based on their expected frequencies, with the highest 20 clusters in the first group, the next 20 in the second group and so on. Pierrehumbert (1994) found that the number of attested clusters in each group decreases steadily with its ranking and that the top 10 groups covers all attested word-medial clusters with few exceptions. Following the idea of considering phonotactics as probabilistic, this method of using expected frequencies to predict the attestation or un-attestation of clusters implies that just like the gradience in well-formedness, the surfacing of sequences in the lexicon is also dependent on their phonotactic probabilities. Thus, certain gaps in the lexicon are not merely accidental, but might be due to low phonotactic probabilities.

However, Gorman (2013) used both expected frequencies and MaxEnt OT to predict the un-attestation of all possible English word-medial clusters and found that neither method is sufficiently accurate in predicting whether a cluster would be unattested or not. Gorman (2013) argues that due to the small size of the lexicon and the skewed distribution of onsets and codas in the lexicon, the prevalence of accidental gaps is inevitable. Frisch (1996) also questioned the result of Pierrehumbert (1994) and pointed out that even though it is normal for individual clusters in low-ranking groups as categorized by Pierrehumbert (1994) to not occur, the number of low-frequency clusters in the lexicon should match the aggregated expected frequencies of clusters in low-ranking groups. Thus, it is questionable that none of them are attested. In other words, expected frequencies alone would not be sufficient in explaining the un-attestation of accidental gaps.

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. This conclusion is consistent with examples discussed in 2.1, where occurrences of patterns in the lexicon (e.g. /ʌf/, /æl/) do not always conform to their expected frequencies calculated based on probabilities of individual segments. In a broader sense, this hypothesis implies that the relationship between the probability of attestation and the phonotactic probability of sequences is not linear. The lexicon might avoid combinations of low-frequency segments or sequences more than expected by probability-based phonotactic models.

Martin (2007) also demonstrated that in English, Navajo and Turkish, compound words have dis-preferences for patterns that are illicit within stems. For example, English allows geminates across morpheme boundaries (e.g. *bookcase*), but noun-noun compounds and words with *-ness*, "*less*" and "*-ly*" suffixes all have less geminates in inter-morphemic clusters than expected possibly due to geminates being illicit within English stems. In other words, compound words also have a tendency to conform to phonotactics pertaining to stems, which further indicates the advantage of preferred structures in the forming of lexicons.

2.3 Superadditivity and constraint conjunction

Along the lines of inconsistencies between expected frequencies and attested frequencies, Albright (2009a) looked into in Lakhota where clusters, fricatives, aspirates and ejectives do not tend to co-occur in the initial and medial onset positions. Analysis of Lakhota mono- and di-syllabic roots revealed that almost any combination of these structures in the initial and medial onset positions occur less often in the lexicon than

expected. For example, according to frequencies of clusters respectively in the initial and medial onset positions, the expected frequency of clusters in both positions (e.g. [gleʃka] 'spotted') is 80 out of 1924 disyllabic words. However, the observed occurrences of such a combination is 54. Compared to 54/80, the ratio between observed and expected frequencies is even lower for combinations of clusters and aspirates (11/47), clusters and ejectives (1/10) and ejectives and aspirates (0/10). Since clusters and fricatives are in general more frequent in Lakhota lexicon than aspirates, and aspirates are in turn more frequent than ejectives, the degree of under-attestation of onset combinations is correlated with frequencies of individual structures[2].

These findings are consistent with the "poor getting poorer" effect proposed by Frisch (1996) as mentioned in Section 2.2. It is unlikely that these are caused by specific constraints since structures such as fricatives and clusters are not usually considered in long-distance dependencies. Albright (2009a) attributed the under-attestation of these co-occurrences to "superadditive" effects of independent markedness violations. More specifically, clusters, fricatives, aspirates and ejectives are relatively dispreferred (marked) by the grammar, and this dispreferrence is correlated with their in-frequencies in the lexicon. These structures together would induce "superadditive" penalties, which lead to the under-attestation of their co-occurrences. Thus, the lexicon would not allow structures that are more complex than a certain complexity threshold. In terms of a MaxEnt model, this effect on a combination of low-frequency structures would be interpreted as two weaker constraints that target individual structures cumulatively outweigh a dominant constraint to lower the probability of its surfacing. In other

[2]It was also found that combinations with nasals do not conform to the same underattestation pattern as other structures. Despite that nasals occur less frequently in the lexicon than aspirates, combinations involving nasals all have frequencies around expected values. Albright (2009a) suggested that nasals are not in themselves dispreferred in Lakhota, they occurred less frequently in the lexicon because there are only 2 nasals in the phonemic inventory, whereas fricatives, aspirates and ejectives all involve at least 4 phonemes.

words, the two individual constraints would gang up to beat a stronger constraint. As a result, input with such combinations would be less likely to surface in the lexicon even though words with a single marked structure can surface as expected.

Compared to a grammar which evaluates input on the basis of "the winner takes it all", the mechanism that leads to superadditive effects in Lakhota described in Albright (2009a) deviates from classical OT mainly in two ways. Firstly, as argued in both Section 2.1 and Section 2.2, the discrepancy between observation and expectation and the gradient nature of under-attestation of different combinations call for a probabilistic rather than categorical description of the lexicon. Secondly, the effects are explainable with ganging-up cumulativity rather than classical OT that does not allow weaker constraints to gang up and beat a stronger constraint. Simply put, all constraints can have an effect on the output, not just dominant constraints (see Jäger and Rosenbach 2006). The idea of weighted constraints and ganging-up cumulativity is hardly new when taking into consideration Harmonic Grammar (e.g. Legendre, Miyata, and Smolensky 1990) and MaxEnt grammar. Ganging-up cumulativity supposedly shows additive effects of constraints rather than "superadditive" effects. Nevertheless, the proposed model in Albright (2009a) does demonstrate "superadditive" penalties of constraints because it is described as a hierarchical model where input to the evaluation process has already been filtered by "baseline constraints", thus any penalty incurred by ganging-up cumulativity is considered additional penalty.

