The organization of lexicons: A cross-linguistic analysis of monosyllabic words

Shiying Yang, Chelsea Sanker, and **Uriel Cohen Priva**
shiying_yang,chelsea_sanker,uriel_cohen_priva@brown.edu

Abstract

Lexicons utilize a fraction of licit structures. Different theories predict either that lexicons prioritize contrastiveness or structural economy. Study 1 finds that the monosyllabic lexicon of Mandarin is no more distinctive than a randomly sampled baseline using the phonological inventory. Study 2 finds that the lexicons of Mandarin and American English have fewer phonotactically complex words than the random baseline: Words tend not to have multiple low-probability components. This suggests that phonological constraints can have superadditive penalties for combined violations, consistent with e.g. Albright (ms.).

1 Introduction

Lexicons can be considered mappings between word meanings and phonotactically-valid sequences of phonemes. There are several dimensions of forces shaping lexicons, based on the frequency of each item and its phonetic distinctiveness from similar items, as well as the phonotactic probability of the phonological sequences within each item. For instance, underlying pressures on the lexicon influence the frequency distribution of items within a lexicon (Zipf, 1929; Piantadosi et al., 2009); Zipf's law predicts that frequent words should be preferentially mapped to shorter segmental sequences.

In the absence of other pressures, syllables and words should be maximally distinct from one another, in order to minimize ambiguity and potential for confusion. This pressure has been demonstrated within phonological inventories; vowel systems tend to maximize the distance between vowels (Flemming, 2004), though other work has found a tendency for economy, in which each feature tends to be used for multiple contrasts, particularly among consonants (Clements, 2003; Dautriche et al., 2017). Wedel et al. (2013) demonstrate that contrastiveness is important in shaping lexicons; phonological mergers are less likely when more lexical contrasts depend on the phonological contrast. The pressure for contrastiveness has been demonstrated in various experiments, in which words with higher neighborhood density are identified more slowly than words with lower neighborhood density (Luce and Pisoni, 1998). If lexicons are not maximizing how distinct lexical items are, there must be other pressures outweighing contrastiveness.

Using two computational studies, we examine some of the factors influencing the shapes of items within lexicons, by comparing actual lexicons to generated lexicons given the same phonotactic restrictions.

1.1 Competing pressures in a lexicon

Zipf (1929) proposed the principle of least effort as a primary force shaping phonological inventories, claiming that the frequencies of sounds within a language are negatively correlated with their articulatory and perceptual complexity, given a set number of contrasts. Thus, the probability of a sound would reflect its overall per-

ceptual and articulatory cost. Consistent with this proposal is the strong correlation between the cross-linguistic frequency of phonemes (i.e. what percentage of languages in UPSID have them) and their frequency within particular languages (Sanker, 2016).

In line with this functional view, emphasizing the communicative goal of language, Flemming's (2004) Dispersion Theory of contrast translated the trade-off between speaker and listener into three conflicting goals: "maximizing the distinctiveness of contrasts," "minimizing articulatory effort" and "maximizing the number of contrasts." He proposed that a phonological inventory would strike a balance between these goals, providing the most distinctive vowel system possible with a given number of contrasts, with articulatory effort only as motivated by achieving distinctiveness. This principle should also extend to lexicons: All else being equal, lexicons should be maximally distinct. This is additionally supported by perceptual evidence that dense lexical neighborhoods slow down processing (Luce and Pisoni, 1998).

However, lexicons seem to be less dispersed than would be expected from the pressure of maximizing contrastiveness. Dautriche et al. (2017) looked at the lexicons of four Indo-European languages and found that they were more regular ("clumpy") than expected by chance. Words were more similar to each other in these languages than in generated phonotactically-controlled baseline lexicons. This result parallels some work in the segmental domain, which shows that languages tend to reuse phonological features (Clements, 2003). However, one potential limitation of this study is that the phonotactic restrictions were tightly controlled, with environments extending out to four segments, which could have constrained the generated lexicons beyond just capturing the intended phonological constraints; in long words, it can be unclear what segmental range best captures the inherent phonotactic patterns.

In order to expand the data into an unrelated language and in particular address whether the lexicon would pattern differently in a language with shorter words and a denser lexicon, we compared generated phonotactically-controlled lexicons to the real lexicon of Mandarin Chinese in Study 1.

