Word forms reflect trade-offs between speaker effort and robust listener recognition

Stephan C. Meylan

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Thomas L. Griffiths

Departments of Psychology and Computer Science, Princeton University

Abstract

How do cognitive pressures shape the lexicons of natural languages? Here we reframe Zipf's (1935) "law of abbreviation" within a more general framework that relates it to cognitive pressures that affect speakers and listeners. In this new framework, speakers' drive to reduce effort (Zipf's proposal) is counteracted by the need for low frequency words to have word forms that are sufficiently distinctive to allow for accurate recognition by listeners. To support this framework, we replicate and extend recent work using the prevalence of sub-word phonemic sequences (phonotactic probability) to measure speakers' production effort in place of Zipf's measure of length. Across languages and corpora, phonotactic probability is more strongly correlated with word frequency than word length. We also show this measure of ease of speech production (phonotactic probability) is strongly correlated with a measure of perceptual difficulty that indexes the degree of competition from alternative interpretations in word recognition. This is consistent with the claim that there must be trade-offs between these two factors, and is inconsistent with a recent proposal that phonotactic probability facilitates both perception and production. To our knowledge, this is the first work to offer an explanation why long, phonotactically-improbable word forms remain in the lexicons of natural languages.

Keywords: word recognition, corpus linguistics; communicative efficiency; language universals; spoken word recognition; Bayesian inference; information theory

Word forms reflect trade-offs between speaker effort and robust listener recognition

1 Introduction

While natural languages are highly diverse in many respects, they display striking structural regularities (Greenberg, 1963; Evans and Levinson, 2009; Futrell et al., 2015). How these structural regularities relate to human cognitive and physiological constraints—which affect both speaking and listening—remains an open question with implications for linguistics, psychology, and neuroscience (Hauser et al., 2002; Evans and Levinson, 2009; Kemp and Regier, 2012; Fedzechkina et al., 2012). Insights about languages have the potential to inform our knowledge about language users as well as vice versa: regularities in languages may provide evidence of cognitive pressures, and increasingly well-specified models of cognitive processes generate new hypotheses about the structure of languages.

One of the most robust statistical laws that describe human languages is the relationship between word frequency and word length, often called Zipf's *Law of Abbreviation*: that the "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences" (Zipf, 1935). This relationship is thought to reflect a tendency towards compression to support efficient communication (Ferrer-i Cancho et al., 2013; Mahowald et al., 2018). To date, this basic relationship between words and their corresponding word forms (*i.e.,* the sounds that compose them in speech) has been demonstrated to hold in all of the approximately one thousand languages that have been tested, with no known counter-examples (Bentz and Ferrer-i-Cancho, 2016).

Despite its robustness, questions remain as to how Zipf's Law of Abbreviation may emerge, or how it relates to the cognitive challenges inherent in both speaking and listening. Word lengths and frequencies are only two attributes in a complex network of correlated properties in the lexicon: previous work has found robust correlations between many pairs of variables among word frequency, average in-context predictability of words

(Piantadosi et al., 2011), neighborhood density (number of perceptually similar words; Luce et al., 1990), phonotactic probability (Vitevitch et al., 1999), age of acquisition (Kuperman et al., 2012), number of distinct word senses (Baayen and del Prado Martín, 2005; Casas et al., 2019), concreteness (Brysbaert et al., 2014), centrality in a semantic network (Vincent-Lamarre et al., 2016), and rate of word change (Pagel et al., 2007; Bybee, 2003). Furthermore, these word-level properties are reflected in language processing, including recognition rates in noise (Luce and Pisoni, 1998), response times in lexical decision tasks (Balota et al., 2007), reading times (Smith and Levy, 2013) and measurements of neural activity (*e.g.,* DeLong et al., 2005; Weissbart et al., 2020; Shain et al., 2020). Establishing precisely which correspondences among these variables are the most robust across languages will help identify which ones should be considered the core manifestations of cognitive pressures, and in turn improve our understanding of other aspects of language structure (though see Ladd et al., 2015 regarding the limitations of correlational studies).

To that end, in the current work we revisit Zipf's Law of Abbreviation and consider this empirical observation under a more general framework that relates the cognitive processes of speakers and listener. First, we generalize Zipf's measure of speaker's production effort as word length to a graded, probabilistic metric of the prevalence of sub-word sequences (phonotactic probability). We then consider how production effort trades off with robustness of word recognition in language processing using a Bayesian model of spoken word recognition. To preview our results: we find evidence that the classic relationship between the length of word forms and their frequency is a special case of an even broader relationship between phonotactic probability and frequency. However, we also find that high phonotactic probability incurs high competition in spoken word recognition, in that such forms are more likely to have more strong competing interpretations. While speakers may prefer to simplify and shorten words to reduce their production effort, they may be limited by listeners' requirements for sufficiently distinctive word forms in order to successfully recognize words.

1.1 Zipf's Law and Speaker Effort

Zipf (1935) originally posited that the relationship between word length and word frequency emerges from speakers' desires to minimize speaker effort to the degree possible by using the shortest form for words that are used most often, following what he later labeled the *Principle of Least Effort* (Zipf, 1949). However, a word's length is just one aspect of its form. Another aspect is the specific sequences of sounds of which it is comprised, *i.e.,* whether it is composed of common sound sequences or rare ones ("something" vs. "xylophone"). This brings the question of how word forms are comprised into contact with the language's phonotactics, or the rules and patterns that govern the arrangement of sounds within words and syllables. Phonotactics can also be used to characterize the ways in which speakers synchronically modify words, especially under- or over- articulating word forms in fluent speech. The process of *phonetic reduction*— the systematic under-articulation, shortening, weakening, or wholesale omission of linguistic material in word forms (Aylett and Turk, 2006; Bell et al., 2009; Gahl et al., 2012; Jaeger and Buz, 2017) results in appreciable variation how words are realized in fluent speech. For example, speakers may use "probly" in place of "probably" at high speech rates in conversational English. On longer, diachronic timescales, reduced or clipped variants of word forms may eclipse their long-form predecessors to become the dominant (or indeed only) form in the lexicon, *e.g., bus* superseding *omnibus* (Mahowald et al., 2013).

One useful way to characterize these processes of reduction is that any of the above edits to a word form has the potential to change its probability. For example, substituting "weaker" phonemes for stronger ones tends to result in more phonotactically probable forms (Jaeger and Buz, 2017). This includes many cases that leave word length (as measured by the number of phonemes or the number of characters) intact. Vitevitch and Luce (2005) found that speakers produce words consisting of common phoneme sequences faster, an effect which is robust for non-words and in the absence of listeners. This idea of probabilistic reduction is also consistent with a work that suggests that repeated access

and production of sound sequences results in automatic reduction, either because of practice effects in articulation or enhanced availability of word form representations in production (Bybee and McClelland, 2005).

If sequence probability is a better measure of speaker effort than word length (or duration), then Zipf's Principle of Least Effort should generalize: a more accurate measure of production cost —*i.e.,* a measure that assigns context-dependent costs to phonemic material—should correlate more strongly with word frequency. Extending work by Landauer and Streeter (1973) and Frauenfelder et al. (1993), Mahowald et al. (2018) found support for this hypothesis across a sample of 97 typologically diverse languages using corpora from Wikipedia. In interpreting this empirical result, they proposed that forms with high phonotactic probability are both easier to produce and comprehend, and that this empirical result provides evidence that languages "maximize the use of good word forms." Languages, they claim, use phonotactically probable forms for highest frequency forms because this confers processing advantages for *both* speakers and listeners. In the current work, we further motivate these claims by relating them to proposals about reduction and conduct a replication on a complementary dataset. We then re-examine their second claim—that high probability forms comprehension—as described in the next section.

1.2 Robust Word Recognition

Of course, speakers' ease of production is not the only pressure on a word's form—it must also be able to be reliably recognized by listeners. Languages have long been thought to reflect a balance between ease of production on the part of speakers and the ease of recognition on the part of listeners. Work by the early German linguist Georg von der Gabelentz in the late 19$^{\text{th}}$ century characterized language change as reflecting the competing pressures of "striving for ease" (*Bequemlichkeitsstreben*; translation in Haspelmath, 1999) and "striving for clarity" (*Deutlichkeitsstreben*, 1901). Subsequent psycholinguistic work has characterized these pressures at the level of individuals as

language processors, motivating speaker choice of hypo- vs. hyper- articulation in terms of what achieves "sufficient discriminability" (Lindblom, 1990).

With this need for sufficient discriminability in mind, we revisit the claim from Mahowald et al. (2018) that phonotactically probable forms are, in addition to being easier to produce, easier for listeners to comprehend. These authors motivated this characterization of high phonotactic probability forms as easier to comprehend by pointing to the finding that phonotactically probable words are more easily recognized than less probable ones Vitevitch et al. (1999). They also noted that children more easily learn higher probability word forms (Coady and Aslin, 2004; Hoover et al., 2010; Storkel, 2004). Interestingly, the framework offered in Lindblom (1990) makes the opposite prediction: high phonotactic probability words should be *harder* to discriminate because of the number of competing intended meanings on the part of the speaker that could have could have plausibly generated the observed acoustic data (see also Gibson et al., 2013; Meylan et al., 2021). Further, neither of the points raised by Mahowald et al. should be taken as conclusive evidence of the ease of processing phonotactically probable sequences. Vitevitch et al. (1999) did indeed find a facilitatory effect of high phonotactic probability on a speeded same-different task, however an auditory lexical decision task revealed slower recognition for nonword stimuli with high phonotactic probability. This is consistent with a number of studies that suggest that words from dense neighborhoods are recognized more slowly and less accurately than ones from sparse neighborhoods (Luce and Pisoni, 1998; Vitevitch and Luce, 1999,9; Goldinger et al., 1989). Likewise, a stronger theoretical motivation is necessary to link improved learnability for higher probability forms among children to more robust spoken word recognition among adults.

In fact, the hypothesis that higher probability word forms facilitate comprehension generates a prediction about the structure of natural languages: if indeed high phonotactic probability word forms were both easier to produce *and* comprehend, languages should then be expected to pack their entire lexicons into the highest probability, shortest word

forms possible (though there may be other constraints as well[1]). However, many phonotactically probable, short word forms are left unused by natural languages; for example, English leaves many possible word forms unused like *dob*, *fiss*, *lub*, and *gep*. This suggests the possibility of some countervailing pressure which limits reduction, for example that excessive reduction may create problems for listeners who are trying to recover speaker's intended words under noisy conditions (see also Piantadosi et al., 2012). This points to the potential utility of a computational model that make predictions about which word forms are likely to lead to recognition failures.

1.3 Current Work

In the current work, we revisit Zipf's Law of Abbreviation and the cognitive processes that might give rise to it. We provide formal characterizations of these processes: quantitative, model-based approximations of production effort and ease of comprehension, formalizing the intuitions of previous work using principles of information theory and (relatedly) Bayesian cognitive science. In the Models section below, we first clarify the relationship between word length and phonotactic probability, which provides a theoretical basis for the utility of the latter as a measure of speaker effort. We then provide an argument *against* the equivalence of easy to produce and easy to comprehend word forms, and propose a method to measure discriminability under a Bayesian model of spoken word recognition. This model shows why excessive reduction should be expected to be problematic for recognition, formalizing and extending ideas set forth in Jaeger and Buz (2017). This model predicts that word forms that are insufficiently *diagnostic* of the speaker's intended word (*i.e.,* their intended message) are likely to lead to transmission failures. Instead, speakers need to produce *sufficiently* distinctive word forms such that each word can be recognized by a listener in spoken word recognition.