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 MaxEnt model for tone alternation in nouns that encapsulates superadditivity without committing to a hierarchical model by independently assigning weights to constraint conjunctions rather than only to individual

constraints. Constraint conjunctions were modeled as interaction terms in regression models in the form of products of constraints (e.g. $C_1 \times C_2$). Shih (2017) compared models with and without constraint conjunctions and showed that the inclusion of constraint conjunctions improves the explanatory power of the grammar without driving up its complexity. These instances together point to the possibility that superadditivity of penalties on certain structures can be prevalent in languages. Languages have a tendency to avoid multiple phonological complexities in a word.

The inclusion of constraint conjunction in terms of superadditivity differs from the standard conception of local constraint conjunction (e.g. C_1 & C_2) with strict locality restrictions (e.g. Łubowicz 2005). Local constraint conjunction was introduced into OT to account for patterns that are similar to ganging-up cumulativity. For example, final devoicing can be explained by the violation of a conjoined constraint both on codas and voiced obstruents. As a result, only voiceless codas can surface. Crucially, the two constraints involved in this conjunction are evaluated on the same segment due to locality restriction on constraint conjunction (see Crowhurst 2011; Itô and Mester 2003).

In other words, local constraint conjunction emphasizes that the violation of constraints in the same place is worse than separate individual violations. This view is compatible with classical OT and additive cumulativity of constraints in assuming that phonological processes affect the phonological pattern independently. Even when local constraint conjunctions are taken into consideration, they only affect one part of the word within a segment or a syllable. However, patterns that demonstrate the "poor getting poorer" effect and the superadditivity effect show coordination across syllable boundaries: the coda of one syllable and the onset of the next can have an impact on the structure of each other (e.g. Pierrehumbert 1994; Frisch 1996), so do onsets of two adjacent sylla-

bles (e.g. Albright 2009a). Therefore, findings regarding superadditivity implies that phonological processes are not independent and that the lexicon might have a general tendency of avoiding the co-occurrences of complex structures, regardless of where they are in a word.

2.4 Processing and well-formedness

2.4.1 Neighborhood density and phonotactic probability

Even though well-formedness judgments have been used as empirical evidence for phonotactic models, another line of theory accounts for the observed gradience only from the perspective of word processing. This line of research stemmed from the idea that the more similar a non-word is to real words, the more readily speakers can accept it as a well-formed potential addition 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 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 thus implies 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 the *wordlikeness effect* (Gathercole and Martin 1996). Non-words with higher phonotactic probabilities also lead to better recognition memory performance (Frisch, Large, and Pisoni 2000). Since both phonotactic probability (see Section 2.1) and neighborhood density have been found to have

a positive correlation with well-formedness judgments, these results suggest that both higher neighborhood density and higher phonotactic probability should have facilitative effects 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, as predicted, 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.

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. Besides spoken word recognition, evidence in word production is also relevant in the representation of the mental lexicon and phonotactic knowledge. Contrary to effects in word recognition where dense neighborhood generally inhibits perception, it has been found that words with fewer neighbors are more prone to speech errors and give rise to higher reaction times in word repetition task (e.g. Vitevitch et al. 1997; Vitevitch 2002; Vitevitch and Sommers 2003). In this case, phonotactic probabilities also have a facilitative effect on production. Moreover, effects of both factors persist when the other is controlled for (Vitevitch et al. 1997; Vitevitch, Armbrüster, and Chu 2004). Th advantage in production of words with denser neighborhoods and higher probabilities was attributed to higher activation level driven by the presence of a larger number of words that are similar to the target or share the same subparts (Vitevitch 2002).

Taken together, it has been established that phonotactic probabilities and neighborhood densities can both influence language processing. Neighborhood densities have varying effects in word recognition and production, but phonotactic probabilities consistently show a processing advantage both in spoken word recognition and in production.

2.4.2 Lexical and phonotactic accounts of well-formedness judgments

Given that distributions of phonological units reflect how representative any sequence is in the lexicon. 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 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, phonotactic probabilities and neighborhood densities both have positive correlations with well-formedness ratings. This correlation between metrics is hardly surprising, since the two approaches represent 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 distinct from each other, words with many neighbors that are one segment away from it would inherently have more prevalent subparts.

As a result, in practice, 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 sig-

nificant 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 previously discussed theory that both lexical and sublexical mechanisms can be adopted in the presence of different stimuli (Vitevitch and Luce 1999); while the effect of phonotactic probabilities was invariant across 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, the independent significant effect of incorporating feature-based probabilities in Albright (2009b) implies 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 analogy-based lexical influences in online processing.

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 effects 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.

2.5 Challenges to the accepted view of phonotactics

Section 2.1 argued for probabilistic phonotactic models over categorical ones. Different ways of quantifying phonotactic probabilities introduced in Section 2.1, despite their distinct choices of levels of phonological representations, recognize that the evaluation of a word is contingent on the additive phonotactic probabilities of all subparts.

Section 2.3 touched on the assumption of independence between phonological processes. Given this assumption, different parts of words are evaluated independently by phonotactic models. However, the gradient patterns discussed in Frisch (1996), Albright (2009a) and Martin (2007) all point to the possibility that there are long-distance dependencies between different parts of words or different constraints even after cumulativity of phonotactic probabilities are accounted for.

Another line of research that challenges the independence assumption comes from the examination of the organization of lexicons. 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 after controlling for phonotactics. They adopted 5-phone phonotactic models over segments for Dutch, English, French and German lexicons and created baseline lexicons assuming that wordforms would be randomly selected from the pool of candidates licensed by the respective 5-phone models 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 observed regularity 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 recognition. Dautriche et al. (2017) argued that these processing preferences could potentially shape a phonotactically-filtered base to become a lexicon that is more clustered than sparse.

Dautriche et al. (2017) attributed the observed clustering in lexicons to a disposition for similarity due to processing advantages. It is plausible to hypothesize that super-additivity can also be a factor in the lexicon's tendency for more easily processable wordforms. Rather than relying entirely on a similarity account, a more refined system of phonotactics which tends to avoid combinations of phonological complexities can also contribute to the discrepancy between the real lexicon and phonotactically-controlled baselines.

