COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2018) 1–19

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12689

# Word Forms Are Structured for Efficient Use

Kyle Mahowald,[a] Isabelle Dautriche,[b] Edward Gibson,[a]
Steven T. Piantadosi[c]

[a] *Department of Brain and Cognitive Science, MIT*
[b] *School of Philosophy, Psychology and Language Sciences, University of Edinburgh*
[c] *Department of Psychology, University of California, Berkeley*

Received 15 May 2015; received in revised form 11 August 2018; accepted 30 August 2018

**Abstract**

Zipf famously stated that, if natural language lexicons are structured for efficient communication, the words that are used the most frequently should require the least effort. This observation explains the famous finding that the most frequent words in a language tend to be short. A related prediction is that, even within words of the same length, the most frequent word forms should be the ones that are easiest to produce and understand. Using orthographics as a proxy for phonetics, we test this hypothesis using corpora of 96 languages from Wikipedia. We find that, across a variety of languages and language families and controlling for length, the most frequent forms in a language tend to be more orthographically well-formed and have more orthographic neighbors than less frequent forms. We interpret this result as evidence that lexicons are structured by language usage pressures to facilitate efficient communication.

*Keywords:* Lexicon; Word frequency; Phonology; Communication; Efficiency

## 1. Introduction

While there is no a priori reason why a *dog* should be called a *dog* and a *tarantula* a *tarantula* in English instead of the other way around, there is evidence that word forms are, at least partially, constrained by their usage. Recent work suggests that pressures from communicative constraints influence the relationship between word usage and word form. One prediction based on information theory (Shannon, 1948) is that the most

Correspondence should be sent to Kyle Mahowald, Department of Brain and Cognitive Science, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail kylemaho@mit.edu

[Correction added on October 10, 2018, after initial online publication: The word “a tarantula” has been added, so that the sentence read as “While there is no a priori reason why a *dog* should be called a *dog* and a *tarantula* a *tarantula* in English instead of the other way around, there is evidence that word forms are, at least partially, constrained by their usage.”]

frequently used words should be more optimized. Zipf (1935) and others (Manin, 2006; Piantadosi, Tily, & Gibson, 2011a) suggest that the length distribution of words is optimized from a language user perspective. Yet pressures stemming from language usage may systematically affect word forms in many other ways. Predictability is one such pressure: People selectively use shorter words (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013)—or omit words altogether—when the context is predictive (Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007).

In this work, we examine another factor that affects word forms: their phonological properties. In particular, we focus on two aspects of the phonological forms of words: *phonotactic probability* and *phonological neighborhood density*.

- Phonotactic probability is a measure of the well-formedness of a string in a given language. For instance, in English, the word *drop* is phonotactically quite probable, *dwop* is less probable but still allowed, and *dsop* has, essentially, zero probability.
- Phonological neighborhood density of a word *w* is the number of words that differ from word *w* by insertion, deletion, or substitution of a phoneme (Luce, 1986; Vitevitch & Luce, 1998). For instance, the neighbors of *cat* include *mat* and *can*.

If a lexicon were structured by language usage pressures, how should word frequency be related to phonotactic probability and neighborhood density? Zipf (1935) claimed that the Principle of Least Effort predicts that easily articulated sounds should be used more often in language than more difficult sounds. While Zipf was referring to individual sounds, there is compelling evidence that phonotactically probable words are easier to produce in language use. For instance, the inventory of sounds in languages evolves to enable easy articulation and perception (Lindblom, 1983, 1990, 1992), and the patterns of sounds observed across languages reflect articulatory constraints (Kawasaki & Ohala, 1980). Therefore, a language whose most frequent words are phonotactically *probable* likely requires less production effort than a language organized such that the most frequent strings are phonotactically *improbable*.

But what about from the listener's perspective? One line of thinking suggests that what is good for the speaker is good for the listener. Just as speakers have an easier time producing frequent sound sequences, listeners are also more adept at perceiving these sound sequences. Phonotactically probable words are more easily recognized than less probable words (Vitevitch, 1999). There is also a learning advantage for probable strings: Probable strings are learned more easily by infants and children (Coady & Aslin, 2004; Storkel, 2004, 2009; Storkel & Hoover, 2010) and infants prefer high-probability sequences of sounds compared to lower probability sequences (Jusczyk & Luce, 1994; Ngon et al., 2013). These lines of evidence suggest a functional advantage for phonotactically probable words not just in production but in comprehension as well. From this, we expect that the most frequent words are phonotactically more probable than the most infrequent strings.

There are conflicting accounts as to how phonological neighborhood density affects language production and comprehension and, therefore, differing predictions about how the lexicon should be organized in terms of phonological neighborhood density. Because everyday linguistic communication takes place in adverse conditions (e.g., environmental noise, errors of production, and perception), words with many neighbors cause increased processing difficulty since they are typically more confusable with other words. And indeed there is evidence that having many neighbors can have an inhibitory effect on lexical access in perception (Luce, 1986; Vitevitch & Luce, 1998), inhibit reading times Magnuson, Dixon, Tanenhaus, and Aslin (2007), and elicit lexical competition that slows down word learning in toddlers (Dautriche, Swingley, & Christophe, 2015; Swingley & Aslin, 2007).

But Sadat, Martin, Costa, and Alario (2014) find that phonological neighborhood density causes longer naming latencies in production and, therefore, has an inhibitory effect. Vitevitch and Stamer (2006), like Sadat et al. (2014), argue that morphologically rich languages like Spanish and French typically show an inhibitory effect for words with many neighbors in naming tasks. Yet it has also been shown that phonological similarity (a) facilitates the ease with which people produce words (Gahl, Yao, & Johnson, 2012; Stemberger, 2004; Vitevitch & Sommers, 2003); (b) supports novel word representation in working memory (Storkel & Lee, 2011), and (c) boosts word learnability in adults (Storkel, Armbruster, & Hogan, 2006). Whether neighborhood effects are facilitative or inhibitory in production may be task dependent (Chen & Mirman, 2012). In such a case, it is difficult to predict how neighborhood density should be distributed with regard to frequency to be optimal for language usage purposes. Following the prediction of information-theoretic accounts (Shannon, 1948), the most frequent words tend to be the most optimized for language communication. Therefore, the directionality of the relationship between neighborhood density and frequency (assuming that it is consistent across many languages) is informative about how neighborhood density should be optimal for language usage.