[1]One obvious constraint is the need to keep consistent morphological paradigms, however such paradigms could easily be maintained while still using shorter and more probable forms. Consider for example replacing all "-ing" forms in English with a single phoneme.

We then use these models to characterize regularities seen in corpus data. In Study 1, we find evidence of a relationship between phonotactic probability and word frequency across a wide range of languages and datasets, consistent with previous work in Mahowald et al. (2018). In Study 2, we test the hypothesis phonotactic probability corresponds with the strength of competition in effects in word recognition. This analysis suggests that uncommon words must have rare word forms in order to maintain a minimum rate of recognition by listeners.

2 Models

Here we present two computational models, the first characterizing speaker effort involved in encoding words, and the second one characterizing how listener's infer which word a speaker meant on the basis of their beliefs and observed data. We present these models in the abstract, using simplified probabilistic models to provide approximate characterizations of complex linguistic structure (first model) and complex behavior (second model). The models are presented at Marr's computational level of analysis (Marr, 1982), and as such are intended to provide abstract, high-level characterizations of the processing problems that need to be solved by a language user. We explain how we parameterize these models on specific languages (*e.g.,* datasets and fitting procedures) in the Methods sections of each particular study.

2.1 A Model of Speaker Effort

The process of speech production requires that speakers retrieve a word form from memory, plan, and then execute motor actions that yield signals observable to listeners (Levelt, 1992). Difficulty in speech production may thus reflect both representational and motoric complexity. In contrast to work that tries to decompose complexity into component parts (*e.g.,* Romani et al., 2017), in the current work we do not try to distinguish between sources of complexity, and instead adopt an information theoretic measure of *information content* that reflects both sources of complexity in the

characterization of speaker effort (for an analogous proposal for information theoretic measures of language comprehension difficulty see Levy, 2008a and Wilcox et al., 2023). We motivate this simplified approach by noting that frequent usage potentially facilitates all three aspects of production (retrieval, encoding, and motor articulation).[2]

In order to have a metric that is both theoretically comparable and on a similar numerical scale to word length, we take the *phonological information content*, or the negative log probability of a phoneme sequence under a probabilistic phonotactic model (Cohen Priva, 2008; Mahowald et al., 2018). Under this model, the difficulty of producing a word corresponds to its phonological information content $PIC(d^w)$ of a word form d^w is defined as:

$$PIC(d^w) = -\log P(d^w) \tag{1}$$

$$= -\log P(l_1, \ldots, l_{|d^w|}) \text{ for } l \in d^w \tag{2}$$

$$= -\sum_{i=1}^{|d^w|} \log P(l_i | l_{i-(n-1)}, \ldots, l_{i-1}). \tag{3}$$

where l are the phonemes that comprise the sequence d^w, $|d^w|$ is the length (in phonemes) of d_w, and n is the sequence length, or order, of the model (*e.g.,* 1 = unigram, 2 = bigram, 3 = trigram). Here we estimate these phonotactic probability by fitting an n-gram model on a list of unique word types in the language. This model thus reflects two consequential decisions: first, to treat words as sequences of phonemes rather than considering more complex hierarchical generative process that consider their sub-word constituents, and second to consider the probabilities of phone sequences under the inventory of unique word forms in a lexicon (*i.e.,* a type-weighted model) rather than the pure frequencies of these sequences in a sample of speech.

Regarding the first decision, a sequence model over phonemes constitute a simplistic

[2]Correspondingly, our usage of the word "phonotactic" refers to processes and graded knowledge at all three levels, rather than the more limited usage regarding representational knowledge that specifies which word forms are well-formed vs. not in a given language, as the term is used in Romani et al. (2017).

measure of phonotactic probability in that it does not explicitly posit sub-word structure. We justify this decision by noting several points. First, n-gram models partially capture higher order structure implicitly, for example that the prevalence of sub-word units is reflected in phoneme transitions. Second, the correct choice of sub-word unit in a hierarchical model is unclear (*i.e.,* morphemes or syllables), and in either case only a subset of languages have large-scale resources to characterize these sub-word segments (*e.g.,* CELEX; Baayen et al., 1995), which would limit the number of languages in our sample (see also Mahowald et al., 2018, who offer a similar justification). Finally, neuroimaging work suggests that phonotactic constraints may emerge from lexical regularities without strong constraints on intermediate representations of sound patterns (Gow et al., 2021,0). We return to the these issues in more detail in the Discussion, particularly in limitations in modeling the recognition of nonwords.

We motivate the second decision—to fit the model on type lists—both theoretically and pragmatically. Our theoretical justification for using the type-weighted model is that such a model reflects that speakers' generalization behaviors are highly sensitive to the composition of lexical entries in the lexicon, moreso than their token frequencies (Pierrehumbert, 2003). Modeling work in the "adaptor grammar" framework offers convergent evidence that both tokens and types have effects on speakers' generalizations of linguistic structure, including at the level of words (Goldwater et al., 2005). Finally, experiments with children also suggests that early phonotactic knowledge reflects the prevalence of patterns across word types (Richtsmeier et al., 2011). Our pragmatic justification is that it helps us avoid a circularity: by building a model of phoneme transition probabilities over unique word types in the lexicon and using only short sequences we avoid the obvious circular relationship between word frequency and token-weighted phonotactic probability (which would necessarily be equivalent).

2.1.1 Phonotactic Probability and Word Length. The above measure of phonotactic probability can be compared with word length to provide a more motivated

understanding of their relationship.[3] The length of a word can be seen as a measure of probability under a highly simplified generative phonotactic model, specifically one that treats phoneme string generation as the result of a memoryless, uniform random process where all phonemes are equiprobable. This kind of random process has long been used as a null hypothesis in statistical language research, often has been characterize as the random typing of "monkeys on typewriters" (Mandelbrot, 1954; Miller, 1957). While notably deficient in capturing the key aspects of human-generated linguistic samples (Ferrer-i-Cancho and Solé, 2002), this simplified model captures the key correspondence that longer word forms are less probable in a language: as long as there is more than one phoneme in the language–such that the probability any phoneme is less than 1– then a word form that is one phoneme longer is necessarily less probable.

Specifically, if a word form d^w is comprised of a string of symbols $d_1^w, ..., d_n^w$ that are equiprobable and independent—$P(d_i^w) = 1/|V|$, where $|V|$ is the number of items in the relevant (phonemic or orthographic) inventory—then PIC under this naive phonotactic model is strictly proportional to the length of the word form:

$$-\log P(d^w) = -\log P(d_1^w) \times ... \times P(d_n^w) \tag{4}$$

$$-\log P(d^w) = -\log P(1/|V|)^n \tag{5}$$

$$-\log P(d^w) = n \times -\log P(1/|V|) \tag{6}$$

Finally because $-\log P(1/|V|)$ can be factored out as a constant scaling factor,

$$-\log P(d^w) \propto n \tag{7}$$

$$PIC(d^w) \propto n \tag{8}$$

This result establishes a correspondence between length and phoneme sequence probability, and implies that Zipf's original formulation may be a special case of a more general

[3] See also Frisch et al., 2000, who investigated the relationship between word length and implicit measures of phonotactic probability in nonwords.)

Table 1

Estimates of the phonological information content (in bits) under a naive model vs. under a probabilistic phonological model for English

"M"	"O"	"T"	"O"	...	Σ Bits
Uniform Character Probabilities					
$-\log_2 P(M)$	$-\log_2 P(O)$	$-\log_2 P(T)$	$-\log_2 P(O)$		
4.700	4.700	4.700	4.700	...	47.004
3-Character Phonological Information Content Model					
$-\log_2 P(M \mid \triangleright)$	$-\log_2 P(O \mid \triangleright M)$	$-\log_2 P(T \mid MO)$	$-\log_2 P(O \mid OT)$		
4.400	2.184	3.217	3.672	...	35.602

Note The negative log probability of a word form under the uniform character probability model is computed assuming 27 characters in the inventory and an end symbol for the word form. The probabilistic language model takes into account sequential dependencies up to length 3, obtained from 25,000 most frequent words in the English Google Books (2012) corpus.

relationship between frequency and phonotactic probability. Phonotactic probability computed under the more sophisticated n-gram model produces some predictions contrary to the length-only model. Depth /dɛpθ/ (*depth*) contains fewer phonemes yet has a higher PIC (28.7 bits) than /gɹɑʊnd/ (ground, 12.89 bits) because the latter is comprised of significantly more probable subsequences. A comparison of phonotactic probability (PIC) estimates from the monkeys-on-typewriters model and a more sophisticated trigram model for the word *motorcycle* is presented in Table 1.

2.2 A Model of Robust Word Recognition

Our second model concerns understanding the role of word form distinctiveness in spoken word recognition. Particularly, we offer a model-based measure of the degree of competition for a given word form using a Bayesian model of spoken word recognition in the vein of Norris and McQueen (2008). This model relates the probability that a listener assigns to the speaker's intended word form to the frequency and the specific word form of a target word, as well as the the frequencies and word forms of competing words in the

lexicon. Critically, this formulation captures the fact that listeners are known to rely on prior probabilities of possible interpretations to infer speaker's intended meanings (Gibson et al., 2013). Under the model, an idealized listener evaluates support for the correct word interpretation w^* of the received phonemic[4] data d, by computing:

$$P(w^*|\ d) = \frac{P(d|w^*)P(w^*)}{\sum_{w' \in V} P(d|w')P(w')}, \tag{9}$$

where $p(d|w^*)$ is the probability that the speaker's intended word w^* would generate the observed data d (the likelihood), $p(w^*)$ is the prior probability of the word, and the denominator reflects the aggregate strength of all possible words in the lexicon V (reflecting both their respective fit to data and their prior probabilities). From this we can thus derive the posterior odds ratio of the listener recovering the speaker's intended word (w^*) vs. recovering some other word ($P(\neg w^*|\ d)$):

$$\frac{P(w^*|\ d)}{P(\neg w^*|\ d)} = \frac{P(d|w^*)P(w^*)}{\sum_{w' \in V, w' \neq w^*} P(d|w')P(w')}. \tag{10}$$

We define the denominator on the right hand side, $\sum_{w' \neg w^*} P(d|w')P(w')$, as a special quantity that we call *aggregate competitor probability*, as it indexes the number of strength of competitor words that could have generated the observed data. This quantity is closely related to the marginal probability of the data, $p(d)$, except that it excludes the contribution of the intended word (*i.e.,* , it only considers competitors). By considering the relationship of a target word's frequency, the number and strength of similar word forms, and bounds on the posterior odds ratio, we can establish an intuitive relationship between the distinctiveness of a word form (as reflected in aggregate competitor probability) and its prior probability of use. The probability that the true word w^* would generate the observed data d, $p(d|w^*)$, can be considered as a constant, and the likelihood that any

[4]In principle, acoustic or even narrowly-transcribed phonetic data would be better, but our available datasets are limited to phonemic transcriptions.

competitor word w' would generate the observed data d, $p(d|w')$, is equal to or less than that constant. We then consider the relationship between the numerator and the denominator from Eq. 10 for a high frequency word w_1 and low frequency word w_2: To have the same posterior odds ratio in spoken word recognition, the lower prior probability word w_2 *must* have a smaller aggregate competitor probability than the higher probability word w_1. This requires that w_2 has a word form d^w which is more *diagnostic*: that there are few competing forms w' in the lexicon that have a high likelihood of generating the form d^w. If $p(d|w')$ is too high for too many competing word forms, the posterior probability of the correct interpretation for w_1 would fall below some threshold, making it impossible for listeners to correctly identify it, and it would, in theory, exit the language (as it would be identified as another word or construction). By contrast, the high frequency word w_2 can still be recognized at the same rate when there are more / higher frequency similar words, by virtue of its higher prior probability $P(w_2)$. The abstract relationship between likelihood, prior, aggregate probability of competitors, and the posterior probability of the correct hypothesis regarding word identity hold regardless of specific form of prior and likelihood, but for concreteness we now operationalize these terms.