3 A study about the organization of lexicons from the perspective of phonotactics

3.1 Proposal

3.1.1 Issues and hypotheses

The current study set out to explore how the real lexicon differs from what would be expected based on standard phonotactic models. More specifically, the main ques-

tion is whether superadditivity in phonotactics can be observed by comparing the real lexicon to lexicons generated from phonotactic models that only account for additive penalties induced by subparts. If a probabilistic phonotactic model which assumes independence between phonological processes is adequate in accounting for wordforms in the real lexicon, baseline lexicons randomly generated from this model should have distributions of phonotactic probabilities that are more or less the same as the distribution of probabilities based on the real lexicon. This served as the null hypothesis in data analysis. On the other hand, if there are superadditivity effects in a language, lexicons generated based on probabilities from an additive phonotactic model would include expected occurrences of sequences that are supposed to be under-attested in the real lexicon due to superadditive penalties. Thus, the probability distribution of phonotactic probabilities of the real lexicon would be expected to shift to the end of higher probabilities from distributions of baseline lexicons.

Moreover, since superadditivity effects are usually associated with (marked) 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. Therefore, a heavier tail would be expected from distributions based on baseline lexicons.

3.1.2 Approaches to the problem

A cross-linguistic study was conducted to examine whether the proposed pattern of superadditivity in discrete cases can be generalized to a variety of languages from different language families. Similar to the choice made by Dautriche et al. (2017), phonotactics of each language was based on a language model that conditioned on a certain number of segments to ensure the phonotactic acceptability of candidate wordforms. Consis-

tent with both the approach in Dautriche et al. (2017) and the hierarchical model proposed by Albright (2009a), baseline lexicons were generated by randomly sampling from these candidate wordforms according to their assigned phonotactic probabilities. For each language, the target for comparison were probability density distributions of log phonotactic probabilities of words from these generated lexicons and that of the real lexicon.

The null hypothesis in question is whether words in baseline lexicons are sampled according to the same probability distribution as those in the real lexicon, which translates into hypothesis testing for distributions rather than hypothesis testing for parameters in each distribution. As will be shown later in Section 3.3, 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 variables such as 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. More details about distributions of these parameters and the hypothesis testing procedure will be discussed in Section 3.3.

3.2 Methods

3.2.1 Materials

Six languages from different language families were used in the current study. 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.

Table 1: Information on filtered lexicons

	Size of lexicons	% of word types in lexicons	% of word tokens covered in lexicons	Cutoff frequency	Size of enumerated lexicons
English	5030	15.76 %	96.47 %	94	4164
German	5004	5.64 %	84.60 %	279	3921
Spanish	5001	3.16 %	82.05 %	12496	3402
Arabic	6003	33.26 %	88.66 %	3	4668
Japanese	5153	100.00 %	100.00 %	1	4179
Korean	5003	27.06 %	94.38 %	78	4001

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

1997). The Korean Telephone Conversations Lexicon (Han 2003) was used for Korean, but word frequencies were taken from the OpenSubtitle Corpus (Lison and Tiedemann 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). Phonetic representations were used for all languages other than English with phonemic representations[3]. Sketches of phonological inventories of these languages according to respective lexicons are presented in Appendix B.

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

[3]This inconsistency was mainly due to the unavailability of easily accessible lexicons of phonemic representations in most languages. Theoretically the choice between phonetic and phonemic representations should not change relative probability differences of different sounds or patterns in the model. To further eliminate this concern, for Japanese and Korean where phonemic representations were available, both types of representations were processed and analyzed. Results from phonemic representations did not change the interpretation of the reported results based on phonetic representations.

of the lexicon. Since I only used phonological representations of words, homophones were counted once in the lexicon with all frequencies summed up. For each language, wordforms whose token frequencies were greater than or equal to the $5000th$ highest frequency in the spoken corpus were selected to constitute the training dataset. After this filtering, only type frequencies would be relevant for the following study.

Table 1 presents the information of lexicons used in the current study after this filtering process. The column of "Cutoff frequency" lists frequencies of the $5000th$ most frequent word in the corpus of each language after taking out function words and proper names, this number determines the size of core lexicons taken out from each language which varies from 5001 to 6003. Percentages of remaining word types in lexicons after filtering indicate that filtered lexicons used for the study contain only a fraction of word types from original corpora. Despite their relatively small sizes compared to the number of all attested words from each corpus, these filtered lexicons still represent the majority of word tokens occurring in these corpora as indicated by percentages of word tokens in Table 1.

3.2.2 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[4]. 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

[4]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. For example, the probability of the word "movement" (/muvmʌnt/) under the current 4-phone model can be illustrated as:[5]

$$\begin{aligned} Pr(/mu_1vm\Lambda_0nt/) = {} & Pr(m|\_) \times Pr(u_1|\_m) \times Pr(v|\_mu_1) \times Pr(m|mu_1v) \times \\ & Pr(\Lambda_0|u_1vm) \times Pr(n|vm\Lambda_0) \times Pr(t|m\Lambda_0n) \times Pr(\_|\Lambda_0nt) \\ = {} & \frac{\#\_m}{\#\_} \times \frac{\#\_mu_1}{\#\_m} \times \frac{\#\_mu_1v}{\#\_mu_1} \times \frac{\#mu_1vm}{\#mu_1v} \times \\ & \frac{\#u_1vm\Lambda_0}{\#u_1vm} \times \frac{\#vm\Lambda_0n}{\#vm\Lambda_0} \times \frac{\#m\Lambda_0nt}{\#m\Lambda_0n} \times \frac{\#\Lambda_0nt\_}{\#\Lambda_0nt} \end{aligned}$$

Phonotactic probabilities were computed in log format where log probabilities of sub-

[5]Number 0 and 1 mark unstressed and stressed vowels.

parts were combined additively. In later comparisons between phonotactic probability distributions, log probabilities were also used. Besides the technical advantage of avoiding 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).

3.2.3 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.