To evaluate the extent to which the phonological forms of words may be explained by word usage, we analyzed the lexicons of 96 typologically diverse languages downloaded from Wikipedia. For each word form composing these lexicons we calculated its orthographic probability (a proxy for phonological probability) and its orthographic neighborhood density (a proxy for phonological neighborhood density). We investigated (a) the relationship between frequency (by token) and orthographic probability (as measured over word types) and (b) the relationship between frequency and neighborhood density. In all of these analyses, we compared only words of the same length so that any of resulting effects are not driven by word length. To assess significance, we used the method described in Dautriche, Mahowald, Gibson, Christophe, and Piantadosi (2017), whereby we generate "null" lexicons using an n-phone model trained over unique word forms in the real lexicon. Thus, for each language, we generate N simulated lexicons with the same frequency distribution as the real lexicon and ask whether we see the same correlation between the factors above in the simulated lexicons and in the real lexicon.

Given previous experimental results showing an advantage for phonotactically well-formed word forms, information-theoretic accounts predict that there should be a consistent positive correlation between frequency and orthographic probability across languages. The experimental results on the influence of neighborhood density in language processing are not clear and could go either way, yet a *consistent* positive or negative correlation between frequency and neighborhood density would be informative about how neighborhood density should be optimal for language usage. However, if we observe no consistent correlations between orthographic probability and frequency and between frequency and neighborhood density, it would suggest either that languages differ in how they have evolved or that these two variables are not optimized for language usage.

These sorts of correlations have been examined in the literature before, but only for a small number of languages. Landauer and Streeter (1973) performed a similar analysis for English, and Frauenfelder, Baayen, and Hellwig (1993) for English and Dutch. All found that the most frequent words in the language have higher phonotactic probability and more phonological neighbors than more infrequent words. But it is difficult to draw conclusions on the functional nature of these correlations based on just a small number of languages. It is particularly difficult to do so given the large differences that exist across language families. For instance, there are meaningful differences in the distribution of word length or in the size of the phoneme inventory that may jointly influence the contribution of neighborhood density and orthographic probability in different languages. Vitevitch and Stamer (2006) suggest that differences in morphological complexity between Spanish and English, for instance, lead to different neighborhood effects.

Therefore, in this work we take a breadth-based approach and examine a large range of typologically different languages using an orthographic corpus. Crucially, unlike previous studies of the relationship between word form and frequency, these results include a very wide range of languages with a correspondingly wide range of morphological complexity and structure. Thus, if the previously attested relationship between phonotactic probability and frequency is simply a byproduct of, say, shared morphological processes or shared word formation processes common to the Germanic and Latin languages most commonly studied in previous research, then we would expect to see different results in the typologically varied languages that we consider.

## 2. Method

### 2.1. Lexicons

We used the lexicons of 96 languages extracted from Wikipedia. The details on these lexicons, including the typological details and our corpus cleaning procedure, are explained in Appendix A. The languages analyzed included 62 Indo-European languages and 34 non-Indo-European languages. Of the non-Indo-European languages, 12 language families are represented as well as a creole. The languages analyzed are shown in Tables 1 and 2.

For this analysis, we selected the set of the 20,000 most frequent unique orthographic word forms (word types) in a given language. From this set, we defined as a lexicon all words of length three to seven letters for each language (in characters for orthographic lexicons and in phonemes for phonemic lexicons).

The number of tokens in the original Wikipedia corpus for each language ranged from 118,800 tokens for the language with the smallest Wikipedia corpus to 14.4 billion tokens for English. The median language contained 7.7 million tokens. After the restrictions described above (focusing on the top 20,000 most frequent word types and then words of

Table 1
List of Indo-European languages used, with language families in bold

**Albanian**: Albanian; **Armenian**: Armenian; **Baltic**: Lithuanian, Latvian; **Celtic**: Breton, Irish, Scottish Gaelic, Welsh; **creole**: Haitian; **Germanic**: Afrikaans, Alemannic, German, English, Luxembourgish, Low Saxon, Dutch, Scots, West Frisian, Yiddish; **Hellenic**: Greek; **Indo-Aryan**: Fiji Hindi, Marathi, Urdu, Bosnian, Croatian, Punjabi, Serbian; **Iranian**: Central Kurdish, Persian, Kurdish, Mazandarani, Tajik; **Italic**: Latin; **North Germanic**: Danish, Icelandic, Norwegian (Nynorsk), Norwegian (Bokmal), Swedish; **Romance**: Aragonese, Asturian, Catalan, Spanish, French, Galician, Italian, Lombard, Neapolitan, Occitan, Piedmontese, Portuguese, Romanian, Sicilian, Venetian, Walloon; **Slavic**: Belarusian, Bulgarian, Macedonian, Czech, Polish, Russian, Serbo-Croatian, Slovene, Slovak, Ukrainian

Table 2
List of non-Indo-European languages used, with language families in bold

**Afro-Asiatic**: Arabic, Amharic, Egyptian Arabic, Hebrew; **Altaic**: Mongolian, Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Turkish, Tatar, Uzbek; **Austronesian**: Minang, Malagasy, Indonesian, Malay, Sundanese, Cebuano, Tagalog, Waray-Waray, Buginese, Javanese; **Austroasiatic**: Vietnamese; **Kartvelian**: Georgian; **Niger-Congo**: Swahili, Yoruba; **Quechuan**: Quechua; **Tai-Kadai**: Thai; **Uralic**: Estonian, Finnish, Hungarian; **Vasonic**: Basque

only length three to seven), we were left with between 6,408 word types and 18,240 word types to analyze.

