Word forms—not just their lengths—are optimized for efficient communication

Stephan C. Meylan
Department of Psychology
University of California, Berkeley
smeylan@berkeley.edu

Thomas L. Griffiths
Department of Psychology
University of California, Berkeley
tom_griffiths@berkeley.edu

Abstract

The inverse relationship between the length of a word and the frequency of its use, first identified by G.K. Zipf in 1935, is a classic empirical law that holds across a wide range of human languages. We demonstrate that length is one aspect of a much more general property of words: how distinctive they are with respect to other words in a language. Distinctiveness plays a critical role in recognizing words in fluent speech, in that it reflects the strength of potential competitors when selecting the best candidate for an ambiguous signal. Phonological information content, a measure of a word’s string probability under a statistical model of a language’s sound or character sequences, concisely captures distinctiveness. Examining large-scale corpora from 13 languages, we find that distinctiveness significantly outperforms word length as a predictor of frequency. This finding provides evidence that listeners’ processing constraints shape fine-grained aspects of word forms across languages.

Despite their apparent diversity, natural languages display striking structural regularities (Greenberg, 1963; Evans and Levinson, 2009; Futrell et al., 2015). How such regularities relate to human cognition 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). Prominent among these regularities is the well-known relationship between word length and frequency: across languages, frequently-used words tend to be short (Zipf, 1935). In a classic work, Zipf (Zipf, 1935) posited that this pattern emerges from speakers minimizing total articulatory effort by using the shortest form for words that are used most often, following what he later called the Principle of Least Effort (Zipf, 1949). While the underlying cause has been the subject of debate (Yule, 1944; Miller, 1957; Cancho and Sole, 2003; Conrad and Mitzenmacher, 2004; Piantadosi, 2014), this relationship between word length and frequency remains one of the most robust statistical laws that describe human languages.

Here, we propose a generalization of Zipf’s analysis and present two possible listener-focused explanations. We show that a word’s frequency is inversely related to its distinctiveness—how easily it can be identified as the intended message for a given speech signal. While speakers prefer easier-to-produce less distinctive forms, they are constrained by listeners’ need for sufficiently distinctive forms to differentiate each word from others, especially if a speaker’s intended word has higher-frequency competitors. If word recognition is modeled as Bayesian inference (Norris and McQueen, 2008; Balling and Baayen, 2012), the probability of successful recognition depends on both the prior probability of the intended word and on the number and strength of alternative words (“competitors”). We define a statistical measure of distinctiveness that succinctly captures the diagnosticity of a
word form by assessing the aggregate strength of competitors in the language. We then show that distinctiveness should be inversely related to frequency, if languages are constructed to equalize error rates for low and high frequency words.

Importantly, distinctiveness subsumes Zipf’s observation regarding the relationship of length and frequency as a special case. Length is a naïve approximation of the distinctiveness of a word form if longer strings are simply less probable. We demonstrate that a more comprehensive measure of distinctiveness that takes into account the sound-to-sound and letter-to-letter sequences in a language accounts for significantly more frequency-related variance than does length across a broad sample of natural languages. This relationship between frequency and distinctiveness adds to a growing body of evidence that cognitive constraints influence the structural properties of natural languages (Hawkins, 1994; Fedzechkina et al., 2012; Futrell et al., 2015).

Model

We define a probabilistic language model to characterize the distinctiveness of word forms in terms of their constituent sound-to-sound or letter-to-letter transitions. This model can be fit using a large written sample of a particular language. The starting point for the model is formalizing the task of the listener as a rational statistical inference.

Bayesian inference and distinctiveness

Upon hearing a string of sounds \( s \), a listener has to infer what word \( w \) was intended by the speaker. This can be formulated as a problem of Bayesian inference. The listener should calculate a posterior distribution \( P(w|s) \) over words based on the sounds. Applying Bayes’ rule, this is given by

\[
P(w|s) = \frac{P(s|w)P(w)}{P(s)}
\]

(1)

where \( P(s|w) \) is the probability of hearing \( s \) if \( w \) is the intended word, \( P(w) \) is the prior probability of the word \( w \) intended by the speaker, and \( P(s) \) is the probability of hearing \( s \).