Thus, real lexicons to be later compared to phonotactically-controlled baseline lexicons only contain words up to 8 segments. The number of word types in these lexicons are listed in the last column of Table 1.

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 3.1) that the lexicon of a language is formed by randomly selecting from all phonotactically-plausible candidates based on their probabilities. As discussed in Section 3.1, 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.

3.3 Results

As stated in Section 3.1.2, quartiles of the distribution of phonotactic probabilities are parameters used for the comparison between the real lexicon and sample baseline lexicons. No assumptions can be made about distributions of each parameter. Their cumulative distribution functions were thus 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).

Taken from distributions of phonotactic probabilities of all words from $10,000$ baseline lexicons, Table 2 illustrates for each quartile values within the 95% interval. Values of each parameter in the real lexicon is compared to sampling distributions of the same parameter to pinpoint where it falls compared to the generated baseline (as shown in the "percentile" column).

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

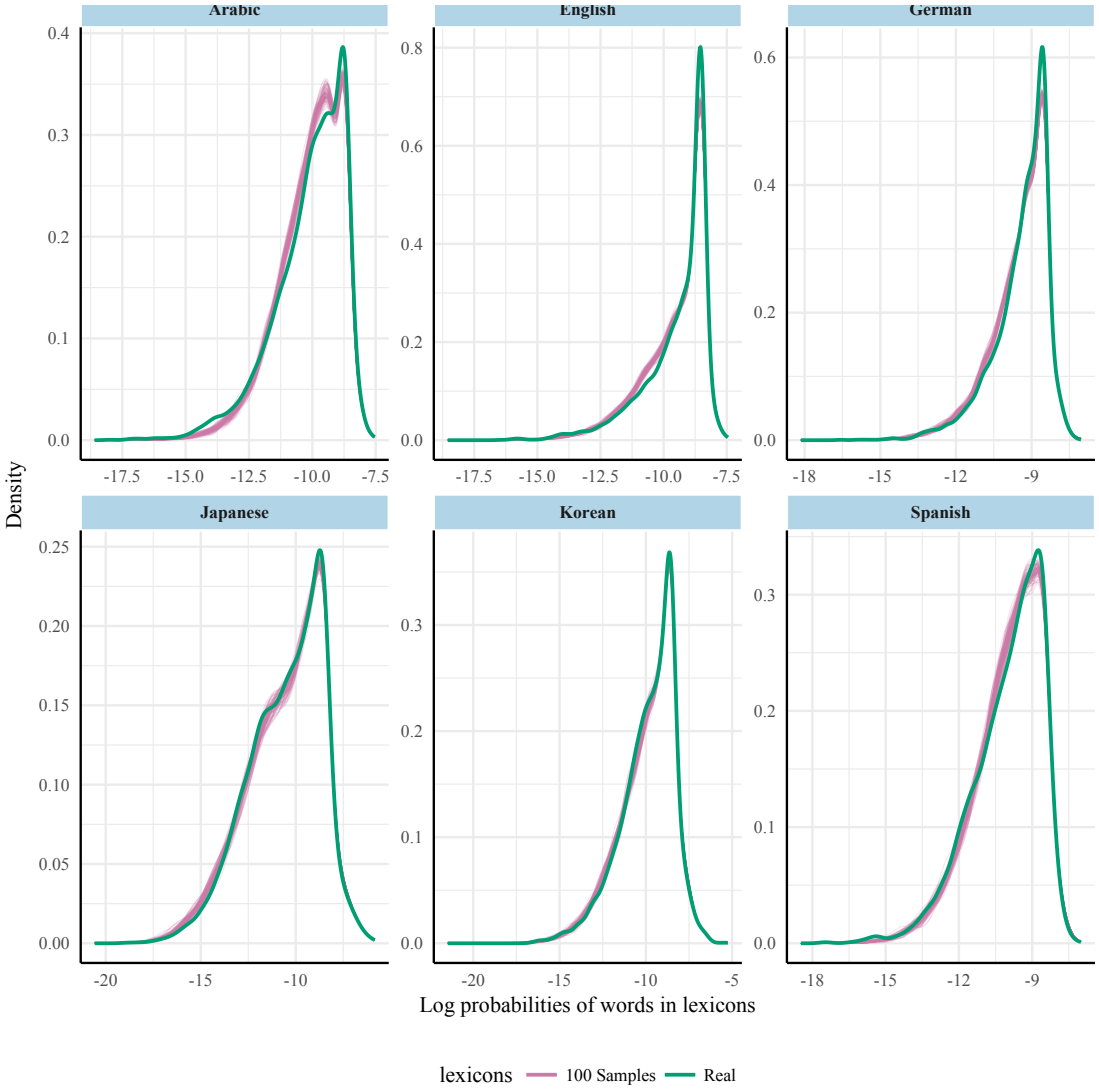

Figure 1: Probability density distributions of log probabilities from the real lexicon and 100 simulated sample baseline lexicons

As far as 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 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 shifted to higher values than their baseline counterparts. Such a trend is captured in Figure 1 where density distributions of 100 sample lexicons and the real lexicon for each language are represented[6]. The x-axis represents lower to higher values of log probabilities, the y-axis marks the density words at each probability. 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. Overall, real English and German lexicons are more shifted to higher probabilities and have thinner tails than distributions of sample lexicons, suggesting less low-probability words than expected.

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%). Similarly, the Spanish Q1 of the real lexicon is significantly lower than expectation for a significance level of 0.05, which, as in the Arabic distribution, corresponds to a heavier 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. For both Arabic and Spanish, 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

[6]Scales for density distributions of each language are adjusted for a clearer presentation of data. The unadjusted version can be found in Figure 2 in Appendix A.

to the right than expected.

In terms of Japanese, the difference between Q1 of the real lexicon and the baseline is close to being significant (at 91.15%). Q2 of the real lexicon is representative of that of the baseline, but Q3 again falls onto the higher end of the estimation (at 99.7%). For Korean, the only quartile of the real lexicon that stands out from the sample lexicons is Q1, which is significantly higher. The other two quartiles for Korean does not provide sufficient evidence to say that the real lexicon has a distribution significantly different from the baseline distribution.