1.2 An explanation of phonologically clustered lexicons

In contrast to the dispersion account which bases the drive for distinctiveness on communicative efficiency, Dautriche et al. (2017) attributed their findings to a pressure for regularity that is driven by the goal of lowering cognitive costs in language acquisition and lexical access. A different possibility is that our understanding of the forces driving a language's phonotactics are flawed.

Within phonological theories that address gradient phenomena, models are generally multiplicative. For instance, in MaxEnt, as presented by Hayes and Wilson (2008), the probability assigned to a phonotactic form is e raised to the negative sum of the weighted constraint violations. Calculated differently, this is the product of the probability of each individual violation occurring.

Thus, MaxEnt treats constraints as being independent (Hayes and Wilson, 2008). However, multiple languages have constraint combinations which are more limited in combination than would be predicted from their independent probabilities (Albright, ms; Green and Davis, 2014; Shih, 2016). For example, English /æ/ and coda /z/ are attested with somewhat low frequency, but their combination is extremely uncommon, far below the product of their independent probabilities (Kessler and Treiman, 1997). Such patterns have been explained as "superadditivity"(Albright, ms) or "supercumulativity"(Shih, 2016), a phenomenon in which combinations of marked structures incur additional penalties, though their co-occurrence is not categorically disallowed.

The superadditivity effect might underlie some of the patterns of lexicons, as it would produce a faster drop-off in the occurrence of low probability forms, resulting in more clustering around higher probability forms than is predicted by models in which all phonotactic

constraints are independent. Study 2 was devised to test the null hypothesis of a multiplicative grammar, in which the probability of a certain form appearing as a word is the product of the probabilities of each of its components, against a counter-hypothesis of a grammar including additional penalties for combinations of low-probability sequences; see section 3.2.

1.3 The null hypotheses: A lexicon selected by chance

Similar to the resampling procedures used by Dautriche et al. (2017), sample lexicons were generated to estimate statistics of a baseline population distribution as predicted from the phonotactic constraints and lexicon size of Mandarin, to be tested against measurements of the real lexicon. A lexicon can be thought of as a set of word forms drawn from a pool of all forms that are licit within the phonotactic constraints of a language. To draw a lexicon with k contrasting items from a constrained pool of n licit shapes, there are $\binom{n}{k}$ possibilities for lexicons; generated lexicons are drawn from this pool of possibilities.

If the lexicon is not under any pressure to maximize either distinctiveness or regularity, the sampling procedure from the pool of candidate word-forms will be random; Study 1 tests the predictions made by random sampling.

If independent phonotactic constraints are sufficient to capture well-formedness and thus predict frequency distributions in lexicons, probabilities of phonemic shapes will follow from probabilities of their subparts (Albright, ms). Study 2 tests the predictions made by independent evaluation of constraints; if constraints are independent, generated lexicons that are randomly sampled from the pool of forms based on probabilities produced by the phonotactic constraints of a language without any constraint interaction should have distributions similar to the real lexicon.

Both Study 1 and Study 2 are based on constructing phonotactically-constrained pools of words from which generated lexicons are sampled. The word-pools and the artificial lexicons are generated according to the parameters laid out in the following sections, to create baselines for evaluating what factors are influencing the real lexicons. We aim to show that real lexicons cannot be explained by randomly sampling from a constrained phonological space and that the constraints on the phonological space call for a model that includes superadditivity.

2 Study 1: Evaluating the distinctiveness of Mandarin monosyllabic lexicon

2.1 Background

Study 1 investigated monosyllabic words in Mandarin Chinese. Mandarin has a dense phonological space and limited licit syllable structures, which make it possible to enumerate all phonologically permissible forms with relatively few assumptions.

Mandarin syllables are limited to a structure with at most four phonemes: CGVX (Li and Thompson, 1987). C stands for a consonant in the onset position; G stands for a glide; V stands for a vowel; and X can either be a nasal /n/ or /ŋ/, or the off-glide of a diphthong. Every syllable must have a vowel, but all other positions can be empty (Duanmu, 2009). In addition, each syllable has one of four phonological tones. Given only these structural constraints, the phonological inventory would allow 7,600 possible syllables (Duanmu, 2009). Most words in Mandarin are monosyllabic or disyllabic, so limitations in licit syllables result in a rather small number of possible words.