Assuming that sounds are produced faithfully, such that \( P(s|w) \) is close to 1 for a particular string \( s_w \) for each \( w \) and close to 0 otherwise, we obtain the approximation

\[
P(w|s_w) \approx \frac{P(w)}{P(s_w)}
\]

(2)

which expresses the probability that the word \( w \) is correctly identified as a function of its normalized frequency, \( P(w) \), and the probability of the string \( s_w \) in the language, \( P(s_w) \). We define the distinctiveness of a word to be inversely related to \( P(s_w) \): intuitively, a word that shares the same sound sequences with many other words is less distinctive.

Following this logic, we use the phonological information content (PIC) of a word to measure the distinctiveness of its phone-to-phone (sound-to-sound) transitions, or its approximation in character-to-character transitions (Cohen Priva, 2008). Analogous to the metric of lexical surprisal (negative log conditional probability) used to measure the predictability of a word given preceding words (Levy, 2008; Piantadosi et al., 2011; Smith and Levy, 2013), PIC is the negative log probability of the sequence of phones (distinct meaningful sounds) or letters in the string comprising a word. To obtain a compact representation of the probabilities of various phone or character sequences in a language we estimate an \( n \)-phone or \( n \)-character model (analogous to an \( n \)-gram model over words (Manning and Schütze, 1999) but computed over individual phones or characters) from a corpus sample. To support a stronger test of the relationship between distinctiveness and frequency, we avoid the circularity that more probable words necessarily contain more probable sequences by estimating the transition probabilities using the type inventory (unique words) in the language.

Under this model, the phonological information content \( PIC(w) \) of a word is defined as:

\[
PIC(w) = -\log P(s_w)
\]

(3)

\[
= -\log P(l_1, \ldots, l_{|s_w|}) \text{ for } l \in s_w
\]

(4)

\[
= -\sum_{i=1}^{\text{|s_w|}} \log P(l_i|l_{i-1}, \ldots, l_{i-1}).
\]

(5)
Table 1: Estimates of the information content of the word `motorcycle` under uniform character probability (proportional to word length) and under a model that takes into account sequential dependencies up to length 5 obtained from the 25,000 most frequent words in the English Google Books (2012) corpus.

|     | “M”   | “O”   | “T”   | “O”   | ... | Σ Bits |
|-----|-------|-------|-------|-------|-----|--------|
| **Uniform Character Probabilities** | log \( P(M) \) | log \( P(O) \) | log \( P(T) \) | log \( P(O) \) |     | 47.004 |
|     | 4.700 | 4.700 | 4.700 | 4.700 |     |        |
| **5-Character Phonological Information Content Model** | log \( P(M) \) | log \( P(O|M) \) | log \( P(T|M) \) | log \( P(O|MOT) \) |     | 26.327 |
|     | 5.358 | 3.223 | 3.451 | 6.110 |     |        |
Figure 2: Frequency demonstrates a stronger negative correlation with distinctiveness in all datasets analyzed; this difference is significant in all cases but one (Russian in Books 2012). Bars indicate Spearman’s ρ for the two variables for the n = 25000 most frequent words in each dataset. Gray lines indicate 99% bootstrapped confidence interval. Black lines indicate the correlation with the other measure of word difficulty partialed out. After distinctiveness is partialed out, positive partial correlations are obtained for word length in the English and Dutch 1T, and German and Hebrew Google Books 2012, and Hebrew and Dutch OPUS datasets. Circles indicate the number of tokens in each dataset.

However, this approximation fails to capture regularities in the lexical substructure observed in natural languages in two obvious ways. First, symbols are not equiprobable: across the word types in the English lexicon, w is substantially less common than e. Second, the sound symbols are not independent: the sound t in English is followed more frequently by i or e and very rarely—if ever—by b or g. People have rich knowledge of the relative prominence of these sequences in their respective languages—just as they have rich knowledge of inter-word statistical dependencies—and can call upon this knowledge in spoken word recognition (Vitevitch and Luce [1999], Luce and Large [2001]). Ample psycholinguistic evidence also suggests that people are capable of using sub-word information, for example using sounds from the beginning of a word to predict possible continuations (Marslen-Wilson and Welsh [1978], Marslen-Wilson [1987], Zwitserlood [1989], Eberhard et al. [1995]).