To assess whether the Wikipedia corpus (which uses orthographic forms and contains morphologically complex words) is a good proxy for a more controlled corpus that uses phonemic representations and is restricted to monomorphemic words, we also analyzed phonemic lexicons derived from CELEX for Dutch, English, and German (Baayen, Piepenbrock, & Gulikers, 1995) and Lexique for French (New, Pallier, Brysbaert, & Ferrand, 2004). The lexicons were restricted to include only monomorphemic lemmas (coded as “M” in CELEX; I.D., a French native speaker, identified monomorphemes by hand for French). That is, the lexicons contained neither inflectional affixes (like English plural *-s*) nor derivational affixes like the English *-ness*. In order to focus on the most used parts of the lexicon, we selected only words whose frequency is greater than 0. (The CELEX database includes some rare words listed as having 0 frequency, which were not in the original CELEX sample.) Since we used the surface phonemic form, when several words shared the same phonemic form (e.g., “bat”), we included this form only once.

All three CELEX dictionaries were transformed to make diphthongs into two-character strings. In each lexicon, we removed a small set of words containing foreign characters. This resulted in a lexicon of 5,459 words for Dutch, 6,512 words for English, 4,219 words for German, and 6,782 words for French. The resulting lexicons are available at https://osf.io/rvg8d/.

### 2.2. *Variables under consideration*

For each word in each language we computed the word’s:

- *Word length:* for orthographic lexicons, in characters; for phonemic lexicons, in phones (so that we can compare only words of the same length)

- *Token frequency:* for orthographic lexicons: across all the Wikipedia corpus of the language; for phonemic lexicons: using the frequency in CELEX or Lexique.
- *Orthographic or phonotactic probability:* for orthographic lexicons, we trained an n-gram model on characters ($n$ = 3 with a Laplace smoothing of 0.01 and with Katz back-off in order to account for unseen but possible sound sequences; see (Dautriche et al., 2017)) on each lexicon and used the resulting model to find the probability of each word string under the model. Table 3 shows examples of high and low probability English words under the English language model.

  For phonemic lexicons, we proceeded the same way but the n-gram model was trained on phones rather than characters.

- *Orthographic or phonological neighborhood density:* for orthographic lexicons, we calculated the orthographic neighborhood density of words (as a proxy for phonological neighborhood density) and for phonemic lexicon we calculated their phonological neighborhood density. Orthographic/Phonological neighborhood density is defined for each word as the number of other words in the lexicon that are one edit (an insertion, deletion, or substitution) away in orthographic/phonological space (Luce, 1986; Luce & Pisoni, 1998). For instance, “cat” and “bat” are phonological neighbors, as well as minimal pairs since they have the same number of letters and differ by 1. “Cat” and “cast” are neighbors but not minimal pairs. We will only focus on minimal pairs, as opposed to neighbors, in order to avoid confounds from languages having different distributions of word lengths.

Table 3
Phonotactically likely and unlikely words in English with their log probabilities

| Word | Log probability |
|---|---|
| reed | −3.69 |
| shed | −3.75 |
| mention | −4.63 |
| comment | −4.68 |
| tsar | −8.64 |
| iowa | −9.47 |
| kremlin | −11.53 |
| tsunami | −12.90 |

## 3. Results

### 3.1. Large-scale effects of frequency on 96 languages

#### 3.1.1. Correlational analysis

Figs. 1 and 2 shows correlations for each language and length (from four to six letters) separately, between (a) orthographic probability and frequency and (b) minimal pairs and