If there is a pressure towards contrastiveness within the lexicon, it should be particularly apparent in a language with such a small number of phonotactically licit forms. Thus, our prediction was that the real Mandarin lexicon would be more dispersed than the randomly sampled generated lexicons.

2.2 Methods

For Study 1, we used the LDC Mandarin Lexicon and the corresponding frequency data from the LDC Mandarin Callhome training transcripts (Huang et al., 1997). Words which include the 5th tone ('neutral tone') or lack a nu-

clear vowel were excluded from analysis, to avoid clitics (Chao, 1968), which would be outside the scope of this analysis.

The crucial aspect of lexical contrast is phonological form, so we based perceptual distinctiveness on phonemic representations rather than phonetic measurements, using features to calculate distance between consonants and distance between formants to calculate distance between vowels. Based on misperception studies, perceptual distance between the presence and absence of a segment is highly sensitive to the segment and its environment (Tang, 2015; Sanker, 2016); such differences do not clearly fit into the same system as contrasts between phonemes, so words with different syllable structures were considered separately. Focusing just on the monosyllabic lexicon of Mandarin, we looked at CV (open syllable) and CVX (closed syllable) structures.

2.2.1 Defining the licit structures

In order to sample generated lexicons of Mandarin from the hypothesized phonological space described in 1.3, a list of well-formed syllables was generated for CV and CVX forms, to represent candidate word-forms. In order to generate such lists, all combinations of CV and CVX structures were laid out, based on the phonological segment inventory of Mandarin; then the licit word-forms of the two structures were filtered through phonotactic models, using n-grams for phonological sequences (Jurafsky and Martin, 2008).

For CV words, well-formedness was determined using a phonological bi-gram (bi-phone) model, in which the probability of a word was defined as the product of the individual probabilities for all segments given the phoneme immediately preceding each; the probability of the tone was conditioned on the vowel. CVX words were evaluated similarly, but with a tri-phone model instead of a bi-phone model due to the extra degree of freedom induced by the coda. The probability of a word was defined as the product of the individual probabilities for all segments given the two phonemes preceding each, and tone was still conditioned on the vowel. Because only monosyllabic words were considered, there is no possibility of long-distance dependencies. Segment probabilities were based on all attested syllables in the LDC lexicon. Under this model, words with probabilities higher than 0 were considered well-formed. Beyond that, the probabilities produced by this model were not used for Study 1.

The resulting lists of forms contain 304 CV syllables (out of 360 structurally possible combinations) and 544 CVX syllables (out of 1440 structurally possible combinations), which represent the number of phonotactically licit syllables of these shapes. Of these, there are 187 monosyllabic words with CV structure attested in the LDC Mandarin Lexicon and 327 words with CVX structure. The two filtered lists of words serve as phonologically licit pools of words for the sampling procedure described in 2.2.3.

2.2.2 Defining the distinctiveness of lexicons

In evaluating dispersion within lexicons, the distinctiveness between any two segments σ_k and σ_v is denoted as $d(\sigma_k, \sigma_v)$. Comparisons were conducted with corresponding segments from the syllables being compared, e.g. comparing onsets to onsets.

In order to reflect the perceptual differences between segments, the metrics for distinctiveness differed for consonants and for vowels. For consonants, the distinctiveness between each pair of sounds was determined by the number of featural differences, which has been shown to correlate with perceptual measures of distinctiveness (Bailey and Hahn, 2005; White and Morgan, 2008). For example, $d(/p^{\mathrm{h}}/, /t^{\mathrm{h}}/) = 1$, because the two phonemes differ only in place of articulation; $d(/f/, /\widehat{\mathrm{ts}}^{\mathrm{h}}/) = 4$, because the two phonemes differ in place, continuance, delayed release and aspiration.