When the probability of sub-word sequences is taken into account, sequences of the same length can vary markedly in distinctiveness: while xylophone and something are both nine letters long, the latter is comprised of significantly less common subsequences. When phonological information content is computed with a more accurate model of the structure of words, it differs from length because it captures deviations from independence (see Table 1).

**Benefits of distinctiveness**

Conceiving of words in terms of their distinctiveness has several benefits compared to length. First, distinctiveness (as measured by PIC) is a much more fine-grained measure of word form complexity, which generates predictions contrary to word length in many cases. A shorter word may contain a relatively low probability phone sequence (e.g., depth, 4 phones/5 letters, PIC = 17.86 bits), while a longer word may contain a higher probability, less informative sequence (e.g., ground, 5 phones/6 letters, PIC = 12.85 bits).

Second, PIC is closely related to metrics of lexical neighborhood density used in psycholinguistic models of spoken word recognition. Neighborhood density reflects how many words have a similar form to a given word; while proposals vary on how to measure neighborhood density, they share the intuition that words with more similar word forms (or “neighbors”) are harder to recognize because there are more competitors consistent with a given signal. PIC is a kind of type frequency-weighted neighborhood density (Luce and Pisoni [1998]), however it measures the number of competitors at each successive phone—a feature consistent with empirical results suggesting incremental phone-by-phone processing (Eberhard et al. 1995). PIC thus constitutes a more detailed measure of neighborhood
density than the canonical measure of Coltheart’s \( N \) (Coltheart et al., 1977), the number of words within a edit distance of one of the target word. In particular, PIC is sensitive to competition effects from hearing the partial word form: while there are approximately 40 possible candidates upon hearing /\( \text{Thi} \)/ in \textit{thesis} in a large sample of English, neighborhood density as assessed by Coltheart’s \( N \) is much lower (the only competitors by this criterion are \textit{theses} and \textit{Theseus}).

Finally, PIC has the added benefit that it can be interpreted as a straightforward measure of processing difficulty. Accounts of human sentence parsing have shown that relative entropy (surprisal) approximates the difficulty of processing a single word given its sentential context (Hale, 2001; Levy, 2008; Smith and Levy, 2013). We propose that this same relationship might hold at a lower level of language structure: the difficulty of processing a single phone given the preceding phones is the relative entropy of the observed distribution with respect to the expected distribution.

A relationship between distinctiveness and frequency

We provide two listener-centric explanations as to why there might be an inverse relationship between distinctiveness and frequency. First, distinctiveness provides a way to measure the effort that listeners expend in comprehending speakers; total comprehension effort is minimized when the most frequent words are the least distinctive. Zipf’s original explanation for the inverse relationship between the frequency of a word and its length was based on the idea that languages are shaped by the desire of speakers to expend the least effort in producing words. If longer strings are more effortful to produce, it makes sense that they should be associated with less frequent words. However, we can imagine a similar argument being applied on the part of listeners: that languages are shaped by the desire of listeners to minimize the effort they expend in comprehending speakers. Distinctiveness more accurately indexes comprehension effort, above and beyond word length.

Second, an inverse relationship between distinctiveness and frequency can also be derived from an invariance argument. Recent work in computational psycholinguistics has discovered that a variety of syntactic phenomena can be predicted from the Uniform Information Density hypothesis: that languages are structured such that each word in an utterance provides approximately the same amount of information (Genzel and Charniak, 2002; Aylett and Turk, 2004; Levy and Jaeger, 2007; Piantadosi et al., 2011). A similar invariance principle that might be relevant at the level of the words themselves is Uniform Recognizability: that the probability any word is successfully recognized is approximately the same. Under Equation 2, the probability of successfully recognizing \( w \) from its associated string \( s_w \) is \( P(w)/P(s_w) \). Making this constant across words means that we should expect \( P(w) \) and \( P(s_w) \) to be directly related, and hence \( P(w) \) and PIC\((w) \) to be negatively correlated.