2.2.1 Operationalizing the Prior. Our primary analyses here operationalize the prior probability of the target word, $P(w^*)$, as the normalized frequency of the speaker's intended word in a large corpus; we explain how we use word frequencies from specific corpora in the Methods. We also consider an alternative of in-context predictability, $P(w^*|c)$, that has been treated in detail in previous work. Questioning the relevant aspect of word predictability in predicting word forms, Piantadosi et al. (2011) demonstrated that word length is better predicted by average in-context probability (*i.e.,* predictability under an n-gram model at the level of words) rather than frequency alone across the lexicons of eleven European languages. This empirical pattern is consistent with the theory of Uniform Information Density (UID) that speakers modulate the information rate of their utterances to maximize information transfer without exceeding channel capacity (Levy and

Jaeger, 2007; Jaeger, 2010). Meylan and Griffiths (2021) and Levshina (2022) both suggest limitations in the corpus analyses in Piantadosi et al. (2011) and find that frequency is the better predictor of word length in many language samples. On the other hand, Mahowald et al. (2013) obtain consistent experimental results that speakers choose long or short variants of words (*e.g., rhino* vs. *rhinoceros*; *info* vs. *information*) as a function of in-context predictability, as predicted by UID. Seyfarth (2014) also found that speakers moderate the acoustic duration of individual realizations of a word form according to their contextual predictability, also consistent with the predictions of UID. This leaves open the possibility that a word's predictability in longer preceding contexts helps to determine its word form. We thus test the correlations between average in-context predictability of words and the phonotactic probability of their corresponding wordforms.

2.2.2 Operationalizing the Likelihood. The likelihood, $P(d|w^*)$ can be realized as any function that yields the highest value for data d^* corresponding to w^* (up to 1) and decreases with the dissimilarity of the expected phonemic form for the word and the observed data. A common choice in noisy channel models of language acquisition is "negative exponentiated edit distance" (*e.g.,* Levy, 2008b) where zero edits results yields a likelihood of 1 (e^0) and an increasing number of edits yields an exponentially decreasing probability.

This simple edit-distance based likelihood has the drawback of equating the probability of all insertions, deletions, and substitutions, when in fact speakers are much more likely to delete phonemes than add them, and that certain phoneme substitutions are much more likely than others (/p/ is likely to be confused with /b/, but not with /m/). A more powerful parameterization is thus the exponentiated path weight under a weighted finite state transducer (WFST; Mohri et al., 2002). This can be seen as a generalization of edit distance, where the cost of edits, deletions, and substitutions may all be different for each phoneme (or pair of phonemes, in the case of substitution). In the study presented here, FST weights are set using phoneme confusion probabilities obtained from a lab-based

experiment with adults identifying phonemes in noise, as well as some free parameters (see Study 2: Methods).

2.2.3 Neighborhood Density. The likelihood function presented above is closely relate to a word's neighborhood density, or the number of words with similar word forms, which is known to have an effect on word recognition (Coltheart et al., 1977; Luce and Pisoni, 1998; Suárez et al., 2011). While proposals vary on how to best operationalize the definition of similarity, proposals share the intuition that words with more similar word forms ("neighbors") are harder to recognize because there are more competitors consistent with a given received acoustic signal (or visual form in the case of reading; here we focus on the case of spoken word recognition). The most common definition of a "neighbor" is a word within an edit distance of one phoneme or letter (one substitution, deletion, or insertion) (Coltheart et al., 1977). Neighbors defined in this way have a special status under the Bayesian spoken word recognition model above: under the exponentiated negative edit distance likelihood above, they represent the set of highest likelihood competitors. Intuitively, the number of such competitors is higher for short sequences; however, competition dynamics may be more complex when word recognition is considered as a realtime process that unfolds within a word (*e.g.,* Luce and Pisoni, 1998 Study 1; Strand and Liben-Nowell, 2016).

Previous work (Vitevitch et al., 1999; Dautriche et al., 2017; Mahowald et al., 2018), has demonstrated a strong correlation between phonotactic probability and neighborhood density. Against this basic pattern, neighborhood density and phonotactic probability are known to decouple in certain ways. First, it is possible to find words with high phonotactic probability but no neighbors within a single edit for words of moderate length. Previous work has used this distinction to investigate differences between listeners' phonological and word-level expectations in word recognition (Storkel et al., 2006). Second, cross-linguistic work suggests another dissociation in that neighborhood densities may be on average higher in natural languages than expected under a lexicon-wide phonotactic model

(Dautriche et al., 2017). In the current work, we investigate the degree to which metrics of neighborhood density track the proposed quantity of aggregate competitor probability under the Bayesian spoken word recognition model.

2.3 From Cognitive Processes to Language Change

The above models help characterize word forms along two gradients: how easily they are produced by speakers, and how susceptible to mishearing they are by listeners. In the current work we refrain from providing an explicit process model for language change because lexicons undoubtedly reflect a variety of other other concerns (transparency of morphological paradigms, exogenous social factors related to language contact, interactions of existing items in the lexicon with lexicalization processes by which multi-word combinations attain the status of single words). Though we don't provide such a full account, we offer a sketch of how these two competing pressures could interact to influence lexicons over time: First, a speaker wanting to transmit a message (a specific word) considers alternative word forms, each of which has an associated production cost and a posterior probability of recovery by an idealized listener. The speaker chooses the word form with the lowest production cost which can still be recognized by the listener as the intended word with some error rate (this could vary by word; some words may have higher and lower thresholds, but the error rates are in some way constrained for each word). All speakers in a language choose word forms in this fashion, seeking to reduce their articulation costs while remaining intelligible to listeners given the lexicon and their understanding of the use frequencies of words at that point in time. This moves the language towards the shortest possible word forms that are still intelligible to listeners — though of course such strong optimization is unlikely given the other pressures noted above that are likely to influence languages.

3 Study 1: Phonotactic Probability and Frequency

Study 1 tests the hypothesis that the phonotactic properties of word forms reflects variation in frequency above and beyond that captured by word length. To approach this problem empirically, we examine the strength of the relationship between word frequency and phonotactic probability using the speaker model above across a wide range of languages and datasets. If indeed phonotactic probability captures fine-grained changes in word reduction on the basis of frequency, we expect it to outperform word length as a correlate of word frequency. The null hypothesis is that the correlation between frequency and phonotactic probability is no stronger than the correspondence between frequency and word length. We evaluate this hypothesis across 13 languages drawn from three large-scale corpora.

3.1 Methods

To provide a brief overview of this process: we compute new frequency estimates for web-scale corpora in 13 languages and three datasets. For each corpus, we take the top 25,000 most frequent words and construct a type-weighted phonotactic probability model following the reasoning in "A Model of Speaker Effort," above. We then use the model for each language to produce model-based information content estimates — negative log probability of each word form under the phonotactic model (see Estimating Phonotactic Probability) – for those 25,000 highest-frequency word forms. We then evaluate the correspondence between frequency and phonological information content across the words and corresponding word forms in each corpus.

3.1.1 Datasets. The Google Web 1T datasets were downloaded from the Linguistic Data Consortium (Brants and Franz, 2006,0); the Google Books 2012 datasets were downloaded from **storage.googleapis.com/books/ngrams/books/datasetsv2.html** (Michel et al., 2011), and OPUS (2013) from **opensubtitles.org** (Tiedemann, 2012). All

punctuation-only word tokens were discarded, and punctuation marks appearing with other text, with the exception of apostrophes, were removed. We make the simplifying assumption that the tokenized written forms correspond to lexical items used by speakers; while this assumption may not hold for all forms (*e.g.,* German compound nouns, French contractions), it holds for the vast majority of word forms in the analysis (see Baayen et al., 2016 for further discussion of the importance of variation in orthographic segmentation conventions for frequency analyses). Following best practices for web-based corpus analysis proposed in Meylan and Griffiths (2021), all tokens were converted to lowercase using the relevant POSIX locale; US English and European Portuguese were used for English and Portuguese, respectively. In the case of Google Books 2012, part-of-speech tags were discarded, and instances from earlier than 1800 removed from the analysis. UTF-8 encoding was maintained throughout for all languages and datasets; Hebrew strings were represented with right-normalized forms.

3.1.2 Estimating Phonological Information Content. For each language and dataset, a three-character transition model (see "A Model of Speaker Effort") was estimated using the 25,000 most frequent in-dictionary words also appearing in the corresponding OPUS subtitle corpus (following Piantadosi et al., 2011). Diphthongs (vowel sequences) were treated as sequential discrete vowels. We also produced analogous three-phone transition models (excluding the Hebrew OPUS and Hebrew Google Books 2012 datasets) using the International Phonetic Alphabet (IPA) transcriptions from an automatic speech synthesizer, eSpeak. While these broad phonological transcripts are imperfect, using IPA representations for words accounts for language-specific variation in orthographic conventions. For example, written Spanish includes accents only when the placement of prosodic stress cannot be deduced from more general rules in the language. Using an IPA transcription avoids the need for developing language-specific processing rules, for example deciding whether 'a' vs. 'á' should be merged or kept as separate orthographic variants in Spanish. Diphthongs were treated as multiple distinct phonemes.

Loan words and acronyms can greatly affect the obtained transition probabilities for the phonotactic model, especially because types contribute equally to the transition weighting (*e.g.,* the transitions in "Okeechobee," "mañana," and "ACLU" would be as heavily weighted in a phonotactic model of English as the transitions in "they" and "will"). To minimize these effects, we used only non-capitalized word types present in the relevant Aspell dictionary to build sound and character transition models for each language (with the exception of German, in which nouns, which are capitalized by convention, were retained).

In that phonological inventories are much smaller than lexicons, sparsity is less problematic for phoneme-level language models than for word-level ones. Nonetheless, to avoid overfitting among higher order sequences, phone and character transition probabilities were computed with Witten-Bell smoothing (Chen and Goodman, 1999) with interpolation on transitions of order 3 using the SRILM toolkit (Stolcke, 2002), as is commonly used for character-level language models. Finally, some amount of probability mass must be assigned to unseen phonemes, in case they are encountered in the test set. While this is not a concern for the core phonemes that comprise a phonological inventory, loanwords may contain singleton phonemes (e.g. the final vowel in "rapprochement" in English). Here we map these out-of-vocabulary phonemes to an unknown phoneme token, which is assigned a small probability mass; this unknown token is then treated as any other by the smoothing scheme.

Each word's phonotactic probability was calculated as the product of the probabilities of each symbol given the preceding symbol string up to two symbols, including a start symbol $\triangleright$ and an end symbol $\triangleleft$,
e.g., $P(the) = P(t|\triangleright) \times P(h|\triangleright t) \times P(e|th) \times P(\triangleleft|he)$. We convert this sequence probability to phonological information content—which keeps the same directionality and approximate range of word length—by taking the negative logarithm.