3.4 Discussion

3.4.1 Preliminary conclusion and implications

Analysis of 6 languages shows that more than half of languages tested in the current study have lexicons with clearly more higher-probability words than expected by 4-phone phonotactic model over segments. For Korean and Japanese whose lexicons do not show a consistent predisposition for higher-probability words, the distribution of log probabilities of these lexicons still show some shift to the higher-probability end in at least one parameter.

The clear 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 potential cross-linguistic existence of such a trend does provide evidence for the prevalence of superadditivity effects in languages.

This result can also be linked to processing advantages of words with higher phonotactic probabilities as discussed in Section 2.4.1. Despite the correlation between neighborhood density and phonotactic probabilities, it seems more plausible to attribute the processing advantage to phonotactic acceptability rather than the similarity account. As discussed in Section 2.4.2, phonotactic probabilities and phonological similarity separately contribute to the metrics such as well-formedness judgments. They have distinct effects in spoken word recognition, yet higher phonotactic probabilities consistently have facilitative effects in both word recognition and production. Moreover, from a parsimony point of view, compared to similarity, it is easier to relate phonotactic probability to other phonological constructs such as markedness and ease of articulation.

Additionally, as illustrated in Table 2 and Figure 1, English and German lexicons show similar patterns and results in terms of both the distribution of the real lexicon and distributions of baseline lexicons. Notably results from Dautriche et al. (2017) which showed the clustering of lexicons were based on Dutch, German, English and French. A consistent result would be expected from these languages due to how closely related they are. Similarities between English and German lexicons and dissimilarities between lexicons of other languages found in the current study point to the importance of establishing a theory across language families.

3.4.2 Limitations of the phonotactic model

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 suggests some systematic predisposition that calls for the exploration of more languages. With such restrictions on 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, smoothing would be applied to language models to take away probabilities from frequent patterns and assign them to unattested patterns in training data (see Jurafsky and Martin 2008, chap. 4). This modification was not conducted

in the current study. If some probabilities were assigned to zero-probability patterns from higher-probabilities words, distributions of baseline lexicons would shift further to the left. The fact that distributions of real lexicon probabilities in several languages are already on the right side of those of baseline lexicons without smoothing further demonstrates the validity of current results.

3.4.3 Possible explanations and error analysis

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.

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. 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 little 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.

Despite the general shift of real lexicon distributions towards higher probabilities shown by statistical results, Figure 1 seems to show distributions of real lexicons have higher peaks than baseline lexicons, indicating that words in real lexicons are particularly over-attested around certain higher probabilities rather than being over-attested around a range of probabilities. This is due to the presentation of the data. Density distributions shown in Figure 1 are smoothed over very skewed distributions which leads to

distinction between curves concentrating on peaks. Frequency polygons of these distributions are shown in Figure 3 in Appendix A to demonstrate more intuitively how skewed distributions of phonotactic probabilities are.

Another interesting observation in Figure 1 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 4 in Appendix A). 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.

A look into attested English words with log probabilities below -13.5 reveals that almost all (54 out of 64) attested words with such low phonotactic probabilities are of Latin and French origin. Out of these words, some of them seem to only appear in the 5000 most frequent words due to specific contexts used for counting frequencies (e.g. *"pasta", "solitaire", "satellite", "recipes", "spaniel", "oxygen", "patriot", "matrix"*). Others are frequently used words with combinations or stress patterns that are rarely observed in other parts of the lexicon (e.g. *"honesty", "extent", "capital", "medicine", "orange", "semester", "penalty", "fantasy", "guarantee", "agencies", "anxious", "mechanic", "transit"*). Therefore, it is possible to attribute the over-attestation of very low-probability words to borrowings that did not fully assimilate to the phonotactics of the language. Reasons for the existence of such idiosyncrasies in the lexicon is beyond the scope of the current study. They could also have contributed to the especially low Q1s of Arabic and Spanish shown in Table 2. But this made little impact on the overall trend of these two lexicons leaning towards higher-probability words.

4 General Discussion

Given that words in lexicons are sampled based on how acceptable they are as a word in the language, previous studies showed superadditivity effects of separate phonotactic complexities in several languages: co-occurrences of some distinct complex structures are under-attested in the lexicon compared to their expected frequencies. The current study suggests that superadditivity effects can apply cross-linguistically. In spite of varying degrees, distributions of real lexicons differ from phonotactically generated baselines and demonstrate a general preference for words of higher phonotactic probabilities. Moreover, these effects are directly driven by dispreferences for combinations of non-local structures since local combinations are controlled for in the baseline phonotactic model.

The reason for lexicons to penalize structures with multiple subparts of low phonotactic probability ("poor getting poorer") might be the processing advantage of words that are phonotactically more acceptable. Since speakers are quicker to respond to and produce acceptable words, their facilitative effects in both word recognition and production can indicate that these words are more likely to be used and preserved in the speech community. This possible mechanism is similar to the theory proposed in Martin (2007) from the point of view of linguistic evolution. Martin (2007) argues that the speech community would more easily preserve phonotactically preferred patterns due to their advantage in production. More specifically, several synonyms would compete with each other since the speech community tends to converge to a single word for each concept. In this competition, a word with higher phonotactic probability would be selected more often due to its processing advantage. This preference in use across the community would in turn increase the resting activation level of the word and its subparts which again drives the word's processing advantage. Therefore, with the competition won

by many phonotactically preferred words, the entire lexicon would show a preference for items with higher phonotactic probabilities and a dis-preference for phonotactic complexities.

Together with studies about superadditivity effects in individual languages (Albright 2009a; Green and Davis 2014; Shih 2017) and results about clustering in lexicons from a similarity perspective (Dautriche et al. 2017), current findings challenge the assumption of locality in phonotactics that different parts of a word are evaluated independently from each other. Particularly, evidence presented in the current study shows an overall bias against combinations of low-probability structures, suggesting that non-local superadditivity effects are not isolated incidents, but an inherent predisposition of lexicons.