Methods

Datasets for Frequency and Surprisal Estimates

The Google Web 1T datasets were downloaded from the Linguistic Data Consortium (Brants and Franz, 2006, 2009); 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 \( n \)-grams with punctuation-only words were discarded, and punctuation appearing with other text, with the exception of apostrophes, was removed. All characters 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, records with part-of-speech tags were discarded, along with records from earlier than 1800. UTF-8 encoding was maintained throughout for all languages and datasets. Hebrew strings were represented with right-normalized forms. Counts were stored using ZS, a specialized file format for efficient retrieval of \( n \)-gram counts (Smith, 2013).

Estimating Sentential Information Content

Following Piantadosi et al. (2013) we analyze a wordlist constructed from the most frequent words in each dataset, filtered in each case by the largest available collection of movie subtitles to better reflect spoken usage. Token frequencies were computed from the 2013 release of the the OPUS subtitle corpus. For each language, the list of unique types were filtered by those words recognized by the UNIX utility Aspell for the relevant locale. After filtering, we took the top 25,000 lexical items.
For words with the same number of meaningful sounds (in this case seven), distinctiveness measured with Phonological Information Content accounts for additional variance in frequency. The red line indicates the best linear fit for all words with seven sounds. Though data presented here reflect phonetic transcriptions, words are labeled in standard English form. Labeled points (black) are randomly sampled from the $n = 3,415$ words with phonetic transcription of length seven.

in each dataset and computed for each word $w$ the negative log unigram probability (proportional to log normalized frequency) and negative mean log trigram probability across contexts, following (Piantadosi et al., 2011), $-\frac{1}{N} \sum_{i=1}^{N} \log P(W = w | C = c_i)$, where $c_i$ is the context for the $i$th occurrence of $w$ and $N$ is the frequency of $w$ in the dataset.

Estimating Phonological Information Content

A five-character transition model was estimated for each language using the 25,000 most frequent in-dictionary words from the corresponding OPUS subtitle corpus. We also produced a five-phone transition models for all languages with the exception of Hebrew using IPA transcriptions from an automatic speech synthesizer, eSpeak. Using IPA representations for words accounts for language-specific variations 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 decisions, for example deciding whether ‘a’ vs. ‘á’ should be merged or kept as separate orthographic variants in Spanish. Loan words and acronyms can greatly affect the obtained transition probabilities, especially when the transitions observed in all types are equally weighted (e.g., if the transitions in “Okeechobee,” “mañana,” and “ACLU” are as heavily weighted in a phonotactic model of English as the transitions in “they” and “will”). To minimize these effects, we use only non-capitalized types present in Aspell dictionaries to build sound and character transition models for each language. To avoid overfitting among higher order sequences, phone and character transition probabilities were computed with modified Kneser-Ney smoothing (Chen and Goodman, 1999) with interpolation on orders 3, 4, and 5 using the SRILM toolkit (Stolcke, 2002).

Results

We investigated the correlation between distinctiveness and frequency in large corpus samples (43m to 266b words) in 13 languages. We limit our analysis to languages with phonetic scripts. In each case we compute the frequency and in-context predictability (mean trigram surprisal, or negative mean log conditional probability under a trigram model) for each word, and measure its PIC under an $n$-phone model of the sound transitions in the language, as well as an $n$-character model. If length is a reduced-resolution approximation of distinctiveness, then we expect an even stronger relationship between distinctiveness and frequency across languages (see Materials and Methods).
Figure 4: Comparison of correlations between combinations of sentence-level predictability (unigram surprisal vs. mean trigram surprisal) and word-level measure (word length or phonological information content, PIC). Bars indicate Spearman’s ρ for the two variables for the n = 25,000 most frequent words in each dataset. A. PIC is more strongly correlated with unigram surprisal (negative log unigram probability) than mean trigram surprisal (negative mean log conditional probability) across a broad range of languages and datasets. B. The correlation between PIC and frequency is stronger than the correlation between mean trigram surprisal and length in all but one dataset (English Google 1T). The latter is the key relationship identified in Piantadosi et al. (2011). C. Among words found in a relevant language-specific dictionary, unigram surprisal is a better predictor than mean trigram surprisal of word length in most cases. Notable exceptions are English and French in Google 1T and English and German in Google Books 2012. Correlations of OPUS datasets are omitted because these datasets are too small to reliably estimate mean trigram surprisal.