3.1.3 Evaluating Correlations. Following the basic methodology adopted in Piantadosi et al. (2011), we examine the correlation between log frequency and each of two quantitative measures of structural form (either word length or PIC) for the 25,000 most frequent types in each language. Unlike that work, we limit our analysis to in-dictionary types, thereby excluding person names, place names, acronyms, and loan words from the analysis (Meylan and Griffiths, 2021).

We evaluate the strength of each of these correlations using Spearman's rank correlation coefficient, which evaluates the degree of monotonicity of the function rather than linear correspondence. Word form measures included the length in phonemes, the length in characters, the phonological information content as estimated under an trigram model of phoneme transitions, and the approximation of phonological information content as estimated under a trigram model on character transitions. The statistical significance of the difference between correlations is evaluated in each case using bootstrapped estimation of the difference scores (orthographic PIC vs. length; phonemic PIC vs. length) and comparing the resulting distribution to 0.

To evaluate the relationship of PIC and frequency in the absence of word length, we also partial out word length (i.e., use the residuals from predicting PIC from word length with a linear model) and compute the correlation with frequency (following Piantadosi et al., 2011); following the recomendation of an anonymous reviewer, we also present an analysis where we partial word length out of both PIC and frequency and compute the correlation among the yielded residuals. We perform the analogous operation on length, and compute the residuals when predicting length from PIC, and test the correlation with word frequency. This provides a strong test of the unique predictive value of each of these variables.

Finally, we compute correlations between frequency and PIC for three random baselines for each language in Google 1T to confirm that the obtained correlations are nontrivial. In the first set of baseline models, the assignment of word form to frequency is

randomly permuted in each baseline lexicon, such that a short word like "the" could have a rare and unlikely wordform like "giraffe" (we call this a "Shuffled Lexicon"). In the second set of baseline models, word forms are drawn from the phoneme inventory of the language (without replacement), maintaining each word form's length in the source language but not the inventory of characters for a word ("Drawn with same length"). In the third set of baseline models, the order of phonemes or characters for each word are permuted recomputed ("Shuffled within word form"). All controls maintain the same unigram phoneme statistics, the second two baselines perturb the higher-order phoneme transitions in the natural language language. For this reason, we refit the relevant phonotactic model on each sample and recompute PIC over all word forms in the sample. We expect the correspondence in the actual language to exceed all three of these baselines.

Models and analyses can all be accessed through OSF: `https://osf.io/wza8k/?view_only=7ef28c51ceba46f88835e5a9bc3a3851`.

3.2 Results

We investigate the correlation between word frequency and two different measures of speaker effort: word length and phonotactic probability in a sample of large corpora (43m to 266b words) in 13 languages, across three large-scale datasets from different linguistic sources. If a word's phonotactic probability is indeed a stronger correlate of frequency (technically, that phonological information content is a stronger negative correlate of log word frequency), then we may conclude that speaker effort is better reflected in the phonotactic probability of word forms than their length alone, providing evidence of an important generalization of Zipf's Law of Abbreviation.

3.2.1 Phonotactic Probability vs. Word Length. Across languages, we obtain a systematically stronger negative correlation between log frequency and PIC than log frequency and word length (Figure 1). Building the model from phonemic transcriptions, this pattern holds in 11 of 11 languages in the Google 1T datasets, 4 of 6

languages from Google Books 2012, and 12 of 12 languages from the 2013 OPUS corpus (phonemic transcriptions were not available for Hebrew). Building the model from characters—as an approximation of phonological forms—this pattern holds in all languages from Google 1T, 4 of 7 (results are consistent in 2 of the remaining languages, but fail to reach significance), and all languages from the 2013 OPUS corpus. Partial correlations reveal a significant length-related contribution to PIC and vice versa, though in some cases PIC with length partialed out is a stronger correlate of frequency than length is, *e.g.,* English 1T when PIC is computed over IPA representations.[5] In other words, among words of the same length in a given language, PIC explains substantial additional variance in word frequency (Figure 2): high frequency words contain higher phonotactic probability (lower PIC) sound sequences.

[5]While we follow the partial correlation procedure in Piantadosi et al. (2011), we thank an anonymous reviewer for noting some of the shortcomings with this approach for computing partial correlations (Wurm and Fisicaro, 2014). We provide an auxiliary set of analyses looking at the correlation between word frequency and PIC after residualizing both measures with respect to word length in the Supplementary Information.

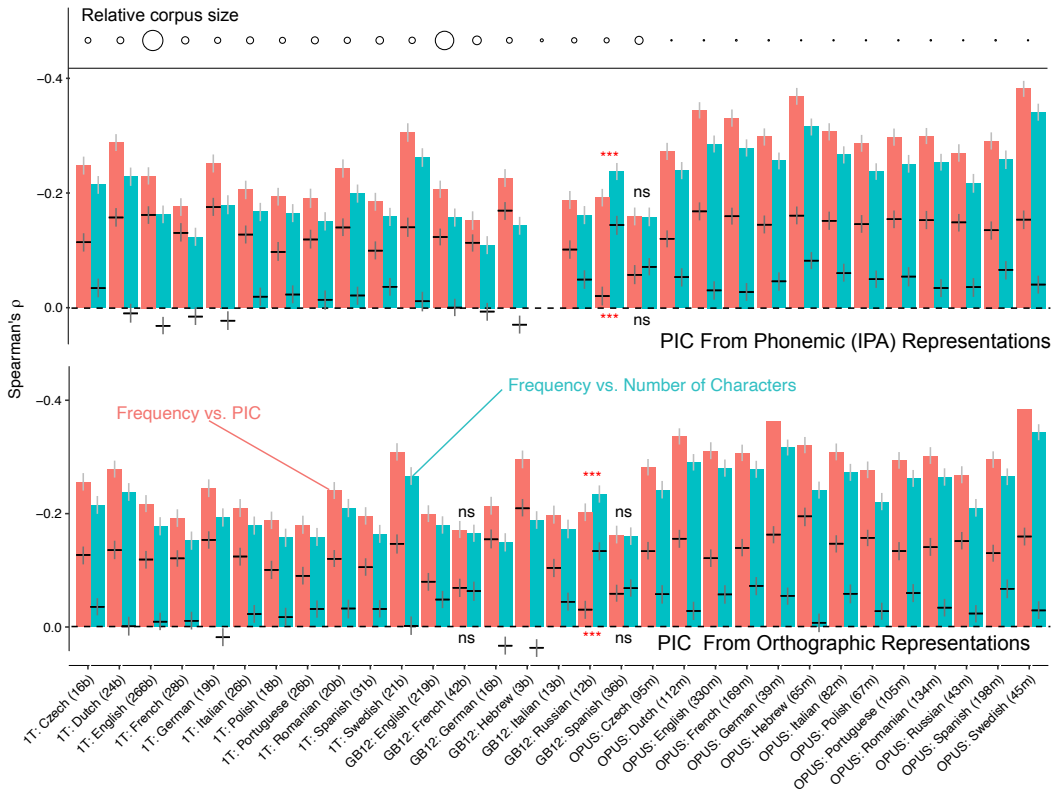


Figure 1. Spearman's ρ between frequency and phonological information content (red) and frequency and number of characters (green). Frequency exhibits a stronger negative correlation with PIC than length for the $n = 25,000$ most frequent words in each dataset. Gray lines indicate the 99% bootstrapped confidence interval. Black horizontal lines indicate the correlation with the other measure of word difficulty partialed out, and are show with corresponding 99% bootstrapped confidence intervals. All comparisons are statistically significant ($p < .01$) unless otherwise noted: "ns" indicates non-significant contrasts and contrasts that are significant in the opposite direction to that predicted (Russian in both datasets) are marked in red.

A similar pattern of results emerges regardless of whether PIC is computed over characters or phonemic representations. The Russian Google Books 2012 dataset is the only dataset showing consistent evidence in favor of a stronger relationship between length and frequency—however, this is contrary to the results of the Russian OPUS results, which reflect the prevailing dominance of PIC. Looking across datasets, Russian shows the lowest

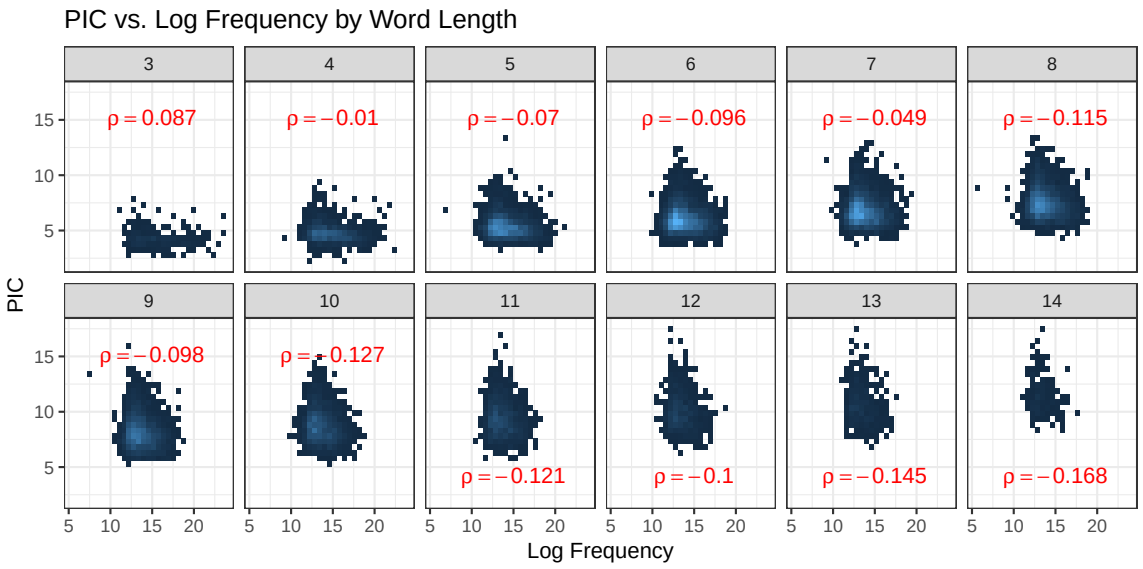


Figure 2. Among words with the same number sounds, more frequent words have higher phonotactic probability as measured with PIC. This figure shows this relationship in the English Google Books (2012) dataset. Correlations display Spearman's ρ among words of the same length in phonemes.

correlation between frequencies between Google Books and OPUS (Pearson's $r = .48$), as well as the lowest correlation between PIC estimates derived from Google Books and those derived from OPUS (Pearson's $r = .63$). Across languages, models built over phonemes and character transitions provide similar estimates of PIC (Pearson's r between .789 and .919 across languages, median = .874). While more research is required to extend these findings beyond Germanic, Romance, and Slavic languages, Hebrew provides an important first test of whether this relationship holds in the lexicon of a language from Afroasiatic language family.