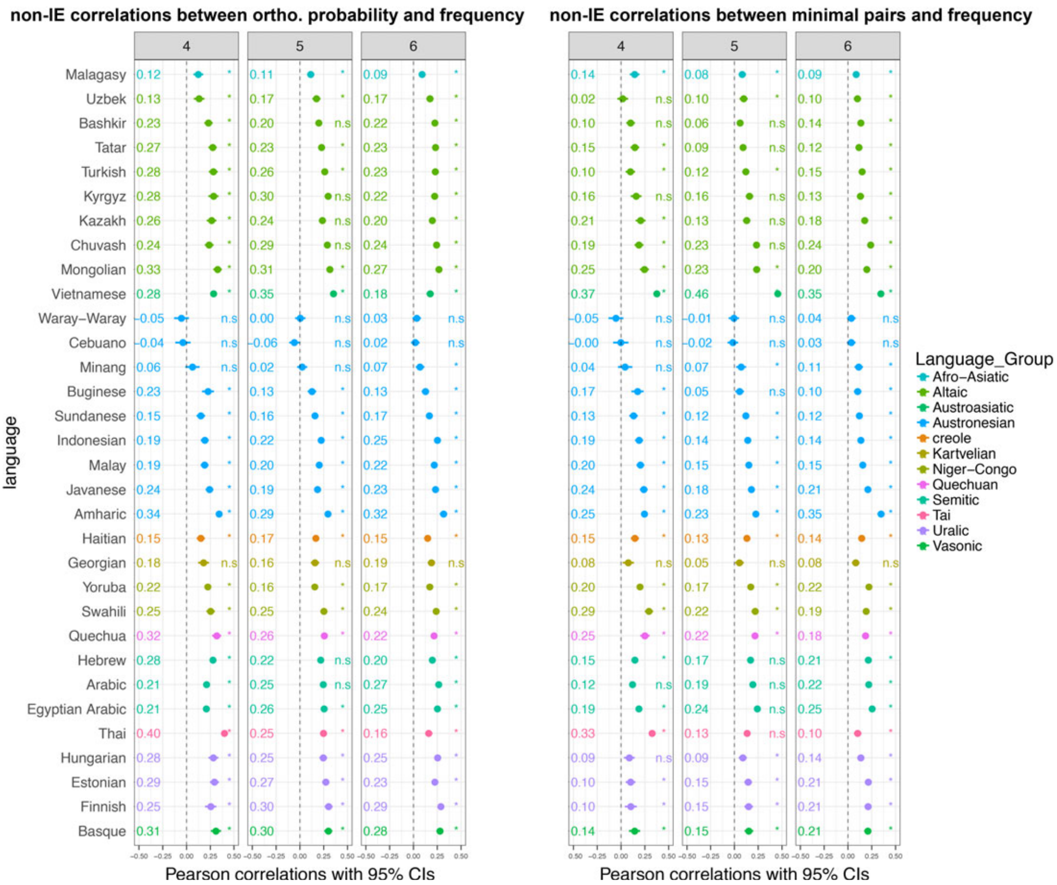

Fig. 1. Correlation coefficients between (left) orthographic probability and frequency and (right) minimal pairs and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length four to six letters for non-Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the *y*-axis are the Pearson correlations. Text and points are colored by language family. The presence of a star on the right of the graph indicates that the correlation is significant at $p < .01$ compared to the distribution of correlations obtained across 1,000 simulated lexicons.

frequency for non-Indo-European languages (Fig. 1) and for Indo-European languages (Fig. 2). Dots to the right of the dotted line at 0 show a positive correlation. Almost all languages show a positive correlation.[1]

To evaluate whether the correlations between frequency and orthographic probability and between frequency and orthographic neighborhood are driven by language usage pressures above and beyond what would be expected by chance, we need to compare these correlations to a baseline. This baseline would reflect what these correlations would be like in the absence of language usage pressure. We created such a baseline by following the procedure used by Dautriche et al. (2017). We selected the orthographic model that best reproduces the orthographic processes which are at play in each language[2] and

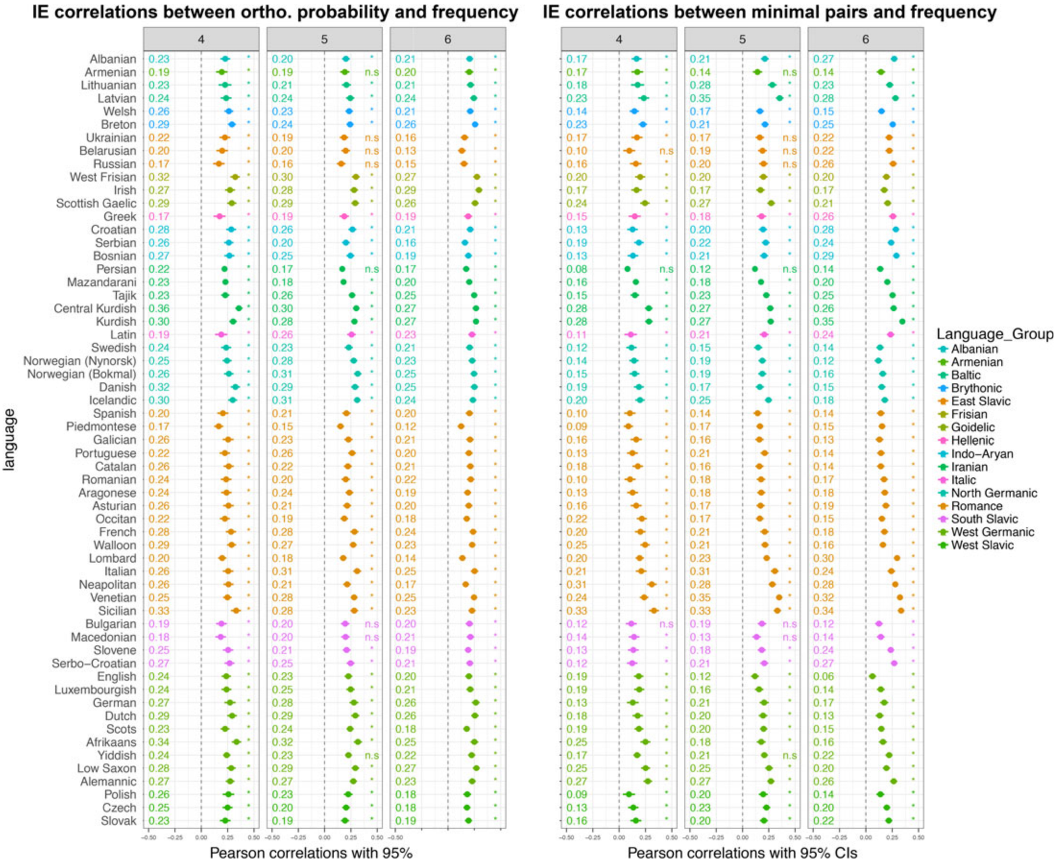

Fig. 2. Correlation coefficients between (left) orthographic probability and frequency and (right) minimal pairs and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length four to six letters for non-Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the *y*-axis are the Pearson correlations. Text and points are colored by language family. The presence of a star on the right of the graph indicates that the correlation is significant at $p < .01$ compared to the distribution of correlations obtained across 1,000 simulated lexicons.

used that resulting language models to generate words for 1,000 simulated lexicons for each language. The number of words in each simulated lexicon was matched to the number of words in each of the real lexicons (20,000 unique strings) and respected the word length distribution in each language. For each simulated lexicon, we then randomly reassigned the frequencies of the words in the real lexicon to the words of the simulated lexicon of the same length, in order to preserve the frequency-length distribution observed across languages. Because our simulated lexicons are generated independently of the properties we are interested in (correlations between frequencies and neighborhood density and between frequency and orthographic probability), they can be used as a statistical baseline with which the real lexicons can be compared. For each simulated lexicon, we computed Pearson correlations between log frequency and orthographic probability and between log frequency and the number of minimal pairs. We computed a $z$-score using the mean and standard deviation of the transformed correlation scores estimated from the 1,000 simulated lexicons for each language. The $p$-value reflects the probability that the real lexicon correlations could have arisen by chance.[3]

Analyzing each length separately and focusing on words of three to seven letters, we found a significant correlation between log frequency and orthographic probability in most languages (see Table 4). For instance, for the four-letter words, 94 out of 96 languages showed a positive correlation and 92 of these correlations were significantly positive at $p < .01$ when compared to the simulated baseline.