Examining the top 25,000 most frequent in-dictionary types in each language, we obtain a systematically higher correlation between frequency and distinctiveness, as measured by PIC, than frequency and word length (Figure 5). At each word length, distinctiveness explains additional variance in word frequency (Figure 5). Building the model from phonetic transcriptions, this pattern held in 11 of 11 languages in the Google 1T datasets, 6 of 7 languages from Google Books 2012, and all 13 languages from the 2013 OPUS corpus. Building the model from characters, this pattern held in all cases. In many cases (9 of 11 languages in Google 1T, 4 of 7 in Google Books 2012, and 2 of 13 in OPUS) the partial correlation of PIC and frequency—with word length partialed out—is higher than that of word length. The obtained correlations are even stronger than those between word length and mean contextual predictability (Piantadosi et al., 2011).

A similar pattern of results emerges regardless of whether the model is computed over characters or phones; the sole exception is the Russian corpus from Google Books 2012 where word length is a stronger predictor when the model is computed over phones. However, this dataset is an outlier in two notable ways. Russian shows the lowest correlation between frequencies obtained from 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 phone 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 test of whether this relationship holds in languages with extensive nonconcatenative morphology (word root modification rather than prefixes and suffixes).
Figure 5: For words that have entered English more recently, the correlation between frequency and PIC is higher than the correlation between frequency and word length. This is consistent with the hypothesis that PIC can capture finer-grained changes in word forms—which are more likely on shorter timelines—besides changes in length. Error bars show bootstrapped 99% confidence intervals.

Discussion

The relationship between word length and frequency is one of the most robust empirical findings regarding the structure of natural languages. Treating word recognition in terms of Bayesian inference, we show that the requirement for successful recognition not only motivates the above relationship, but generates additional strong predictions regarding the probability of the phone and character sequences that make up word forms. Analyses of large-scale corpora from 13 languages across three datasets (web pages, books, and movie subtitles) substantiate these claims. In the remainder of the paper we consider some of the implications of these results.

Linking psycholinguistics and language change

The approach that we have taken in this paper potentially provides a link between the psycholinguistic processes at work in producing and perceiving speech and the factors that shape human languages. Previous research has demonstrated that speakers may “reduce”—underarticulate, shorten, weaken, or omit altogether—frequent or predictable words, for example swapping “probly” for “probably” in fluent speech (Aylett and Turk, 2006; Bell et al., 2009; Gahl et al., 2012). On longer timescales, more probable (often shorter) variants may become the dominant form in the language, e.g., English lunch displaced luncheon in the 19th and 20th centuries (oed, 2016).

Whereas word length can only reflect deletion and epenthesis (the insertion of material), PIC is sensitive to changes that maintain the same length, including assimilation (a sound becoming more similar to a neighbor), dissimilation (a sound becoming less similar to a neighbor), and metathesis (transposition of proximal sounds). The change from Middle English aks to ask, for example, is reflected in a change in PIC for the two forms though both have the same length.

To examine how a word’s PIC and frequency relate given how long it has been part of the language, we obtained the date of first appearance for 31,027 English types from the Oxford English Dictionary. We stratified these dates by century, assigning pre-700 (corresponding to proto-Old English) words to the 7th century, and computed Spearman’s rank correlation coefficient against current word frequency for the set of words corresponding to each century. The obtained correlation between PIC and frequency is stronger than that between length and frequency for words entering English between the 13th and 20th century (bootstrapped Z-tests of the difference) (Figure 5). This is consistent with the hypothesis that PIC better reflects frequency on shorter timescales because it is sensitive to a wider variety of psycholinguistically-motivated sound changes.