PIC-Frequency correlations computed from three-phone and three-character transition models from natural languages substantially exceed the correlations observed for PIC computed under three random baseline languages (Figure 3). We find that the correlation obtained for natural language are larger than all same-language baselines for every language ($p < .001$, by bootstrapped tests of the difference of correlations between

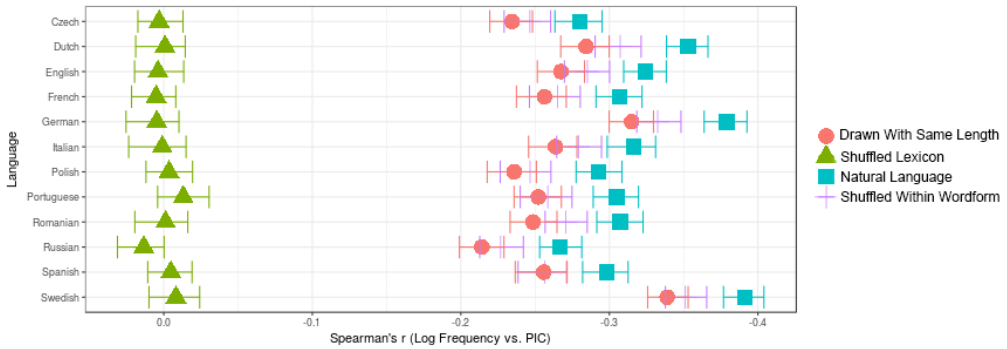


Figure 3. PIC computed from three-character transition models on the lexicons of natural languages ("Natural Language") have a significantly higher correlation with frequency than three baselines: 1) when PIC is computed for randomly shuffled word forms ("Shuffled Lexicon") 2) when the word form is drawn using single-character probabilities ("Drawn with same length") and 3) when the order of characters is shuffled within each word form ("Shuffled Within Word form").

each <baseline, natural language> pair). This shows that alternative possible lexicons do not have the same composition aas natural languages and suggest that the correlations obtained here are nontrivial.

4 Study 2: Phonotactic Probability and Word Recognition

Study 1 yields empirical results that corroborate the stronger correlation between word frequency and word form phonotactic probability (vs. word form length) found by Mahowald et al. (2018). In Study 2, we revisit the explanation offered by Mahowald et al. (2018) for this correlation and offer an alternative explanation following the arguments in "A Model of Roubst Word Recognition" above. In contrast to the claims in Mahowald et al. (2018), our model suggests that high phonotactic probability word forms should be *harder* to recognize because they induce more competition effects from alternative interpretations, and this competition must be proportional to the prior probability of the target word in order for it to be reliably recognized. As an empirical test, we look at the correlation between phonotactic probability and the above term of "aggregate competitor probability"

(ACP), which indexes the degree of competition from all other word interpretations in the language under a Bayesian model of spoken word recognition. We conduct this analysis only for English given the limitations in available data for parameterizing the likelihood in the Bayesian model of spoken word recognition, as detailed below.

To connect with a rich previous literature on the effects of phonemic neighborhood density on spoken word recognition, we additionally compare aggregate competitor probability to two measures of neighborhood density that have been previously in characterizing spoken word recognition: the number of phonemic (Coltheart et al., 1977) neighbors, and average Levenshtein distance to the 20 nearest phonetic neighbors (PLD20, Suárez et al., 2011).

4.1 Methods

To estimate $p(w^*)$ and $p(w^{'})$ under the Bayesian spoken word recognition model proposed in the introduction, we take normalized unigram probabilities from the Google Books corpus (Michel et al., 2011). Likelihood terms, $p(d|w^*)$ and $p(d|w^{'})$, are computed using a weighted finite-state string transducer. We motivate this in relation to a commonly-used likelihood in visual and spoken word recognition of exponentiated negative edit distance (Levy, 2008b):

$$P(d|w) \propto e^{-\text{dist}(d^*:w,\ d)} \tag{11}$$

where $\text{dist}(d^* : w,\ d)$ is the edit distance (or Levenshtein distance, or the minimal number of deletions, insertions and substitutions) between citation form d^* for candidate word w, designated here $(d^* : w)$, and the received phonetic data d. In the case of the target word, the received phonetic data d is the citation form for the target word d^*, yielding a probability of 1 (e^0).

This treatment of edit distance fails to take into account perceptual similarity

between phonemes: some phonemes are more likely not to be pronounced or perceived by the listener, and some are more likely to be confused with one another (*e.g.,* /f/ and /θ/, per Cutler et al., 2004). We account for this by using a weighted finite state string transducer (WFST; Mohri et al., 2002) which encodes variable edit costs, and parameterize the model using estimates of phoneme confusion in noise from a lab-based recognition experiment (Weber and Smits, 2003). For simplicity, we conduct this analysis only in English; to our knowledge, the only other language for which appropriate datasets exist is Dutch, which is sufficiently close in the phylogenetic tree of languages that demonstrating that it holds there would provide little additional evidence of its generality. Probabilities are treated as positive valued costs by taking the negative log transformation.

Following Meylan et al. (2023), we estimate $P(w^*|d)$ in a similar fashion to exponentiated negative edit distance,

$$P(d|w^*) \propto e^{-\ \theta_{\text{WFST}}(d^*:w^*,\ d)}. \tag{12}$$

While Levenshtein distance relies on a single minimum edit distance between the input and output string, the WFST captures the aggregate weight of paths through a finite state machine, $\theta_{\text{WFST}}(d^* : w^*,\ d)$, which captures the many possible edit sequences that would transform the expected form of a word into the received form. We create a finite state machine, $\text{FSM}(d^*, d_n)$, for each combination of the citation form for the target word d^* (from the CMU pronunciation dictionary) with the 10,000 highest frequency words that form the candidate vocabulary (d_n).[6] Computationally, this involves composing a finite state acceptor for the conventional form d_n for each possible word identity with the transducer learned above, and then composing that state machine with the finite state acceptor for the citation phonetic form of the target word, d^*. We then enumerate the weights along all paths through the resulting finite state machine that represent possible

[6]We limit the size of the vocabulary, $|V|$, to 10,000 words for computational tractable, as we need to compose $|V|^2$ state machines over the entire vocabulary.

sequences of edits from the input to output string,

$$\theta_{\text{WFST}}(d^* : w,\ d) = -1 \times \log \left(\sum_{\text{path} \in \text{FSM}(d^*, d)} P(\text{path}) \right). \tag{13}$$

Following the chain rule, each path reflects the product of the conditional probabilities of its component arcs:

$$P(\text{path}) = \prod_{\text{arc} \in \text{path}} P_{\text{FSM}(d^*, d)}(\text{arc}). \tag{14}$$

The phoneme confusion data in Weber and Smits (2003) lacks some critical information for parameterizing the WFST (*e.g.,* the probability of phoneme insertion), and other values may deviate substantially between lab experiments and fluent spoken speech (*e.g.,* the probabilities of phoneme deletions). For these reasons, we use three free parameters to translate from the substitution probabilities in lab experiments to edit probabilities in our likelihood model. The lab-based phoneme confusion data do not contain probabilities of inserting phonemic material, so we leave this as a free parameter ("insertion cost"). This parameter should be expected to be relatively high (assigning low insertion probability) as it is unlikely that listeners think that a short word generated a long word form (the opposite of prevalent reduction processes). Lab-based data also most likely underestimate the rate at which participants fail to detect phonemes in fluent speech, so we include a free parameter ("deletion scaling") that linearly scales the probability of deleting a phoneme (we distribute the remaining probability mass proportional to the original phoneme-to-phoneme transitions). Finally, we experiment with fractional additive smoothing to the phoneme/phoneme edit probabilities, in case real-world substitutions are less peaked than those observed under lab conditions. The value of this parameter indicates the probability mass that we take from the lab-based distribution and spread equally across all other phonemes of the same class (consonants vs. vowels). We find best

values for these three parameters by grid sampling to maximize the correlation between aggregate competitor probability and phonotactic probability. We report the best results below as well as the results across the parameter range, and report the average posterior probability that the model places on the speaker's intended word.

All data, models, and analyses can all be accessed through OSF: `https://osf.io/wza8k/?view_only=7ef28c51ceba46f88835e5a9bc3a3851`.

4.2 Results

In Study 2, we test if the measure of phonotactic probability of a word form from Study 1—phonotactic information content—correlates with the number and strength of competitor words in the lexicon under an idealized Bayesian model of spoken word recognition. If there is a strong positive correlation, it means that, consistent with our hypothesis outlined in the introduction, speaker articulatory cost for word forms trades off with the number and strength of competing interpretations. The proposal in Mahowald et al. (2018) predicts a negative correlation, as the strength of competitors would need to decrease with phonotactic probability in order for the recognition of high phonotactic probability to be easier, as they claim.

The obtained measure of the aggregate strength of competitors under a Bayesian model of spoken word recognition is strongly correlated with phonotactic probability, consistent with our proposal (Spearman's $\rho = .842$, 95% bootstrapped CI = [.835, .848]; Fig. 4).[7] This points to a fundamental constraint on the lexicon: higher phonotactic probability comes at the expense of stronger competition in spoken word recognition from alternative words in the lexicon. Only frequent forms are able to sustain such competition, because listeners are able to use their prior expectations to recognize words with less

[7]This correlation is robust across the range of free parameters in the likelihood function; the lowest obtained correlation was .68. The results reported above reflect a smoothing parameter of .1, an insertion cost of 7.0 and a deletion cost of 1.0 and yield an average surprisal of 1.12 bits. If we instead optimize to maximize posterior probability of the target words, the best model has an average surprisal of .41 bits. In this case the correlation between phonotactic probability and ACP decreases a small amount to .82. The results of the grid search are presented in the SI.

distinctive wordforms.

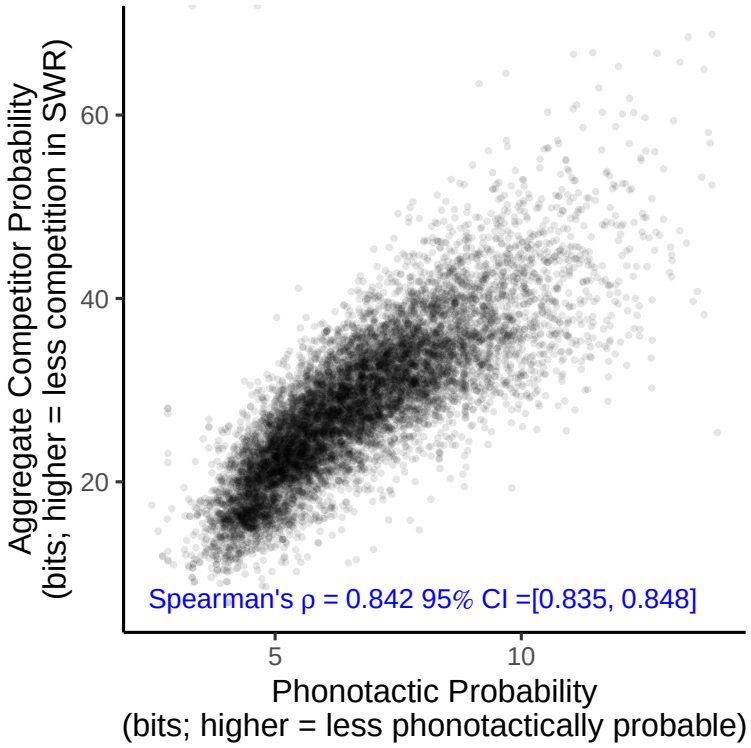


Figure 4. Phonotactic probability vs. the aggregate probability of competitors under a Bayesian model of spoken word recognition for n=10,000 most frequent words in the English lexicon (each point represents a word).