We also found a robust correlation between log frequency and number of minimal pairs for almost all languages, as shown in Table 5.

Additionally, we unsurprisingly find a robust correlation between orthographic probability and number of minimal pairs (mean $r = 0.47$ when we average across the correlations found for each length from 3 to 7; $r = 0.44$ across the simulated lexicons). This result follows trivially from the fact that a phonetically probably word like "set" is more

Table 4
Summary of relationship between orthographic probability and frequency, across languages. Separated by length, (a) the mean correlation across languages for the relationship between orthographic probability and frequency, (b) the proportion of languages that show a positive correlation between orthographic probability and frequency, and (c) the proportion of languages for which this relationship is significantly different from chance at $p < .01$, chance being the correlation obtained in 1,000 simulated lexicons

| Word Length | Mean Correlation | Proportion Showing Positive Correlation | Proportion Showing Significant Correlation |
|---|---|---|---|
| 3 letters | 0.27 | 1 | 0.88 |
| 4 letters | 0.24 | 0.98 | 0.96 |
| 5 letters | 0.23 | 0.99 | 0.81 |
| 6 letters | 0.21 | 1 | 0.97 |
| 7 letters | 0.19 | 1 | 0.76 |

Table 5
Summary of relationship between minimal pairs and frequency, across languages. Separated by length, (a) the mean correlation across languages for the relationship between the number of minimal pairs and frequency, (b) the proportion of languages that show a positive correlation between the number of minimal pairs and frequency, and (c) the proportion of languages for which this relationship is significantly different from chance at $p < .01$, chance being the correlation obtained in 1,000 simulated lexicons

| Word Length | Mean Correlation | Proportion Showing Positive Correlation | Proportion Showing Significant Correlation |
|---|---|---|---|
| 3 letters | 0.19 | 1 | 0.69 |
| 4 letters | 0.17 | 0.98 | 0.88 |
| 5 letters | 0.18 | 0.98 | 0.78 |
| 6 letters | 0.19 | 1 | 0.97 |
| 7 letters | 0.18 | 0.99 | 0.76 |

Table 6
Separated by length, the model coefficient from the full model including random intercepts and slopes for language, subfamily, and family for orthographic probability and number of minimal pairs. Two asterisks means that by a likelihood test, the predictor significantly improves model fit at $p < .01$. Three asterisks means $p < .001$. The coefficients can be interpreted as the following: For four-letter words, a 1 *SD* increase in orthographic probability is predictive of a 0.21 *SD* increase in frequency, and a 1 *SD* increase in number of minimal pairs is predictive of a 0.05 *SD* increase in frequency

| Word Length | Orthographic Probability | Number of Minimal Pairs |
|---|---|---|
| 3 letters | 0.23*$^{*}$ | 0.08*$^{*}$ |
| 4 letters | 0.21**$^{*}$ | 0.05**$^{*}$ |
| 5 letters | 0.19**$^{*}$ | 0.07*$^{*}$ |
| 6 letters | 0.15**$^{*}$ | 0.11**$^{*}$ |
| 7 letters | 0.13**$^{*}$ | 0.11**$^{*}$ |

likely to have more minimal pairs in English than the word "quiz" simply because the letter sequences in "set" are more common and so, probabilistically, there are more opportunities for a word to be orthographically close to "set" than to "quiz."

It follows that the correlations between frequency and phonological similarity that were uncovered previously should be (partly) due to both frequency and orthographic probability being correlated with phonological similarity. Because languages structure the vocabulary in different ways (in particular because they have different phoneme inventories and different word length distributions), the contribution of orthographic probability versus neighborhood density may vary across languages. Thus, the question becomes (a) whether the correlation between frequency and neighborhood density remains after factoring out the effect of orthographic probability and (b) whether the correlation between frequency and orthographic probability remains after factoring out the effect of neighborhood density. If these two correlations remain consistent across languages, then this would suggest that these relationships are the product of language usage pressures, while if languages display different correlations, this would indicate that these are most likely the result of language-specific properties on vocabulary structure.

In addition, many of the languages in this study are highly related, so we need an analysis that generalizes across families and languages to make sure that the effect is not just lineage specific.

### 3.1.2. Mixed effect analysis

We ran a mixed effect regression predicting (scaled) frequency for each word from orthographic probability and number of minimal pairs, where both predictors were normalized for each language and length. We used a maximal random effect structure with random intercepts for each language, language subfamily, and language family and slopes for orthographic probability and number of minimal pairs for each of those grouping factors. In effect, this random effect structure allows for the possibility that some languages or language families show the predicted effect, whereas others do not (see also e.g.,

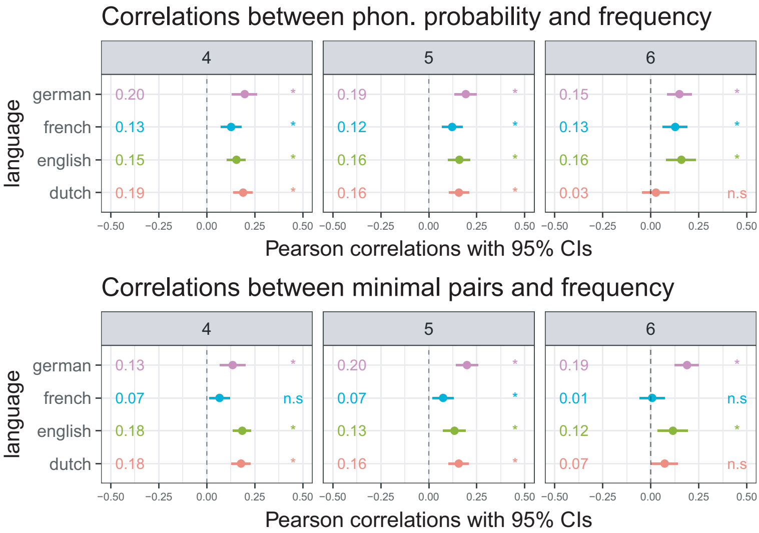

Fig. 3. Correlation coefficients between (a) phonotactic probability and frequency and (b) minimal pairs and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length four to six letters for Dutch, English, French, and German. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the *y*-axis are the Pearson correlations. The presence of a star on the right of the graph indicates that the correlation score is significant at $p < .01$ compared to the distribution of correlation scores obtained across 1,000 simulated lexicons.