For vowels, the distinctiveness was based on the Manhattan distance between each vowel pair in the three-dimensional vowel space defined in Flemming (2004), where F1, F2 and F3 values are mapped onto a set of integers in each dimension, given the number of cross-linguistically possible contrasts making use of

each dimension.[1] This choice of metric, rather than a feature-based metric, was due to perception studies suggesting that acoustic differences provide a better model for vowel perception than a feature model does (Ettlinger and Johnson, 2009). In order to have equal weighting of contrasts between consonants and contrasts between vowels, measurements of vowel distinctiveness were scaled down by 1/3.[2]

Each tone was treated like a distinct segment, but with a binary measure of distinctiveness: 1 (different) or 0 (the same). This decision was based on the paucity of available data on tone misperception patterns among native speakers of Mandarin and based on the variation in what distinctiveness patterns are suggested by results from different tasks (Huang and Johnson, 2010).

The distinctiveness of a word from each other word was measured with the log-transformed sum of each segment's distinctiveness from the corresponding segment in the other word. For example, for a word of CVX structure S_r, a segment σ_k in position Φ_n of the syllable is denoted by σ_{r,Φ_n}. The distinctiveness between 2 syllables S_r and S_t, as denoted by $d(S_r, S_t)$, is the log-transformed sum of $d(\sigma_{r,\Phi_{pos}}, \sigma_{t,\Phi_{pos}})$ for all positions of the syllable POS (2.1). In addition, a 1 was added to the sum before log-transformation so that minimal pairs would have a distinctiveness score larger than 0.

(2.1) The distinctiveness between word S_r and S_t

$$d(S_r, S_t) = \log[\sum_{pos \in POS} d(\sigma_{r,\Phi_{pos}}, \sigma_{t,\Phi_{pos}}) + 1]$$

An average distinctiveness of all pairs of words in the given phonological system M' was calculated for each generated lexicon (2.2). The higher this number is, the more distinctions in the possible phonological space the lexicon has used.

[1]For example, /u/ is represented in the vowel space as [F1: 1, F2: 1, F3: 1] and /i/ is represented as [F1: 1, F2: 6, F3: 3], so /u/ and /i/ differ by 7 units in total.

[2]This scale was based on aligning the featurally-defined distinctiveness of the three Mandarin glides (/w/, /j/, and /ɥ/) with the formant-based distinctiveness of the three corresponding vowels (/u/, /i/ and /y/).

(2.2) The average distinctiveness (by word pairs) of a size N lexicon from a given system M'

$$D_{M'} = \frac{\sum_{w \in W_{M'}} \sum_{w \in W_{M'}} d(S_w, S_w)/2}{P(N,2)/2}$$

2.2.3 Generating baseline inventories

Baseline inventories for each syllable structure were generated in three steps. First, the summed frequency of CV (or CVX) words in the real monosyllabic lexicon was used to generate a random set of words within the phonologically licit space defined in 2.2.1. The generated lexicons were then optimized to minimize differences from the real Mandarin lexicon in lexicon size, word frequency distribution, and individual segment frequencies. By minimizing the differences in these parameters, we ensured that the generated baseline lexicons would be comparable to the Mandarin lexicon. Finally, the generated lexicons were filtered to further ensure a close match with these parameters, limiting the generated lexicons to those with a size within 5% of the original lexicon size and a correlation of at least 0.95 between their segment frequencies and the segment frequencies of Mandarin, and between their word frequency distribution and that of Mandarin.

These parameters served to hold articulatory effort constant, with variation only in distinctiveness, based on the assumption that the overall effort of a language is the mean of the effort needed for all words of the language and the effort associated with each word is the sum of the effort associated with all of its segments.

2.3 Results

Consistent with the central limit theorem and the independent sampling process, the distribution of distinctiveness scores of the generated lexicons of both CV and CVX structures conform to normality, as confirmed by the Kolmogorov-Smirnov test (CV: $p = 0.963$, CVX: $p = 0.881$).

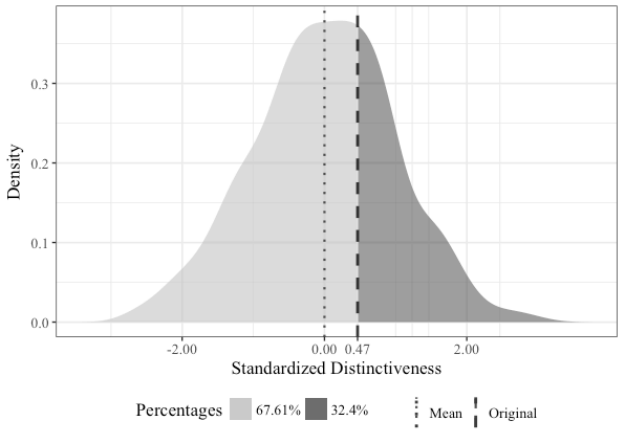

Figure 1: Standardized *D.*(distinctiveness) of 389 generated CV monosyllabic lexicons.