Relationship to Preceding Context

Piantadosi et al. (2011) recently found that taking into account contextual predictability (mean trigram surprisal, or average in-context conditional probability under an n-gram model) better
predicts word length than using frequency (operationalized as unigram surprisal, or negative log probability). However, we find a qualitatively different pattern of results for phonological information content. Examining the same set of words we find that the correlation between frequency and PIC is higher than the correlation between mean trigram surprisal and PIC in all cases (Figure 4, A). The correlation between frequency (negative log unigram probability) and PIC is also greater than the correlation between mean trigram surprisal and word length in all but one case (Figure 4, B). We find substantially attenuated support for the principal claim in (Piantadosi et al., 2011), in that we find unigram frequencies better predict word length than does mean trigram surprisal (Figure 4, C). This discrepancy seems to reflect refinements to the list of lexical items analyzed, improvements in the data preparation and analysis method, and underlying issues in the construct validity of mean surprisal under an \( n \)-gram model (see SI: Preceding Context Revisited).

In contrast to the original formulation by Zipf which focused on minimization of speaker effort, in-context predictability and word form distinctiveness both implicate listener-oriented pressures in the relationship between frequency and word form. The two proposals make contrasting predictions regarding what factors are most relevant to generalizing the relationship, however. In-context predictability generalizes the notion of a listener’s prior beliefs regarding the probability of a given word, replacing frequency with predictability. Distinctiveness provides a more detailed characterization of word form complexity. In principle, both proposals could be true: the strongest relationship could be between in-context predictability and distinctiveness, though we do not find empirical support for this in the datasets analyzed here (though see SI: Computing Lexical Information Content).

**Limitations and Future Directions**

The statistical properties of the everyday language environment of speakers may differ substantially from the corpora examined here, which were constructed from webpages (Google 1T), books (Google Books 2012) and movie subtitles (OPUS). Lexical surprisal estimates for the lowest rank items may be biased, given relatively small datasets for languages other than English. Nonetheless, these corpora constitute the best approximation to naturalistic language use of sufficient size to support lexicon-wide analysis, particularly the calculation of mean information content for relatively rare words. As more corpora are made available, we hope to extend this analysis to more diverse languages.

We make two key simplifications in the current work with respect to the measurement of PIC: we use a linear, symbolic representation of the phonemes that comprise a word rather than hierarchical, feature-based representation, and we compute PIC with respect to the true preceding context within each word. A generative model that assumes a rich hierarchical representation of word form structure can account for regularities than an \( n \)-phone model cannot, leading to a modest improvement in predicting a held-out set of words (Futrell et al., 2016). Adapting such a model may lead to improved estimates of word form probability—and hence PIC—in a language.

With respect to the second simplification, listeners do not have access to the true identity of preceding phones, and instead likely marginalize over a distribution of preceding sounds within a word in the process of word recognition. This may mean that listeners have comparatively peaked estimates of the preceding sounds towards the ends of longer words. In future work we intend to investigate whether this uncertainty accounts for additional variance in word frequency.

**Acknowledgements**

This material is based upon work supported by the US National Science Foundation Graduate Research Fellowship under grant no. DGE-1106400 and NSF grant no. SMA-1228541. Special thanks to Steven Piantadosi for sharing materials, helpful commentary on early drafts from Terry Regier, and members of the Computational Cognitive Science Lab at UC Berkeley.

**References**

(2016). OED Online.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang. Speech*, 47(1):31–56.
Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *J Acoust. Soc. of Am.*, 119(5):3048–3058.

Balling, L. and Baayen, R. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, 125(1):80–106.

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

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.

Cancho, R. F. i. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. U.S.A.*, 100(3):788–791.

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

Cohen Priva, U. (2008). Using information content to predict phone deletion. In *Proc. of the 27th West Coast Conf. on Formal Ling.*, pages 90—98, Somerville, MA. Cascadilla Proceedings Project.

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

Conrad, B. and Mitzenmacher, M. (2004). Power laws for monkeys typing randomly: the case of unequal probabilities. *IEEE Trans. Inf. Theory*, 50(7):1403–1414.

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., and Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *J. Psycholing. Res.*, 24(6):409–436.

Evans, N. and Levinson, S. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.*, 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.