In future work, more languages will be tested to discover if the "poor getting poorer" effects can be established in more languages. Also, since the explanation for the preference of phonotactically acceptable words is concerned with the evolution of lexicons, investigation into diachronic data of languages can inform the hypothesis of lexicons converging to higher probabilities.

In short, this paper provides insight into the distribution of phonotactic probabilities of the real lexicon. Real lexicons have a dis-preference for lower-probability words beyond expectations of additive phonotactic models, indicating superadditivity effects of non-local phonological complexities.

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): 9–41.

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

Berent, I., D. Steriade, T. Lennertz, and V. Vaknin. 2007. "What We Know About What We Have Never Heard: Evidence from Perceptual Illusions." *Cognition* 104 (3): 591–630.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

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

Clements, G. N. 1990. "The Role of the Sonority Cycle in Core Syllabification." *Papers in Laboratory Phonology* 1: 283–333.

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.

Crowhurst, M. J. 2011. "Constraint Conjunction." In *The Blackwell Companion to Phonology*, 1–30. Wiley Online Library.

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

Duanmu, S. 2002. "Two Theories of Onset Clusters." *Journal of Chinese Phonology* 11: 97–120.

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

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): 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.

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): 501.

Gorman, K. 2013. "Generative Phonotactics." PhD thesis, University of Pennsylvania.

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): 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.

Itô, J., and A. Mester. 2003. "On the Sources of Opacity in Ot: Coda Processes in German." *The Syllable in Optimality Theory*, 271–303.

Jäger, G., and A. Rosenbach. 2006. "The Winner Takes It All—Almost: Cumulativity in Grammatical Variation." *Linguistics* 44 (5): 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, P. W., P. A. Luce, and J. Charles-Luce. 1994. "Infants' Sensitivity to Phonotactic Patterns in the Native Language." *Journal of Memory and Language* 33 (5):

630.

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

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

Legendre, G., Y. Miyata, and P. Smolensky. 1990. *Harmonic Grammar: A Formal Multi-Level Connectionist Theory of Linguistic Well-Formedness: Theoretical Foundations*. Citeseer.

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): 1.

Martin, A. 2007. "The Evolving Lexicon." PhD thesis, University of California, Los Angeles.

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): 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): 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): 735.

Vitevitch, M. S., J. Armbrüster, and S. Chu. 2004. "Sublexical and Lexical Representations in Speech Production: Effects of Phonotactic Probability and Onset Density." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30 (2): 514.

Vitevitch, M. S., and P. A. Luce. 1998. "When Words Compete: Levels of Processing in Perception of Spoken Words." *Psychological Science* 9 (4): 325–29.

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

———. 2005. "Increases in Phonotactic Probability Facilitate Spoken Nonword Repetition." *Journal of Memory and Language* 52 (2): 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): 47–62.

Vitevitch, M. S., and M. S. Sommers. 2003. "The Facilitative Influence of Phonological Similarity and Neighborhood Frequency in Speech Production in Younger and Older Adults." *Memory & Cognition* 31 (4): 491–504.

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

Wilson, C., and L. Davidson. 2013. "Bayesian Analysis of Non-Native Cluster Production." In *Proceedings of the Northeast Linguistics Society 40*, edited by S. Kan, C. Moore-Cantwell, and R. Staubs, 265–78. Amherst, MA: Graduate Linguistic Student Association.

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.

Łubowicz, A. 2005. "Locality of Conjunction." In *Proceedings of the 24th West Coast Conference on Formal Linguistics*, edited by J. Alderete, C. Han, and A. Kochetov, 254–62. Somerville, MA: Cascadilla Proceedings Project.

A Supplementary Figures

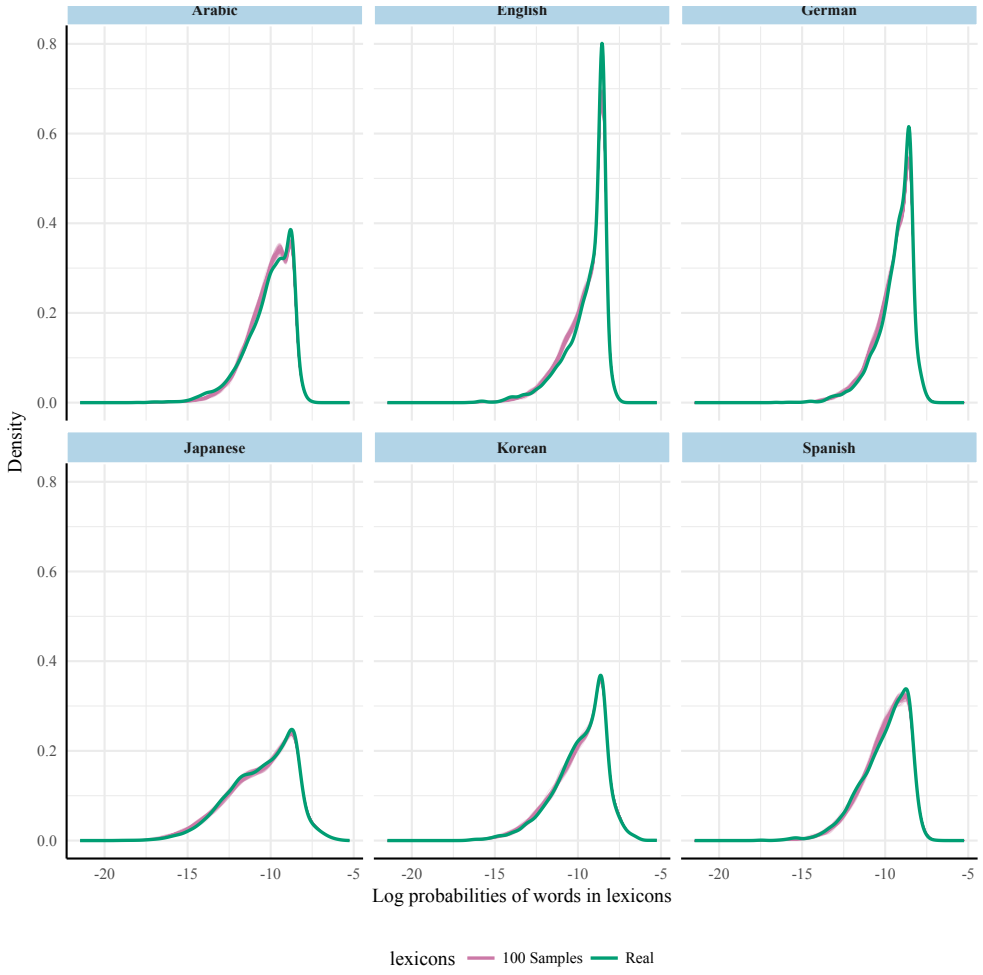

Figure 2: Probability density distributions of log probabilities from the real lexicon and 100 simulated sample baseline lexicons (same scale)