Atkinson, 2011; Jaeger, Graff, Croft, & Pontillo, 2011; for a similar approach). It allows us to test whether the effect exists beyond just language-specific trends. Because of the complex random effect structure and the large number of data points, we fit each length separately and focused on words of length three through seven.

For four-letter words (a representative length), a 1 *SD* increase in orthographic probability was predictive of a 0.21 *SD* increase in frequency; a 1 *SD* increase in number of minimal pairs was predictive of a 0.05 *SD* increase in frequency. To assess the significance of orthographic probability above and beyond the number of minimal pairs, we performed a likelihood ratio test comparing the full model to an identical model without a fixed effect for orthographic probability (but the same random effect structure). The full model was significantly better by a chi-squared test for goodness of fit ($\chi^2(1) = 30.9$, $p < .0001$). To assess the significance of the number of minimal pairs above and beyond the effect of orthographic probability, we compared the full model to an identical model without a fixed effect for the number of minimal pairs using a likelihood ratio test. Once again, the full model explained the data significantly better ($\chi^2(1) = 10.6$, $p < .001$). Thus, both the number of minimal pairs and orthographic probability appear to make independent contributions in explaining word frequency. This effect holds above and beyond effects of language family or subfamily, which are included in the model as random effects. Note that the effect size is larger for orthographic probability than it is for number of minimal pairs and that a model including a fixed effect of orthographic probability but not minimal pairs has a better model fit (AIC = 520,310) than one that includes minimal pairs but not probability as a fixed effect (AIC = 520,330). We find a similar pattern of results for all other lengths examined, as summarized in Table 6. Overall, these results suggest that both the number of minimal pairs and the orthographic probability independently predict frequency but that the effect of orthographic probability is stronger and is likely, in part, driving the neighborhood effect.

### 3.2 *Testing correlation generalizability to phonemic representations*

We used orthographic lexicons because they could be easily extracted for a large number of languages, but orthography is only a proxy for phonetics. Moreover, the Wikipedia dataset does not attempt to exclude morphologically complex words. Both of these factors could add unwanted noise to our analyses.

Therefore, we also tested a subset of languages for which we had more carefully constructed lexicons with both phonemic and morphological information. Specifically, to assess whether the correlation between frequency and neighborhood density and between frequency and phonotactic probability hold in a set of monomorphemic words with phonemic representations, we performed the same analysis using the four phonemic lexicons from Dutch, English, French, and German.

As stated earlier, we calculated the Pearson correlations for each word length between token frequency and phonotactic probability (here approximated by *phonemic* probability

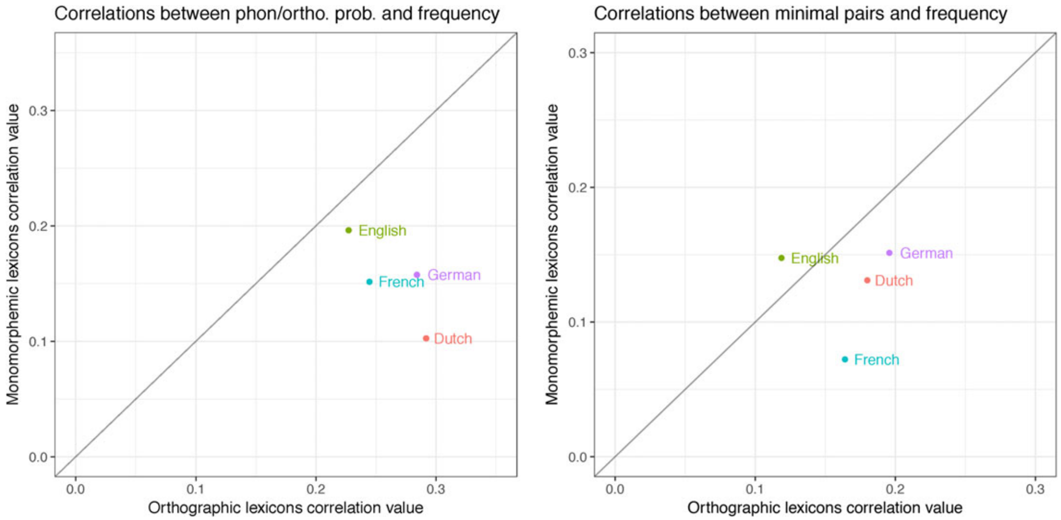

Fig. 4. Comparison of the correlations values obtained for the phonemic lexicons versus orthographic lexicons of Dutch, English, French, and German. The left panel shows the average correlations between phonotactic/orthographic probability and frequency for words of length three to seven letters. The right panel shows the average correlations between minimal pairs and frequency for words of length three to seven letters.

using an n-gram model operating over triphones) and between token frequency and the number of minimal pairs. We compared these correlations to those correlations obtained across 1,000 simulated lexicons following the same methodology as for the wikipedia lexicons. As shown in Fig. 3, the Pearson correlations obtained in these four phonemic lexicons replicated previous correlations with the orthographic lexicons for these languages: All four languages still showed positive correlations for the relationship between phonotactic probability and frequency and between the number of minimal pairs and frequency and these correlations tended to be significantly more positive than the correlations obtained across our simulated lexicons.

In Fig. 4, we compare the correlations between frequency and the number of minimal pairs and between frequency and phonotactic probability obtained in the phonemic versus the orthographic lexicons for Dutch, English, French, and German.