Futrell, R., Albright, A., Graph, P., and O’Donnell, T. (In press). *Transactions of the Association for Computational Linguistics*.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proc. Natl. Acad. Sci. U.S.A.*, 112(33):10336–10341.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *J. Mem. Lang.*, 66(4):789–806.

Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. *ACL*, pages 199–206.

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

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *ACL*, pages 1–8. Association for Computational Linguistics.

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.

Hawkins, J. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge, UK.

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
Levy, R. and Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In Adv. in Neural Inf. Proc. Sys. 19, pages 849–856, Cambridge, MA. MIT Press.

Luce, P. and Large, N. (2001). Phonotactics, density, and entropy in spoken word recognition. Lang. Cognit. Proc., 16(5-6):565–581.

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

Mandelbrot, B. (1954). Simple games of strategy occurring in communication through natural languages. Trans. of the IRE Prof. Group on Inf. Th., 3(3):124–137.

Manning, C. and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA.

Marian, V., Bartolotti, J., Chabal, S., and Shook, A. (2012). CLEARPOND: cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. PLoS ONE, 7(8):e43230.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. Cognition, 25(1-2):71–102.

Marslen-Wilson, W. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognit. Psychol., 10(1):29 – 63.

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. Am. J. Psychol., 70:311–314.

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

Piantadosi, S. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. Psychon. Bull. Rev., 21(5):1112–1130.

Piantadosi, S., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. Proc. Natl. Acad. Sci. U.S.A., 108(9):3526–9.

Smith, N. (Submitted). ZS: A file format for efficiently distributing, using, and archiving record-oriented data sets of any size.

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 ICSLP, volume 2, pages 901–904.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Proceedings of the 8th International Conf. on Lang. Resources and Evaluation (LREC’12).

Vitevitch, M. and Luce, P. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. J. Mem. Lang., 40(3):374–408.

Yule, G. (1944). The Statistical Study of Literary Vocabulary. Cambridge University Press.

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

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

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. Cognition, 32(1):25–64.
Supplementary Material

Preceding Context Revisited

Piantadosi et al. (2011) found that a word’s mean predictability across contexts is a better predictor of length in characters than unigram surprisal (negative log unigram probability), consistent with the theory of Uniform Information Density. Here we add a substantial caveat to this observation: the correlations obtained in that study can be attributed to an interaction between in-dictionary and out-of-dictionary word types, and that this relationship is significantly weakened when each group is analyzed separately (Figure 6). In fact, mean trigram surprisal has no more predictive power with respect to length than unigram surprisal (negative log unigram probability) within 1) lexical items that can be found in a dictionary 2) words borrowed from English into other languages 3) words not found in any dictionary, primarily names, acronyms, and toponyms.

Using the same list of words from Piantadosi et al. (2011), we classified each word into one of the three categories listed above by 1) testing for membership in the relevant GNU Aspell dictionary for each language, and 2) testing for membership in the English Aspell dictionary. Items found both in the language-specific dictionary as well as English, e.g. Spanish pan, were designated as in-dictionary items. The proportion of items found in each category for each language, as well as example classifications from Spanish, are presented in the inset in Figure 7, center.

Using the same set of mean information content estimates, we find that within-group correlations for all three groups are substantially lower than those reported in (Piantadosi et al., 2011) (Figure 6). Instead, the high global correlation between mean information content and word length emerges from the inclusion of out-of-dictionary items and items from English, which are both more more predictable and shorter than in-dictionary words (Figure 7 right). The global correlations between unigram surprisal and word length found in that study were depressed by out-of-dictionary items, which were shorter yet less frequent (higher unigram surprisal) than in-dictionary terms (Figure 7, left).

Because Piantadosi et al. (2011) represented word forms as closest ASCII equivalents (e.g., manana for Spanish mañana), some non-English words cannot be found in the relevant dictionary. This means that for the above analysis some lexical items that should have been classified as in-dictionary have been classified as out-of-dictionary, with unknown implications for the within-group correlations. Because there is no way to systematically restore these word forms, we compare the correlation between mean trigram surprisal and wordlength and unigram surprisal and wordlength using novel lexical surprisal estimates from the main analysis.