4.2.1 Neighborhood Density. How does the new measure here of aggregate competitor probability relate to neighborhood density metrics previously used in spoken word recognition? Aggregate competitor probability shows a strong positive correlation with PLD-20 (Spearman's ρ = .901; 95% bootstrapped CI = [.896, .905]; Fig. 5). The correlation with the number of phonemic neighbors (Coltheart's N; Coltheart et al., 1977) is nearly as strong (Spearman's ρ = -8.43, 95% CI = [-.85, -.835]; Fig. 6). This correlation is striking given that the measures of neighborhood density do not have access to the frequency of the competitors.

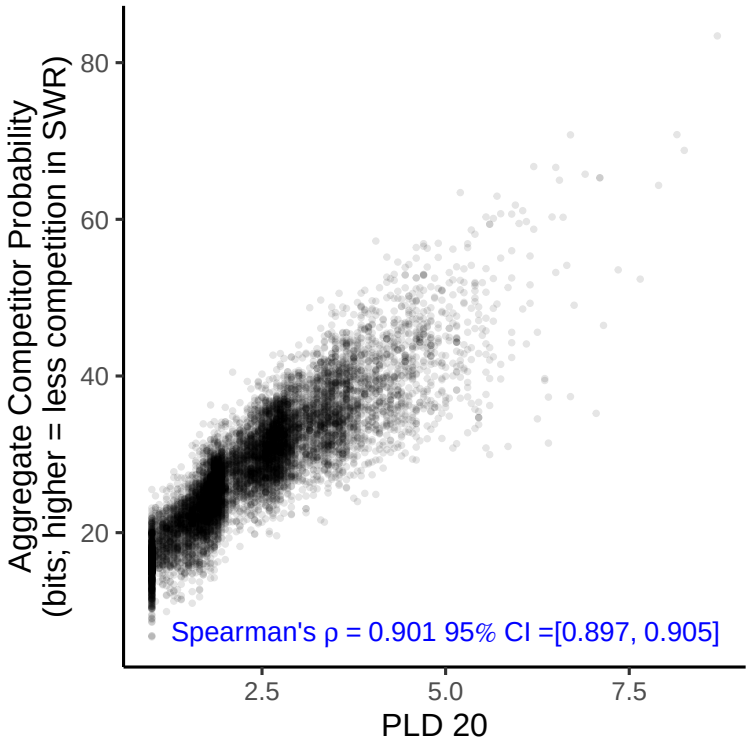


Figure 5. Aggregate probability of competitors under a Bayesian model of spoken word recognition vs. Phonotactic Levenshtein Distance-20 (PLD-20 Suárez et al., 2011) for n=10,000 most frequent words in the English lexicon (each point represents a word). This shows that a more sophisticated neighborhood density metric tracks a key quantity in an idealized mathematical model of spoken word recognition.

5 Discussion

The relationship between word length and frequency (Zipf's Law of Abbreviation; Zipf, 1935) is one of the most robust empirical findings regarding the structure of lexicons. In Study 1, analyses of large-scale corpora from 13 languages across three datasets suggest that word frequency correlates more strongly with phonotactic probability than with word length, suggesting that phonotactic probability better reflects speakers' production. This corroborates empirical findings in (Mahowald et al., 2018). In Study 2, we found in an analysis of the English lexicon that this phonotactic probability systematically trades off with the number and strength of competing words in the lexicon that a listener would need

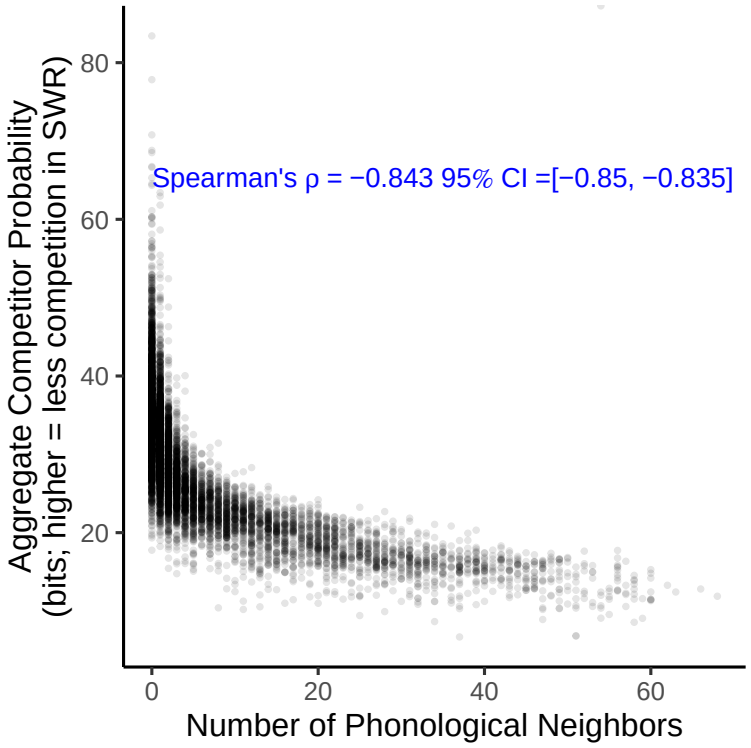


Figure 6. Aggregate probability of competitors under a Bayesian model of spoken word recognition vs. the number of phonological neighbors (Marian et al., 2012) for n=10,000 most frequent words in the English lexicon (each point represents a word). This shows that that even the number of single-edit competitors largely tracks a key quantity of competition in an idealized mathematical model of spoken word recognition.

to distinguish from a target word in the course of spoken word recognition. Taken together, these results suggest a generalization to Zipf's Law of Abbreviation, namely "frequently-used words tend to be phonotactically probable; infrequently-used words tend to be phonotactically improbable so that they can be distinguished the higher frequency words." This pattern in the lexicon is consistent with competing pressures towards speaker- and listener-oriented optimization: phonotactically probable forms are easier to produce, but words must have sufficiently *improbable* word forms in order to be reliably recognized. We now discuss some of the limitations of the current approach and some of the prospects for future research.

5.1 In-context Predictability vs. Frequency

Both of the analyses presented above treat normalized word frequency, or unigram probability, as the relevant feature of words that predict the structure of word forms. An alternative possibility is that average word form predictability—how predictable a word is across the contexts in which it appears—is the stronger determinant of word forms (Piantadosi et al., 2011, but see also Meylan and Griffiths, 2021 and Levshina, 2022). Thus in addition to testing the relationship of frequency and phonotactic probability, we also estimate the overall predictability of each word, operationalized as its average lexical information content, and compute correlations following the same procedure as above. Following Piantadosi et al. (2011), we compute the negative mean log trigram probability across contexts:

$$-\frac{1}{N}\sum_{i=1}^{N}\log P(w|c_i). \tag{15}$$

where c_i is the context for the ith occurrence of w and N is the frequency of w in the dataset. Because estimates of mean information content may be highly biased for small datasets, we do not compute these values for datasets in the OPUS corpus.

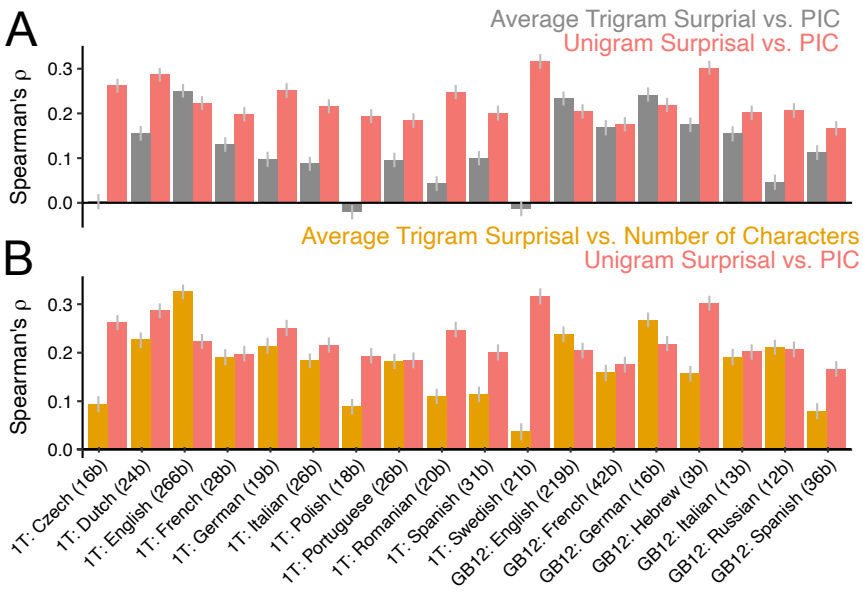


Figure 7. A: Correlations for phonotactic probability and in-context word predictability, treated as average trigram surprisal vs. phonotactic probability and word frequency (unigram surprisal). This shows a stronger relationship between phonotactic probability and word frequency in most datasets. This shows no clear preference for in-context predictability as a better predictor of phonotactic probability. B: A direct comparison of the relationship proposed in Piantadosi et al. (2011) between in-context predictability (average trigram surprisal) and the number of characters, vs. the current work's proposal of unigram probability and phonotactic probability. The mixed pattern of results suggests that languages may not be optimized for frequency or unigram probability (see also, Levshina, 2022).

Whereas Piantadosi et al. (2011) found that taking into account contextual predictability (in the form of mean trigram surprisal) better predicts word length than using frequency (negative log probability), we find that in-context predictability is not a better predictor of phonotactic probability (Fig. 7A). In fact, for all languages except for English in the 1T dataset and English and German in the Google Books corpus, unigram surprisal is a stronger predictor than trigram surprisal for phonotactic probability. This

result corroborates the findings in Meylan and Griffiths (2021) and Levshina (2022) that unigram probability may be a better predictor of word form characteristics than in-context predictability once appropriate controls and data filtering steps are taken into account.

A second question is how the relationship between average information content and word length identified in Piantadosi et al. (2011) compares directly to the relationship between word frequency and phonotactic probability in the current work. This analysis show a mixed pattern of results (Fig. 7B), with alternation across datasets, and even within a single language in the case of German. Variation within a language suggests that datasets may vary in their composition, perhaps according to the language registers that are represented. More research will be required to tease out the relationship between these measures of word form magnitude and expectedness, but we find no clear advantage of in-context predictability.

5.2 Morphology and More Complex Models of Word Forms

As noted above when we introduced a probabilistic model of speaker effort, n-gram models are overly simplistic models of word form probability (Futrell et al., 2017). Further, the commitment to a single-stage model excludes the possibility that neighborhood density effects and phonotactic probability are separable, and can thus make possibly contradictory contributions to different stages of spoken word recognition (Vitevitch and Luce, 1999; Vitevitch et al., 1999). A one-stage model cannot, for example, cannot explain why there might be facilitatory effects of high phonotactic probability in certain tasks, for example the speeded repetition of nonwords (Vitevitch and Luce, 1999, 2005). A two stage-model that includes a separate mapping stage from acoustic input to sub-word units (morphological or otherwise) offers a way for high phonotactic probability to support the recognition of sub-word units, while tasks that involve decisions at the lexical level are impeded by high neighborhood density.

In the current work, we adopted a simpler one-stage model to allow for broad

coverage across languages, albeit at the expense of a complete and accurate model that can account for the breadth of results of word and nonword recognition. However, we briefly describe two alternatives: two more complex probabilistic generative models of word form structure that have been used in linguistics and psycholinguistics that are potentially appropriate for modeling word recognition as a multi-stage process. In principle, both of these model classes can produce better estimates of phonological information content because they learn more abstract regularities in word forms by positing another intermediate level of sub-lexical representation and could be used to model facilitatory effects of phonotactic probability in nonword processing.