The correlations were slightly lower in the more controlled set for the four languages than when using the same measures in the larger dataset: the correlation between minimal pairs and frequencies (across the four languages and word lengths three to seven letters) is, on average, 0.04 lower for the correlation between minimal pairs and frequency and 0.11 lower for the correlation between orthographic/phonotactic probability and frequency. This suggests that part of the effect could be driven by morphology—which is absent in the controlled phonemic lexicons but present in the Wikipedia corpus.

## 4. Discussion

We found that frequent word forms are more likely to be well-formed (phonotactically or orthographically probable) and similar to other word forms (higher neighborhood density) than infrequent ones. These correlations were robustly present across a large number and wide variety of typologically different languages. Just as the Zipfian word frequency distribution allows for functional optimization of word lengths (Piantadosi, 2014; Piantadosi et al., 2011a), this work shows that the frequency profile of words of the same length is structured in a non-arbitrary way so as to maximize the use of high-probability, high-neighborhood-density word forms.

Importantly, we do not believe that the main result of this paper is purely a result of morphological regularity since the same analyses run on monomorphemic words in a subset of languages show the same pattern of results. Moreover, although phonotactic constraints are an obvious and major source of regularity in the lexicon, it is important to note that these results are not likely just the result of phonotactic constraints since the results hold even after controlling for the influence of orthographic or phonotactic probability through simulated baselines.

Furthermore, the present results suggests that word form similarity (as measured by neighborhood density) is a desirable feature for frequent words. This echoes previous results of ours: In Dautriche et al. (2017), we provided evidence that natural lexicons are more tightly clustered in phonological space than would be expected by chance, over and above the constraints imposed by phonotactics. This, taken together with the present results, suggests that languages tend to favor word form similarity in the lexicon.

In addition, while poor perceptual distinctiveness among frequently used forms may look to be disadvantageous during language comprehension, there are reasons to believe that this may be a simplistic story. Many studies in phonology have shown that recurrent sound patterns in the language result from the repeated interaction of perceptual and articulatory needs (e.g., Chomsky & Halle, 1968; Hume & Johnson, 2001; Ohala, 1993). For instance, perception has been shown to be the main explanation for the use of CV syllables patterns cross-linguistically (Steriade, 1997), the nasal place assimilation (Beddor & Evans-Romaine, 1992), and vowel reduction (van Bergem, 1995). This as a whole suggests that there are well-documented influences of perception on phonetics that may interact, in the long run, with the phonological form of words, pushing them to be more similar.

The present results are in line with information-theoretic accounts of efficient communication where communicative efficiency is achieved by trading off between ease of production and transmission accuracy (Lindblom, 1990; Piantadosi, Tily, & Gibson, 2011b; Zipf, 1949) and even with accounts that focus only on transmission accuracy (Ferrer-i-Cancho & Solé, 2003; Pate & Goldwater, 2015) as the same words that are easy to use by speakers may also be easy to process by listeners (Brown, 1991; Ferreira, 2003), increasing the chances of successful transmission of a message.

In this study, we addressed the issue of whether the frequency profile of the phonological forms of words varies systematically across languages, but we leave it to future work to investigate how it got to be that way. Indeed, while language usage has been put forward as a powerful explanatory framework to explain the length distribution of words in the lexicon, it is only recently that experimental studies have shown that such a relationship emerges from pressures for communicative accuracy and efficiency that could be observed in simple language games in the lab (Kanwal, Smith, Culbertson, & Kirby, 2017). A similar approach could be used to test whether the same language usage pressures shape the phonological profile of words. One plausible mechanism for the effects described here is that generations of language users improve on the lexicon, honing it over time by avoiding words that are too strange, complex, or that otherwise do not fit with the rest of the words in the lexicon (following the experimental work looking at the evolution of language showing that language users will preferentially discard forms and structures that are disadvantageous in favor of other, fitter words and phrases; Fedzechkina, Jaeger, & Newport, 2012; Hills & Adelman, 2015; Smith, Kirby, & Brighton, 2003).

## Acknowledgements

K. M. was funded in part by an N.S.F. Dissertation Completion Grant. I. D. was funded by an ESRC Future Research Leaders Award (ES/N0174041/1) and European Research Council (ERC-2011-AdG-295810 BOOTPHON). The authors also thank members of Tedlab and audiences at the CUNY Conference for Human Sentence Processing for helpful comments.

## Notes

1. In order to ensure that any of the observed correlations are not the product of English overlap, we ran the same analyses on the full lexicons as well as on subsets of lexicons that exclude any word that also appears in the English Subtlex subtitles database (Brysbaert & New, 2009). This not only excludes English intrusions but also excludes perfectly good words like *die* in German (which means "the" and is unrelated to English "die") and French *dire* (meaning "to say" and unrelated to the English adjective *dire*). Note that, for all lengths, the results obtained when excluding all English words are similar in terms of overall correlation, as can be seen in Table 5. Because most of the English words excluded are actually not intrusions but are native words that just happen to also be English forms, we include them in all subsequent analyses. Note that this method does not account for the possibility of borrowings in other languages (and indeed many less widely spoken languages will borrow words from nearby major languages and these borrowings may have different phonotactics). We consider this phenomenon, however, to be part of the natural evolution of language and do not attempt to exclude it. In excluding English, we primarily seek to exclude words that appear in Wikipedia due to computer issues (HTML tags such as <head>, <title>, chunks of English erroneously copied into the text of other languages, etc.).
2. For each language, we selected our model by its ability to generate candidate words that are scored to have a high probability in the language considered. Similar to Dautriche et al. (2017), we compared several n-gram models (with back-off and Laplace smoothing of 0.01 based on Dautriche et al. [2017]) over letters where n varied from 1 to 10. Each model was trained on 75% of the lexicon of each language (training set), and evaluated on the remaining 25% of the lexicon (testing set) to determine which model gives the highest sum of log-probability over all words in the training set. This process was repeated over 50 random splits of the dataset into training and testing sets. The seven-letter model gave the best results for 76 of the orthographic lexicons and the nine-phone model for the 20 remaining. The high degree of the best n-gram model is expected given the word length distribution of our lexicons (highly right skewed toward longer words).
3. We also tried another baseline whereby we simply permuted the real lexicon frequencies of words of the same length and then re-examined the correlations. In the interest of space, we do not report those results here. But the findings were similar in that we again found that the real lexicon's correlations between frequency and probability and frequency and neighborhood density were significantly greater than in our simulated baselines.
4. We excluded Gujarati, Telugu, Tamil, Bishnupriya Manipuri, Cantonese, Newar, Bengali, Japanese, Hindi, Malayalam, Marathi, Burmese, Nepali, Kannada, and Korean.