Unigram surprisal is a better predictor of word length in most languages. Key exceptions, however, are consistent with the findings in Piantadosi et al. (2011): English and French in the Google 1T and English and German in Google Books 2012 show a stronger correlation of mean trigram surprisal and word length. English Google 1T and Books 2012 corpora are at least five times larger than the next largest dataset; whether the results obtained reflect cross-linguistic variability or differences in data sparsity will require additional datasets to determine. In the case of Google 1T, this correlation is higher than that of unigram surprisal and PIC, leaving open the possibility that word length is optimized with respect to a different set of constraints than PIC. For further details on the estimation of mean information content see Materials and Methods.

Computing Lexical Information Content

Computing in-context lexical information content using n-gram models may be theoretically unsound in languages with rich morphology. Consider for example that whereas an n-gram model for English would have entries for a handful of forms for the verb sell (e.g., sell, sells, sold, selling), an n-gram model for Spanish, a language with much richer morphology for verbs, needs to have many more entries for the corresponding verb vender owing to the combinatorial space of possible conjugations and clitics (≈160 in Google Books 2012). Though the languages may have similar total lemma frequencies, the computation of mean in-context surprisal is substantially more difficult in Spanish given the sparse support of certain forms (e.g., vèndeselas, corresponding to English imperative you sell them that). Second, if an analysis is conducted on the top n most frequent word forms, high frequency lemmas comprise a greater proportion of the list because they are spread across several discrete entries. Other languages have complex case marking systems for nouns (Russian), Depending
Figure 6: Piantadosi et al. (2011) found higher global correlations between word length and mean trigram surprisal (blue bars in panel 1) than between word length and frequency (unigram surprisal), red bars in panel 1). The correlation is substantially weaker within each of three sub-groups of lexical items: types that are found in the relevant dictionary (panel 2), those found in English (panel 4), and those found in neither dictionary (panel 3).

Figure 7: 2-D density plots depicting the relationship between sentence-level predictor (unigram surprisal or mean trigram surprisal) and word length for the 25,000 most frequent words in 11 languages used in Piantadosi et al. (2011). Words are stratified into three categories: those in a spelling dictionary of the target language (blue), those in a spelling dictionary for English (red), and those not in either (typically names and abbreviations, green). High correlations between mean trigram surprisal and word length, as well as low correlations between unigram surprisal emerge from an interaction between these groups. Marginal boxplots show the median, inter-quartile range (IQR), and 1.5 * IQR for each of the groups. Densities are normalized per group.

on what parts of speech have high morphological complexity, this can lead to substantive differences in the composition of the vocabulary under analysis across languages. Future work will need to investigate how morphological complexity interacts with lexical predictability across languages, and whether different tokenization processes should be used in the estimation of in-context sentential information content (e.g., treating clitics in Spanish as separate words).

Data Issues

Corpus analyses of the scale used here are necessarily noisy. The choice of texts, methods for identifying content (i.e., excluding page numbers, tables, publisher notes etc.), optical character recognition procedure, and tokenization all influence the obtained dataset. Additional data preparation decisions such as text encoding, treatment of punctuation, and treatment of upper and lower case forms can have pronounced effects on lexical statistics.

Despite these challenges, these datasets constitute the best resource available for computing mean trigram surprisal for lower frequency words, given that their huge size provides better evidence of the range of contexts in which they appear. In the course of the analysis we found two notable issues in
both the Google Books 2012 and Google 1T datasets; we include a brief note on each here to alert others using this or similar corpora.

First, we found extensive evidence of issues with word segmentation in the Hebrew corpus of Google Books 2012. Both our novel extraction stack and Google’s public interface reveal a high proportion of the bound morphemes (שה, נח, ל, נ, ב, ו, ה) listed as separate lexical items, accounting for approximately 20% of the total unigram probability mass.

Second, the web-based nature of the Google 1T corpora strongly influences certain lexical surprisal estimates. For example, “Romanian” has very low trigram surprisal in the Google 1T corpus because it is overwhelmingly encountered in language lists or lists of currencies (22% of instances appear in the context “Polish Portuguese Romanian”). As ever, systematic biases in these datasets may lead to spurious conclusions regarding relationships.