The shaded area in Figure 1 demonstrates the distribution of standardized distinctiveness scores of generated lexicons of monosyllabic CV words. Standardized distinctiveness of the real CV monosyllabic lexicon of Mandarin (indicated by the heavy dashed line) was greater than roughly 67.6% of generated counterparts (indicated by the light grey area). While the real lexicon is above the mean, this result is not conclusive evidence that the real lexicon differs from lexicons drawn randomly from the phonological space, given that the real lexicon is not an outlier or at all close to the top or bottom 2.5% of the distribution.

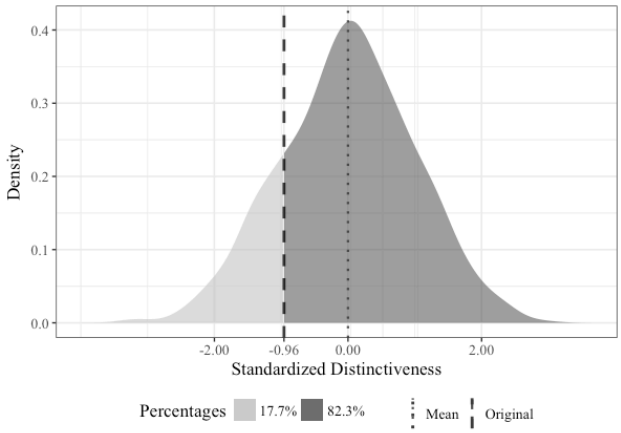

Figure 2: Standardized *D.*(distinctiveness) of 627 generated CVX monosyllabic lexicons.

As indicated by Figure 2, the distinctiveness of the real monosyllabic Mandarin CVX lexicon is better than only 17.7% of baselines. However, the real lexicon is not enough far enough towards the edge of the distribution to demonstrate that it differs from lexicons drawn randomly from the phonological space, because while it is lower than the mean, it is not an outlier.

2.4 Discussion

Mandarin words of both CV structure and CVX structure displayed similarly inconclusive patterns. Compared to randomly generated lexicons following the same parameters of phonology, word frequency, and size, the real lexicon was not an outlier in distinctiveness either in CV or CVX syllables, though the real CV lexicon was slightly better than average among the generated lexicons and the real CVX lexicon was worse. Because the CV phonological space is smaller and more saturated, it is not surprising that CV portion of Mandarin monosyllabic lexicon would be relatively more efficient than its CVX counterpart. However, in general, the results did not support the hypothesis that dispersion plays a large role in shaping lexicons, and are more consistent with the opposite pattern of clumping, as seen in Dautriche et al.'s (2017) results.

The inconclusive results might in part be due to issues with the metrics used for distinctiveness, as perceptual data suggests that different positions in a syllable are not equally salient. At least within English, listeners are most sensitive to mispronunciations in onsets, less so in codas, and least sensitive in nuclei (Franklin and Morgan, 2017), and are more accurate in perceiving onsets than codas, though this can vary depending on listeners' native language, even for the same stimuli (Sanker, 2016). Given such findings, different syllable positions might best be given different weights in distinctiveness when generating sample lexicons. Further research into Mandarin speakers' patterns of misperceptions at the word level and the segment level would further help in accurately quantifying distinctiveness.

3 Study 2: Evaluating the well-formedness of generated Mandarin and English monosyllabic lexicons

3.1 Background

Study 1 shows that a lexicon might not be as dispersed as the functional goal of communicative clarity would predict. In Study 2, we examine whether this lack of dispersion can be partially explained by gradient well-formedness constraints shaping the lexicon, disproportionately favoring words with high-probability sequences.