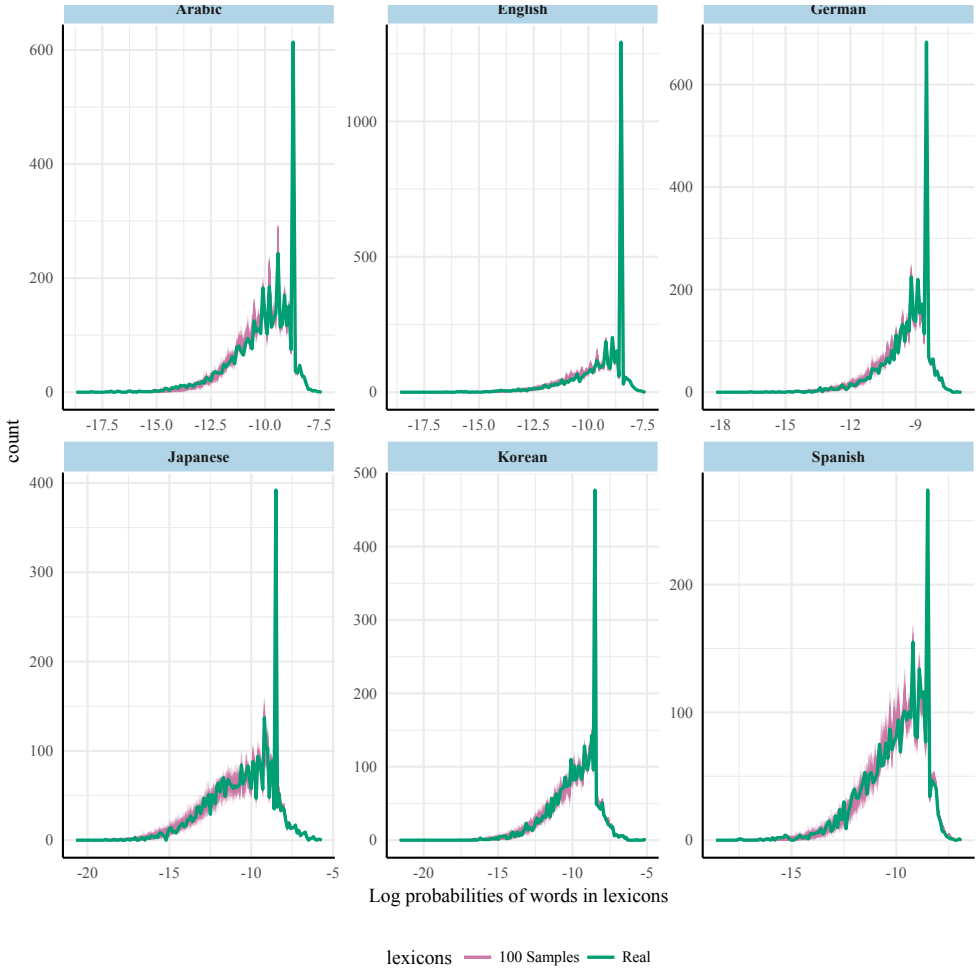

Figure 3: Frequencies of log probabilities from the real lexicon and 100 simulated sample baseline lexicons