## References

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*(6027), 346–349. https://doi.org/10.1126/science.1199295.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (release 2)[cd-rom]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Beddor, P. S., & Evans-Romaine, D. K. (1992). Acoustic-perceptual factors in nasal place assimilation. *The Journal of the Acoustical Society of America*, *91*(4), 2473–2473.

van Bergem, D. R. (1995). Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress. *Speech Communication*, *16*(4), 329–358.

Brown, P. M. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the model of the listener. *Bridges between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman*, *105*, 117–142.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417.

Chomsky, N., & Halle, M. (1968). The sound pattern of English. New York: Harper and Row.

Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, *89*(3), 183–213. https://doi.org/10.1016/j.jecp.2004.07.004.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128–145.

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902. https://doi.org/10.1073/pnas.1215776109.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language*, *48*(2), 379–398.

Ferrer-i-Cancho, R., & Solé, R. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 788.

Frank, A., & Jaeger, T. (2008). *Speaking rationally: Uniform information density as an optimal strategy for language production*. Washington, DC: Cognitive Science.

Frauenfelder, U., Baayen, R., & Hellwig, F. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, *32*(6), 781–804. https://doi.org/10.1006/jmla.1993.1039.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006.

Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition*, *143*, 87–92.

Hume, E., & Johnson, K. (2001). *A model of the interplay of speech perception and phonology*. OSU Working Papers in Linguistics *55*, 1–22.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*(2), 281–319.

Jusczyk, P., & Luce, P. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645. https://doi.org/10.1006/jmla.1994.1030.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Kawasaki, H., & Ohala, J. J. (1980). Acoustic basis for universal constraints on sound sequences. *The Journal of the Acoustical Society of America*, *68*(S1), S33–S33.

Landauer, T., & Streeter, L. (1973 April). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*(2), 119–131. Available at http://linkinghub.elsevier.com/retrieve/pii/S0022537173800015. Accessed (07 July 2014) https://doi.org/10.1016/s0022-5371(73)80001-5

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt & T. Hoffman (Eds.), *Advances in neural information processing systems* (pp. 849–856). Cambridge, MA: MIT Press.

Lindblom, B. (1983). Economy of speech gestures. In *The production of speech* (pp. 217–245). New York: Springer.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Dordrecht, the Netherlands: Springer.

Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. *Phonological Development: Models, Research, Implications*, *131*, 131–163.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception. Technical Report No. 6*.

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

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156. https://doi.org/10.1080/03640210709336987.

Mahowald, K., Fedorenko, E., Piantadosi, S., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, https://doi.org/10.1016/j.cognition.2012.09.010.

Manin, D. (2006). Experiments on predictability of word in context and information rate in natural language. *Journal of Information Processes*, *6*, 229–236.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

Ohala, J. J. (1993). Coarticulation and phonology. *Language and Speech*, *36*(2–3), 155–170.

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17.

Piantadosi, S. (2014 October). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6.

Piantadosi, S., Tily, H., & Gibson, E. (2011a). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526.

Piantadosi, S., Tily, H., & Gibson, E. (2011b). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*, 3526–3529.

Sadat, J., Martin, C. D., Costa, A., & Alario, F.-X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive Psychology*, *68*, 33–58.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 623–656.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386. https://doi.org/10.1162/106454603322694825.

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90* (1), 413–422.

Steriade, D. (1997). Phonetics in phonology: The case of laryngeal neutralization. Unpublished manuscript. University of California, Los Angeles.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(02), 201–221. https://doi.org/10.1017/S0142716404001109.

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, *36*(02), 291. https://doi.org/10.1017/S030500090800891X.

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. https://doi.org/10.1044/1092-4388(2006/085).

Storkel, H. L., & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken American English. *Behavior Research Methods*, *42*(2), 497–506. https://doi.org/10.3758/BRM.42.2.497.

Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. https://doi.org/10.1080/01690961003787609.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99.

Vitevitch, M. S. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1–2), 306–311. https://doi.org/10.1006/brln.1999.2116.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, *31*(4), 491–504.

Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, *21*(6), 760–770.

Zipf, G. (1935). *The psychology of language*. New York: Houghton-Mifflin.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.

## Appendix A: Dataset of 96 lexicons from Wikipedia

We started with lexicons of 115 languages from their Wikipedia databases (https://dumps.wikimedia.org). We then excluded languages for which a spot-check for non-native (usually English) words in the top 100 most frequent words in the lexicon between three and seven characters revealed more than 80% of words were not native. In this way, languages that used non-alphabetic scripts (like Chinese) were generally excluded since the three- to seven-letter words in Chinese Wikipedia are often English. After these exclusions, 96 languages remained.[4] We analyzed the data both with and without these exclusions, and the exclusions do not significantly affect the overall direction or magnitude of the results. The final languages included 62 Indo-European languages and 34 non-Indo-European languages. Of the non-Indo-European languages, there are 12 language families represented.

To get a sense of how clean these Wikipedia lexicons are, we randomly sampled 10 languages for which we then inspected the 100 most frequent words and an additional 100 random words to look for intrusion of English words, HTML characters, or other undesirable properties.

For the top 100 words in the lexicons of the 10 sampled languages, we found at most three erroneous words. For the same languages, we also inspected a randomly selected 100 words and found that the mean number of apparently non-intrusive words was 93.5 (with a range from 85 to 99). The most common intrusion in these languages was English words.