3.2 Methods

In Study 2, English and Mandarin were used as languages for a preliminary cross-linguistic investigation. The same LDC Mandarin Lexicon from Study 1 was used for Mandarin and the CMU Dictionary (Weide, 2008) was used for the phonemic representations for American English. CMU Dictionary entries were spell-checked with GNU Aspell to exclude rare names and borrowings from other languages. Function words and words with the rarest 1% of onsets and codas were also excluded, due to the uniqueness of their phonological structure, as many function words are clitics and can be reduced more than other words, and words with highly unusual sequences are likely to have unique etymologies that do not reflect the overall pressures of the language.

Only monosyllabic words were used. As discussed in 2.2.1, this limitation meant there were no long-distance dependencies that needed to be accounted for. Phonotactics were represented by a tri-phone model of sound sequences (as introduced in 2.2.1), with the predictability of each sound based on the two preceding phonemes. Study 2 focused on the word probabilities assigned to forms within the generated lexicons. Frequencies, as captured by n-gram models in this study, were used to approximate distributional markedness (Albright, ms). The distinction between frequency and markedness is beyond the scope of this paper.

Sampling followed a sampling procedure similar to that of Study 1. First, all combinations of segments in all possible syllable positions were laid out, producing lists of potential words. Then the real monosyllabic lexicons of Mandarin and English were used to train the tri-phone phonotactic models for each language, assigning log probabilities to all forms in the word lists based on the sum of log probabilities of each word's components. The wordlists were then filtered, only retaining forms with probabilities larger than 0, meaning that they were well-formed within the tri-phone model. Finally, in order to generate the artificial lexicons for English, words were randomly taken from the filtered English list according to their weights in the tri-phone model, with the number of words of different lengths kept consistent with the real English lexicon. The same was done for Mandarin. Thus, the generated baseline lexicons had the same distribution of word lengths and the same size as the real lexicons.

Distributions of log probabilities of the baseline lexicons were compared to the real lexicons, to test whether the probability distributions of real lexicons differ from randomly generated lexicons based on phonotactic models which assume independence of subparts more than one segment apart. Logarithmic scales for probability, with probabilities of subparts combined multiplicatively, have been found previously to have a strong positive correlation with gradient well-formedness ratings and decisions about acceptability of nonce words (Frisch et al., 2000; Coleman and Pierrehumbert, 1997), though these studies did not look for patterns in where the data deviated from the model.

3.3 Results

Both in English and Mandarin, the real lexicons exhibited over-representation of high-probability forms and under-representation of low-probability forms.

The independent sampling process was essentially producing replications which could be used to bootstrap variance estimation for the estimators of interest, so standard errors and confidence intervals of variances were constructed with the bootstrap distributions calculated us-

ing the generated sample lexicons.

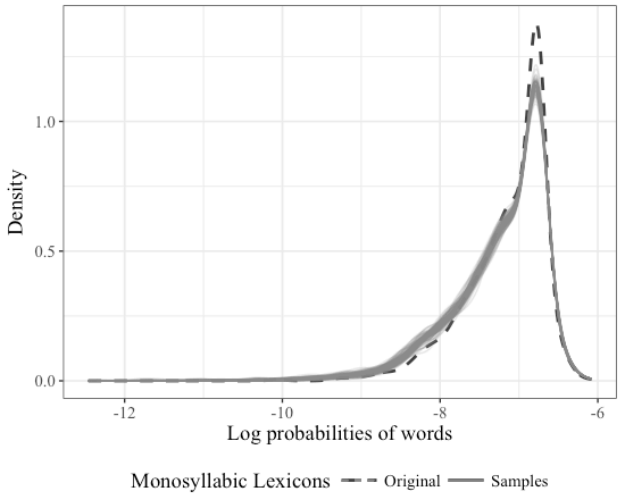

Figure 3: Probability density distributions of the original and 100 generated Mandarin monosyllabic lexicons

Figure 3 illustrates the probability density distributions of the original and generated sample Mandarin monosyllabic lexicons over word probabilities as defined in the phonotactic model. The figure demonstrates that the real monosyllabic Mandarin lexicon (indicated by the dashed line) is more clustered around the higher-probability types than the sample lexicons (indicated by the light solid lines).

Two statistics were used to test whether the distribution of log probabilities in the real lexicon is likely to come from the population distribution. Sample means were tested against the mean of the real lexicon in a two-tailed test ($t = -69.632$, $p \approx 0.000$). This shows that the mean of word probabilities in the real Mandarin monosyllabic lexicon was significantly higher than the true mean of the population distribution generated from the null hypothesis, as measured from the sample lexicons.