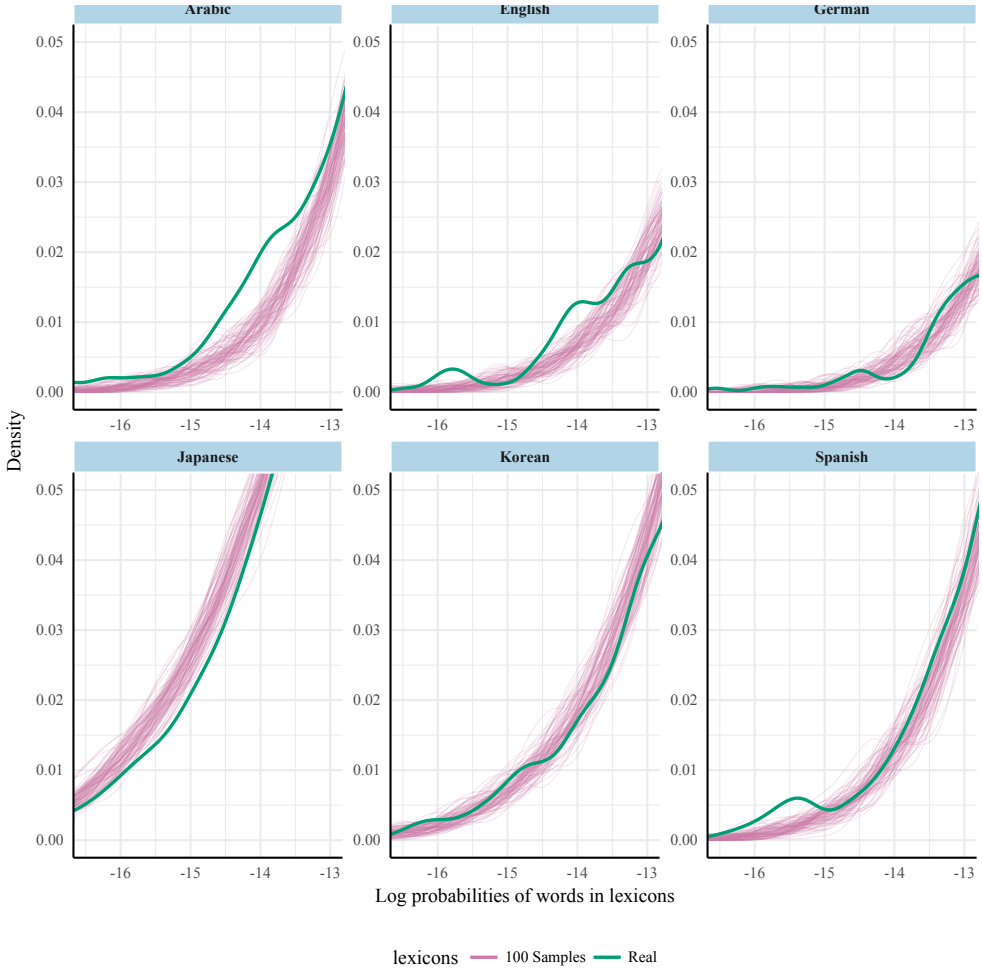

Figure 4: Probability density distributions of log probabilities from the real lexicon and 100 simulated sample baseline lexicons (zoomed)

B Language Sketches

Inventories are extracted based on phonemic or phonetic representations used in each corpus.

1 American English (phonemic)

Consonants

		labial	dental	alveolar	post-alveolar	palatal	velar	glottal
nasal		m		n			ŋ	
stop	voiceless	p		t			k	
	voiced	b		d			g	
fricative	voiceless	f	θ	s	ʃ			h
	voiced	v	ð	z	ʒ			
affricate	voiceless				tʃ			
	voiced				dʒ			
approximant				l	ɹ	j	w	

Vowels

	front	central	back
close	ɪ, i		ʊ, u
mid	ɛ, e	ɝ	ɔ
open	æ	ʌ	ɑ
diphthong		aʊ, aɪ, oʊ, ɔɪ	

Syllable structure:

(CCC)V(CCCC)

e.g. *strengths* /strɛŋkθs/

2 German (phonetic)

Consonants

		labial	alveolar	palatal	velar	glottal
nasal		m	n		ŋ	
stop	voiceless	p	t		k	
	voiced	b	d		g	
fricative	voiceless	f	s	ʃ	x	h
	voiced	v	z	ç, ʒ		
affricate	voiceless	pf	ts	tʃ		
	voiced			dʒ		
trill/tap			ɾ, r			
approximant			l	j	w	

Vowels

	front				central		back	
	unrounded		rounded					
	short	long	short	long	short	long	short	long
close	ɪ, i	iː	ʏ	yː			ʊ	uː
close-mid		eː		øː	ə		o	oː
open-mid	ɛ	ɛː	œ		ɜ		ɔ	
open					a	aː, ã		
diphthong	ɛɪ,	ɑɪ,	ɔø,	ʊɐ,	ɑʊ			

Syllable structure

(CCC)V(CCCCC)

e.g. *k*ä*mpfst* [kɛmpfst]

3 Spanish (phonetic)

Consonants

		labial	dental & alveolar	palatal	velar	glottal
nasal		m	n	ɲ	ŋ	
stop	voiceless	p, pˀ	t		k	
	voiced	b	d		g	
fricative	voiceless	f	θ, s	ʃ	x	h
	voiced	β	ð	ʝ		
affricate	voiceless		ts	tʃ		
	voiced			dʒ		
tap/trill			ɾ, r			
approximant			l	j, ʎ	w	

Vowels

	front	central	back
close	i		u
mid	e, ɛ		ɣ, o, ɔ
open		a	
diphthong		aʊ, aɪ, eʊ, eɪ, oɪ	

Syllable structure

(CC)V(V)(CC)

e.g. *treinta* [tɾeɪnta]

4 Egyptian Arabic (phonetic)

Consonants

		labial	alveolar		palatal	velar	uvular	pharyngeal	glottal
			plain	emphatic					
nasal		m	n						
stop	voiceless		t	tˠ		k	q		ʔ
	voiced	b	d	dˠ		g			
fricative	voiceless	f	s	sˠ	ʃ	x		ħ	h
	voiced	v	z	zˠ		ɣ		ʕ	
tap/trill			ɾ						
approximant			l		j	w			

Vowels

	front		back	
	short	long	short	long
close	i	iː	u	uː
mid		eː		oː
near-open	æ	æː		
open			ɑ	ɑː

Syllable structure

(CC)V(V)(CC)

e.g. [ɕæhɾ]

5 Japanese (phonetic)

Consonants

		bilabial	alveolar	alveolo-palatal	palatal	velar	uvular	glottal
nasal		m	n				(N)	
stop	voiceless	p	t			k		
	voiced	b	d			g		
fricative		ɸ	s, z	ɕ				h
affricate			ts	tɕ, dʑ				
tap/trill			ɾ					
approximant					j	w		

[a] [N] represents a placeless nasal coda

Vowels

	front		central		back	
	short	long	short	long	short	long
close	i	iː			ɯ	ɯː
mid	e	eː			o	oː
open			a	aː		

Syllable structure

(Cj)V(V)(CC)[1]

e.g. [kjaN]

[1] coda must be a nasal or a geminate

6 Korean (phonetic)

Consonants

		bilabial	alveolar	palatal	velar	glottal
nasal		m	n		ŋ	
stop	plain	p	t		k	
	aspirated	pʰ	tʰ		kʰ	
	tense	p’	t’		k’	
fricative and affricate	plain		s	tɕ		h
	aspirated			tɕʰ		
	tense		s’	tɕ’		
liquid			l ~ ɾ			
approximant				j	w	

Vowels

	front		central	back	
	unrounded	rounded		unrounded	rounded
close	i	y		ɯ	u
close-mid	e	ø			o
mid			ə		
near-open	æ				
open			a		

Syllable structure

(CG)V(C)

e.g. [hjuŋnæ]