A hidden Markov Model (HMM) captures an intermediate level of structure by positing an inventory of unobserved states (*e.g.,* vowels vs. consonants or onsets vs. codas), and conditions observed data on the unobserved state. Besides their ubiquitous use in Natural Language Processing and Automatic Speech Recognition, HMMs have recently been used to examine the extent of gradient lexical competition effects (Strand and Liben-Nowell, 2016).

An alternative approach is to allow for full hierarchical structure in word forms. A probabilistic context-free grammar (PCFG) posits that the observed data (*terminals*) are generated from a set of latent states (*nonterminals*) following a set of probabilistic re-write rules. Unlike an HMM, a PCFG is capable of capturing recursion (though, the utility of such may be limited in the case of word forms). While PCFGs alone are not well-suited for capturing the linear structure of phonemes within a word, an adaptor grammar (Johnson et al., 2006) which allows for stochastic memoization allows such a model to capture word form structure. Futrell et al. (2017) show a modest improvement over n-gram models in predicting phonotactic structure in 14 languages in the WOLEX corpus (Futrell et al., 2017).

While useful for illustrating how more sophisticated generative models of word forms may provide better characterizations of the magnitude of words, we adopt smoothed

n-gram models for the analyses presented here. Futrell et al. (2017) show relatively small advantages for a sophisticated feature-interaction PCFG and Dautriche et al. (2016) show slightly worse performance than higher-order n-gram models. Further, supervised PCFG induction requires that word form data be annotated with non-terminal categories, which are not available for many of the languages in the sample examined here. The unsupervised induction of useful PCFGs — which requires learning in a very large hypothesis space — remains an open problem for research.

5.3 Heteroskedasticity & Historical Change

The analysis here was not intended to directly evaluate the relative utility of speaker- and listener-oriented optimization accounts, unlike previous work in Gahl et al. (2012) or Kanwal et al. (2017). However, the obtained results regarding the relationship between phonotactic probability and frequency reveal an intriguing asymmetry: while high frequency words are necessarily short and phonotactically probable, there are also some low frequency words may also be relatively short and phonotactically probable, e.g. English *ewe*, *gut*, *whey*. This heteroskedastic relationship suggests that the pressure from communicative robustness does not result in an augmentation of words to maintain a correspondence between frequency and word form. This may be due to the fact that speakers have a range of options for reducing a word form, but have limited options for augmenting it, such as epenthesis, or varying the duration (Gahl, 2008). Many of these are forms that have been in the language in their current form for a long time—suggesting the possibility that word forms undergo optimization on longer timescales which cannot be undone. Further work may explore the historical processes that affect word forms and their relationship to use frequency.

5.4 Towards a "Complete" Optimization Account for Lexicons

In the current work we provide a framework for thinking about how speaker and listener pressures trade off in languages, but purposefully stop short of offering a full

cooperative optimization account for two reasons. First, there are many other pressures that operate on word forms besides the opposition described here between intelligibility to the listener and the production cost for the speaker. For example, lexicons may have a preference for transparent morphological paradigms (*i.e.,* where most words follow simple combinatorial rules and there are few irregular cases such as those seen in English past tense verbs) as these may facilitate transmission across generations (Dale and Lupyan, 2012). A second broad family of pressures on lexicons may emerge in scenarios of language contact where competing phonologies come into contact for reasons exogenous to the language. For example, the Norman Conquest of England in 1066 set in motion social changes that results in an eventual influx of words from French in the 13th century (Singh, 2013); in this case, some high frequency words had low phonotactic probability word forms borrowed from French. Third, word frequencies may change as their meanings are extended to new communicative contexts (Brochhagen et al., 2023). As such, the trade-off between speaker effort and listener robustness are only two of many pressures that we expect to operate on the word forms that comprise a lexicon (see also Piantadosi et al., 2011 regarding the relatively small proportion of variance in word length that can be accounted for by other lexical features).

Second, even if we make the strong simplifying assumption that the two forces treated here are the only two pressures on language, it is still quite difficult to know how their relative importance across communicative contexts would conspire to produce the lexicons we observe in natural languages. Some words are more critical to help listeners interpret utterances than others, while others contribute less. Relatedly, some social situations require more clarity from speakers, while others stress economy of expression. Further, production costs as well as the costs of failed transmissions vary by modality. As such, we take care in the current work to specify that we believe that the pressures toward economy of effort by speakers and listener robustness place *bounds* on the space of possible word forms rather than strictly determine them: if a word is unlikely to be recovered or is

too much work for the speaker, then it is likely to change or go extinct, however, words should also be expected to move around an "acceptable" envelope of trade-offs, subject to the many competing pressures identified above. The current work provides a quantitative characterization of these bounds and some evidence for them, setting the stage for future rigorous quantitative work that looks at how all of these pressures interact, for example adding a measure of morphological transparency.

6 Conclusion

Here we provide evidence that the canonical relationship between the length of word forms and their frequency of usage is a special case of an even broader relationship between phonotactic probability and frequency. However, phonotactic probability also trades off with the number and strength of competing lexical interpretations for a given word form as measured under an idealized Bayesian model of speech recognition. To maintain a lower bound on the rate of word recognition requires that low frequency word forms have sufficiently distinctive—and thus low phonotactic probability—word forms. Thus while speakers may prefer to simplify and shorten words to reduce production costs, they are limited by listeners' requirements for sufficiently distinctive word forms for successful recognition. The observed correspondences between frequency, phonotactic probability, and aggregate competitor probability in word recognition provide new evidence that the psycholinguistic processes at work in producing and perceiving speech may help to shape human languages at the broadest scales.

References

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5), 3048–3058.

Baayen, H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database. Release 2 (CD-ROM).

Baayen, R. and del Prado Martín, F. (2005). Semantic density and past-tense formation in three germanic languages. *Language*, 81(3), 666–698.

Baayen, R., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30(11), 1174–1220.

Balota, D., Yap, M., Hutchison, K., Cortese, M., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459.

Bell, A., Brenier, J., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), 92–111.

Bentz, C. and Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*: Universitätsbibliothek Tübingen.

Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1 LDC2006T13.

Brants, T. and Franz, A. (2009). Web 1T 5-gram, 10 European Languages Version 1 LDC2009T25.

Brochhagen, T., Boleda, G., Gualdoni, E., and Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, 381(6656), 431–436.

Brysbaert, M., Warriner, A., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3), 904–911.

Bybee, J. (2003). *Phonology and language use*, volume 94. Cambridge University Press.

Bybee, J. and McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition.

Casas, B., Hernández-Fernández, A., Catala, N., Ferrer-i Cancho, R., and Baixeries, J. (2019). Polysemy and brevity versus frequency in language. *Computer Speech & Language*, 58, 19–50.

Chen, S. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computing, Speech & Language*, 13(4), 359–393.

Coady, J. A. and Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of experimental child psychology*, 89(3), 183–213.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics* (pp. 90—-98). Somerville, MA: Cascadilla Proceedings Project.

Coltheart, M., Davelaar, E., Jonasson, J., and Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Lawrence Erlbaum Associates.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (2004). Patterns of english phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, 116(6), 3668–3678.

Dale, R. and Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in complex systems*, 15(03n04), 1150017.

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

Dautriche, I., Mahowald, K., Gibson, E., and Piantadosi, S. (2016). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, (pp. 1–21).

DeLong, K., Urbach, T., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117.

Evans, N. and Levinson, S. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429–448.

Fedzechkina, M., Jaeger, T., and Newport, E. (2012). Language learners restructure their input to facilitate efficient communication. *Proc. Natl. Acad. Sci. U.S.A.*, 109(44), 17897–17902.

Ferrer-i Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., and Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8), 1565–1578.

Ferrer-i-Cancho, R. and Solé, R. (2002). Zipf's law and random texts. *Advances in Complex Systems*, 5(01), 1–6.

Frauenfelder, U. H., Baayen, R. H., and Hellwig, F. M. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, 32(6), 781–804.

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

Futrell, R., Albright, A., Graff, P., and O'Donnell, T. (2017). A generative model of phonotactics. *Transactions of the Association for Computational Linguistics*, 5, 73–86.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences U.S.A.*, 112(33), 10336–10341.

Gabelentz, G. v. d. (1901). *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Weigel.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, 84(3), 474–496.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4), 789–806.

Gibson, E., Bergen, L., and Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences U.S.A.*, 110(20), 8051–8056.

Goldinger, S. D., Luce, P. A., and Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of memory and language*, 28(5), 501–518.

Goldwater, S., Johnson, M., and Griffiths, T. (2005). Interpolating between types and tokens by estimating power-law generators. *Advances in neural information processing systems*, 18.

Gow, D. W., Schoenhaut, A., Avcu, E., and Ahlfors, S. P. (2021). Behavioral and neurodynamic effects of word learning on phonotactic repair. *Frontiers in Psychology*, 12, 590155.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., and Lin, F.-H. (2008). Lexical influences on speech perception: a granger causality analysis of meg and eeg source estimates. *Neuroimage*, 43(3), 614–623.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of Human Language* (pp. 73–113). Cambridge, MA: MIT Press.

Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, 18(2), 180–205.

Hauser, M., Chomsky, N., and Fitch, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579.

Hoover, J. R., Storkel, H. L., and Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, 63(1), 100–116.

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

Jaeger, T. F. and Buz, E. (2017). Signal reduction and linguistic encoding. *The Handbook of Psycholinguistics*, (pp. 38–81).

Johnson, M., Griffiths, T., and Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in Neural Information Processing Systems*, 19.

Kanwal, J., Smith, K., Culbertson, J., and 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.

Kemp, C. and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084), 1049–1054.

Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4), 978–990.

Ladd, D., Roberts, S. G., and Dediu, D. (2015). Correlational studies in typological and historical linguistics. *Annu. Rev. Linguist.*, 1, 221–41.

Landauer, T. K. and Streeter, L. A. (1973). 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.

Levelt, W. J. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition*, 42(1-3), 1–22.

Levshina, N. (2022). Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, 24(2), 280.

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

Levy, R. (2008b). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on Empirical Methods in Natural Language Processing* (pp. 234–243).

Levy, R. and Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 849–856). Cambridge, MA: MIT Press.

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

Luce, P. and Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear Hear*, 19(1), 1–36.

Luce, P. A., Pisoni, D. B., and Goldinger, S. D. (1990). Similarity neighborhoods of spoken words.

Mahowald, K., Dautriche, I., Gibson, E., and Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive science*, 42(8), 3116–3134.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313–318.

Mandelbrot, B. (1954). Simple games of strategy occurring in communication through natural languages. *Transaction of the IRE Professional Group on Information Theory PGIT*, 3(3), 124–137.

Marian, V., Bartolotti, J., Chabal, S., and Shook, A. (2012). Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman and Company.

Meylan, S. C., Foushee, R., Wong, N. H., Bergelson, E., and Levy, R. P. (2023). How adults understand what young children say. *Nature Human Behaviour*, (pp. 1–15).

Meylan, S. C. and Griffiths, T. L. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive science*, 45(6), e12983.

Meylan, S. C., Nair, S., and Griffiths, T. L. (2021). Evaluating models of robust word recognition with serial reproduction. *Cognition*, 210, 104553.

Michel, J. et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.

Miller, G. (1957). Some effects of intermittent silence. *American Journal of Psychology*, 70, 311–314.

Mohri, M., Pereira, F., and Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1), 69–88.

Norris, D. and McQueen, J. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395.