In addition to the mean, the bootstrap percentile confidence intervals (CI) of the variance of the hypothesized Mandarin population distribution was approximated. The variance of the probability distribution of the real lexicon ($\sigma^2_{real} = 0.256$) is significantly lower than the variance of the generated population, which has a 95% CI between 0.31 and 0.42.

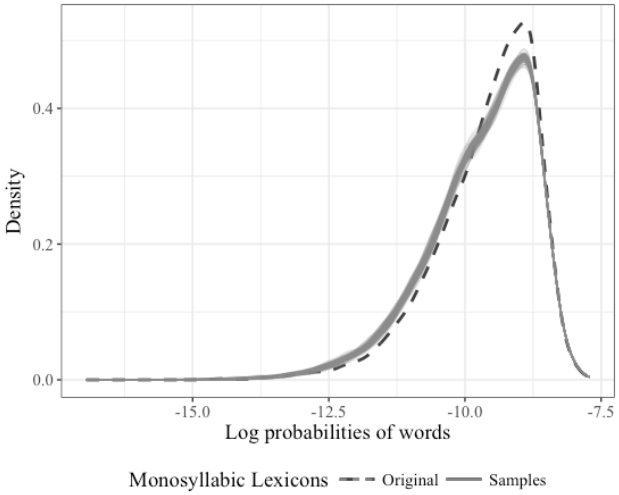

Figure 4: Probability density distributions of the original and 100 generated English monosyllabic lexicons

The English data (illustrated in Figure 4) were analyzed with the same statistical methods as the Mandarin data. As in the Mandarin results, the mean of the probability distribution of the real English monosyllabic lexicon is significantly higher than that of the population distribution generated from the null hypothesis ($t = -155.222$, $p \approx 0.000$).

The comparison in variance between the real English lexicon and the hypothesized population distribution also exhibits a pattern similar to the Mandarin data. The variance of the real lexicon ($\sigma^2_{real} = 0.86$) is significantly lower than that of the hypothesized population which has a 95% CI between 0.93 and 1.02.

3.4 Discussion

If the lexicon is shaped only by local phonological constraints, as controlled for in the phonotactic models used in the current study, sample lexicons generated by the models should follow roughly the same distribution as the real Mandarin and English monosyllabic lexicons. The results of this study provide strong evidence against this null hypothesis.

Within both English and Mandarin, the real monosyllabic lexicons have higher means and smaller variances than the generated baselines, which indicates that the real lexicons make use of more high-probability word types than would

be expected by the phonotactic models, which is also represented by the higher peaks of dotted lines in both Figure 3 and 4. Additionally, as can be seen in both figures, the real Mandarin and English monosyllabic lexicons have thinner left tails than the generated lexicons, indicating a faster drop-off in the frequency of lower probability words. These results seem to suggest that there is a superadditivity effect that penalizes words with multiple low-probability subparts, potentially in combination with a tendency to re-use high-probability sequences, as suggested by Dautriche et al. (2017).

4 General Discussion

Study 1 found that in Mandarin Chinese, dispersion is not a prominent force in the shaping of the lexicon; evidence for a pressure towards clustering was somewhat more suggestive, though not decisive. Study 2 showed that the lexicons of Mandarin and English have more words of higher probability and fewer words of lower probability than would be expected by a phonological model in which constraints are independent. This result reinforces the results in Study 1, indicating a lack of dispersion and instead a trend towards clumping within high probability forms.

These results can fit into Albright's (ms.) proposed grammar of weighted constraints, in which he suggests that inputs with multiple markedness violations have a superadditive effect that can overcome a threshold of well-formedness, resulting in forms which are unattested despite not being directly prohibited. The next step of Study 2 is to expand it to more languages, to test how consistent the effect of superadditivity is cross-linguistically. Future work should also investigate whether the observed patterns in lexicons are driven by the interaction of particular constraints, or if they result from a general pattern in how all constraints combine.

The superadditivity account and the presented evidence are consistent with Dautriche et al.'s (2017) findings that lexicons are more regular than expected. However, a pressure for "clumpiness" and a superadditivity effect make different predictions. According to Dautriche et al. (2017), regularity in the lexicon is due to re-use of phonological patterns, which should produce particularly high peaks among high-probability forms, with less of an effect on the low-probability tail. In the superadditivity account, regularity is due to combinations of markedness violations resulting in such low probabilities that many of them never appear, resulting in a shorter and thinner tail, with less of an effect on the shape of the peak. Both pressures could also co-exist. It would be informative for future work to tease apart the predictions made by each account.

References

Adam Albright. ms. Cumulative violations and complexity thresholds.

Todd M. Bailey and Ulrike Hahn. 2005. Phoneme similarity and confusability. *Journal of Memory and Language*, 52(3):339–362.

Yuen Ren Chao. 1968. *A Grammar of Spoken Chinese*. University of California Press, Berkeley and Los Angeles.

George N. Clements. 2003. Feature economy in sound systems. *Phonology*, 20(3):287–333.

John Coleman and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In John Coleman, editor, *Computational phonology: Third meeting of the ACL special interest group in computational phonology*, pages 49–56, Somerset, NJ. Association for Computational Linguistics.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi. 2017. Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.

San Duanmu. 2009. *Syllable Structure: The Limits of Variation*. Oxford University Press, New York, USA.

Marc Ettlinger and Keith Johnson. 2009. Vowel discrimination by english, french and turkish speakers: Evidence for an exemplar-based approach to speech perception. *Phonetica*, 66(4):222–242.

Edward Flemming, 2004. chapter Contrast and Perceptual Distinctiveness, pages 232 – 276. Cambridge University Press.

Lauren Franklin and James Morgan. 2017. On the nature of vocalic representation during lexical access. *The Journal of the Acoustical Society of America*, 141(5):4038–4038.

Stefan A Frisch, Nathan R Large, and David B Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of memory and language*, 42(4):481–496.

Christopher Green and Stuart Davis, 2014. *Perspectives on phonological theory and development, in honor of Daniel A. Dinnsen*, chapter Superadditivity and limitations on syllable complexity in Bambara words, pages 223–247. John Benjamins, Amsterdam.

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

Tsan Huang and Keith Johnson. 2010. Language specificity in speech perception: Perception of mandarin tones by native and nonnative listeners. *Phonetica*, 67(4):243–267.

Shudong Huang, Xuejun Bian, Grace Wu, and Cynthia McLemore, 1997. *LDC Mandarin Lexicon*. University of Pennsylvania.

Daniel Jurafsky and James H. Martin. 2008. *Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition*. Prentice Hall. Pearson Education, Inc., Upper Saddle River, New Jersey, 2nd edition.

Brett Kessler and Rebecca Treiman. 1997. Syllable structure and the distribution of phonemes in english syllables. *Journal of Memory and language*, 37(3):295–311.

Charles N. Li and Sandra A. Thompson, 1987. *The World's Major Languages*, chapter Chinese. Oxford University Press.

Paul A Luce and David B Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1):1.

Steven T. Piantadosi, Harry J. Tily, and Edward Gibson. 2009. The communicative lexicon hypothesis. In *The 31st annual meeting of the Cognitive Science Society (CogSci09)*, pages 2582–2587.

Chelsea Sanker. 2016. *Patterns Of Misperception Of Arabic Guttural And Non-Guttural Consonants*. Ph.D. thesis.

Stephanie S. Shih. 2016. Super additive similarity in dioula tone harmony. In Kyeong min Kim, Pocholo Umbal, Trevor Block, Queenie Chan, Tanie Cheng, Kelli Finney, Mara Katz, Sophie Nickel-Thompson, and Lisa Shorten, editors, *Proceedings of the 33rd West Coast Conference on Formal Linguistics*, pages 361–370. Cascadilla Proceedings Project, Somerville, MA, USA.

Kevin Tang. 2015. *Naturalistic speech misperception*. Ph.D. thesis.

Andrew Wedel, Scott Jackson, and Abby Kaplan. 2013. Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and speech*, 56(3):395–417.

Robert L. Weide, 2008. *The CMU pronunciation dictionary*. Carnegie Mellon University, 0.7a edition.

Katherine S. White and James L. Morgan. 2008. Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, 59:114–132.

George K. Zipf. 1929. Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, 40:1–95.