Pagel, M., Atkinson, Q., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449(7163), 717.

Piantadosi, S., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the Natuonal Academy of Sciences U.S.A.*, 108(9), 3526–9.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3), 280–291.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech*, 46(2-3), 115–154.

Richtsmeier, P., Gerken, L., and Ohala, D. (2011). Contributions of phonetic token variability and word-type frequency to phonological representations. *Journal of Child Language*, 38(5), 951–978.

Romani, C., Galuzzi, C., Guariglia, C., and Goslin, J. (2017). Comparing phoneme frequency, age of acquisition, and loss in aphasia: Implications for phonological universals. *Cognitive neuropsychology*, 34(7-8), 449–471.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1), 140–155.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. (2020). fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138, 107307.

Singh, I. (2013). *The history of English: a student's guide*. Routledge.

Smith, N. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of 7th International Conference on Spoken Language Processing (ICSLP '12)*, volume 2 (pp. 901–904).

Storkel, H., Armbrüster, J., and Hogan, T. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, 49(6), 1175–1192.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25(2), 201–221.

Strand, J. and Liben-Nowell, D. (2016). Making long-distance relationships work: Quantifying lexical competition with hidden markov models. *Journal of Memory and Language*, 90, 88 – 102.

Suárez, L., Tan, S. H., Yap, M. J., and Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, 18(3), 605–611.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Vincent-Lamarre, P., Massé, Alexandre, B., Lopes, M., Lord, M., Marcotte, O., and Harnad, S. (2016). The latent structure of dictionaries. *Topics in Cognitive Science*, 8(3), 625–659.

Vitevitch, M. and Luce, P. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3), 374–408.

Vitevitch, M. and Luce, P. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of memory and language*, 52(2), 193–204.

Vitevitch, M., Luce, P., Pisoni, D., and Auer, E. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1-2), 306–311.

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

Weber, A. and Smits, R. (2003). Consonant and vowel confusion patterns by american english listeners. In *15th International Congress of Phonetic Sciences [ICPhS 2003]*.

Weissbart, H., Kandylaki, K. D., and Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of cognitive neuroscience*, 32(1), 155–166.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., and P. Levy, R. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*.

Wurm, L. H. and Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of memory and language*, 72, 37–48.

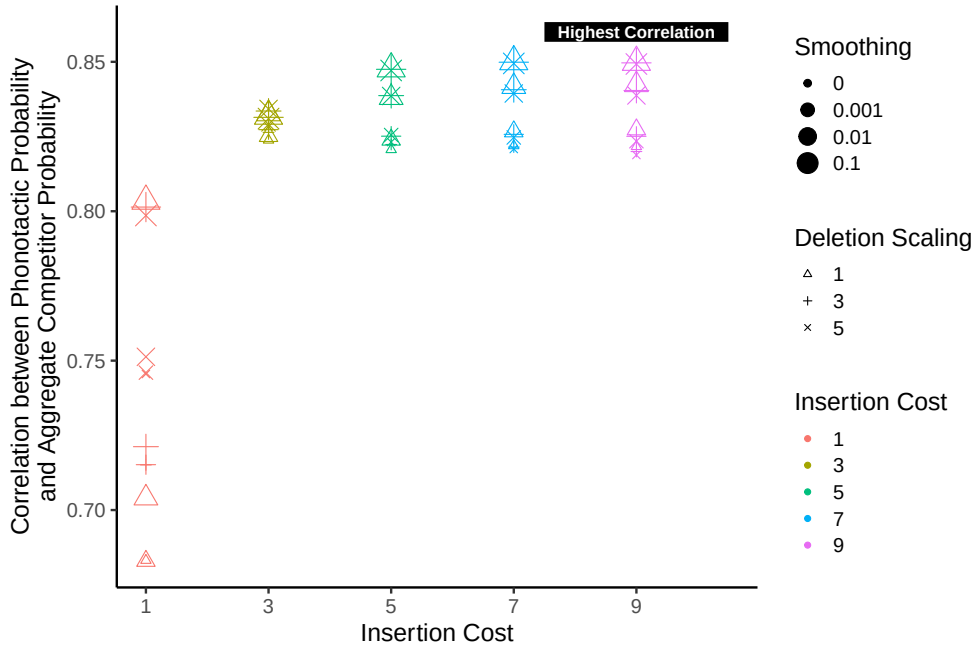
Zipf, G. (1935). *The Psychobiology of Language*. Houghton-Mifflin.

Zipf, G. (1949). *Human Behaviour and the Principle of Least-Effort*. Cambridge, MA: Addison-Wesley.

Supplementary Information

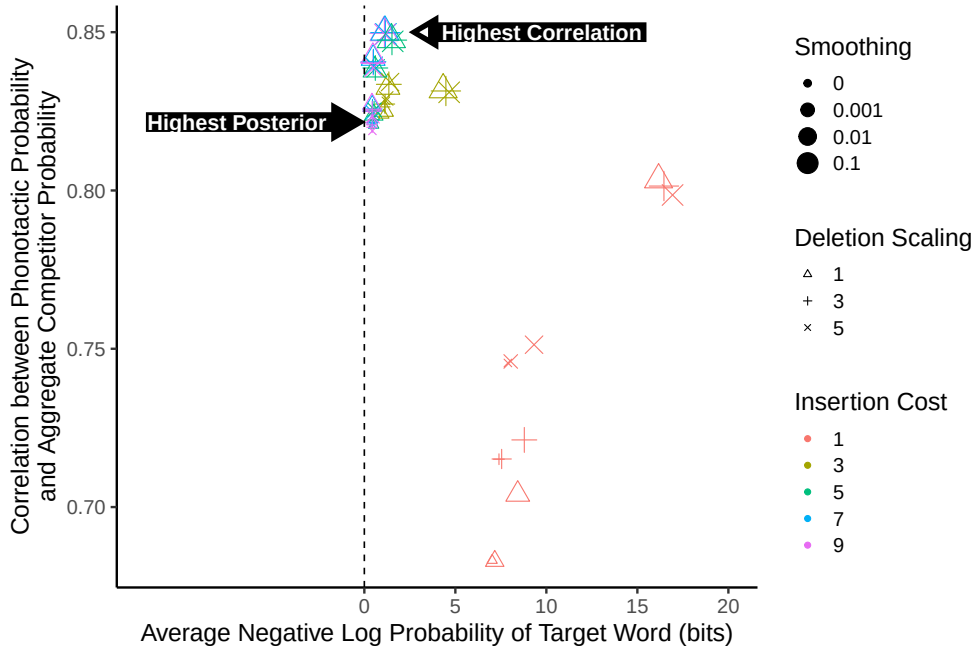
Grid Search for Likelihood Parameters in Study 2

We present here correlations and posterior values across the parameter space used to fit the likelihood of the Bayesian model using Study 2.

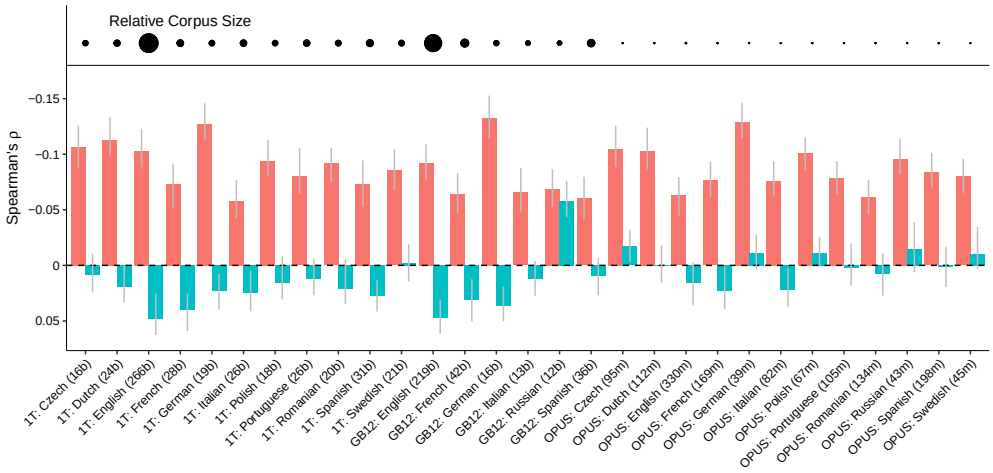


Supplementary Figure 1. The correlation between phonotactic probability and aggregate competitor probability as a function of free parameters in Study 2. Each point represents the results for a single model reflecting a combination of smoothing (probability mass equally apportioned to edits outside of those seen in the lab-based phoneme confusion data), deletion scaling (a multiplier for the rate of deletion seen in the lab-based data), and insertion cost (the negative log probability of inserting an extra phoneme). The highest correlation is found at high insertion costs and high smoothing, though the correlation is positive throughout the parameter space.

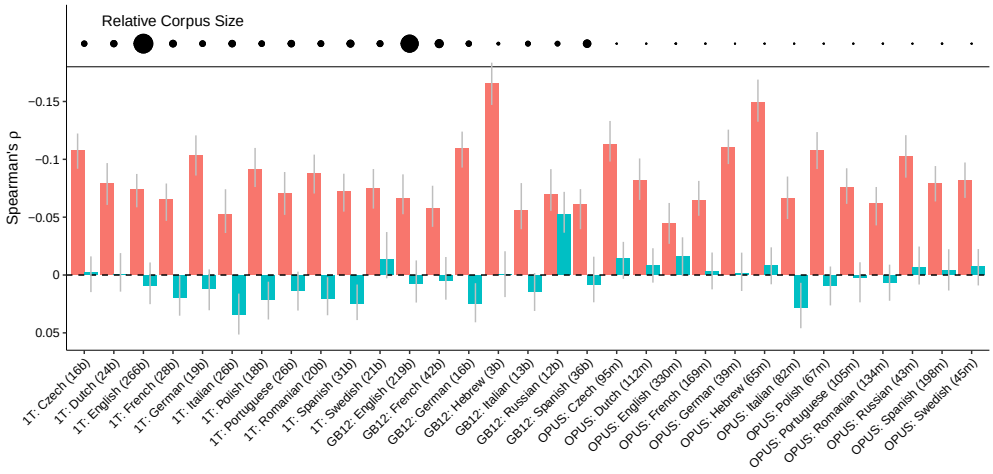
Alternative Test of Partial Correlations



Supplementary Figure 2. The relationship between average posterior probabilities under the model and the correlation between phonotactic probability and aggregate competitor probability across the parameter space in Study 2. Each point represents the results for a single model reflecting a combination of smoothing, deletion scaling, and insertion cost. The highest posterior probability model has lower smoothing values than the model that yields the highest correlation, though the correlation is similar.



Supplementary Figure 3. Partial correlations following Wurm and Fisicaro (2014) computed over phonemic representations (compare with Figure 1, top). Red represents the correlation of residualized PIC with residualized frequency (with both factors residualized on length). Green represents the correlation of residualized length with residualized frequency (with both factors residualized on PIC. Bars indicate Spearman's ρ for the two variables for the $n = 25000$ most frequent words in each dataset. Gray lines indicate the 99% bootstrapped confidence interval. This provides corroborating evidence of the stronger relationship between word frequency and PIC.



Supplementary Figure 4. Partial correlations following Wurm and Fisicaro (2014) computed over orthographic representations (compare with Figure 1, bottom). The plotting conventions are the same as the preceding figure. This provides corroborating evidence of the stronger relationship between word frequency and PIC.