Word lengths are optimized for efficient communication

Steven T. Piantadosi[1], Harry Tily, and Edward Gibson

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited* by Paul Kay, University of California, Stanford, CA, and approved December 17, 2010 (received for review August 23, 2010)

We demonstrate a substantial improvement on one of the most celebrated empirical laws in the study of language, Zipf's 75-y-old theory that word length is primarily determined by frequency of use. In accord with rational theories of communication, we show across 10 languages that average information content is a much better predictor of word length than frequency. This indicates that human lexicons are efficiently structured for communication by taking into account interword statistical dependencies. Lexical systems result from an optimization of communicative pressures, coding meanings efficiently given the complex statistics of natural language use.

information theory | rational analysis

One widely known and apparently universal property of human language is that frequent words tend to be short. This law was popularized by Harvard linguist George Kingsley Zipf, who observed that "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences" (1).

Zipf theorized that this pattern resulted from a pressure for communicative efficiency. Information can be conveyed as concisely as possible by giving the most frequently used meanings the shortest word forms, much like in variable-length (e.g., Huffman) codes. This strategy provided one key exemplar of Zipf's principle of least effort, a grand "principle that governs our entire individual and collective behavior of all sorts, including the behavior of our language" (2). Zipf's idea of assigning word length by frequency can be maximally concise and efficient if words occur independently from a stationary distribution. However, natural language use is highly nonstationary as word probabilities change depending on their context. A more efficient code for meanings can therefore be constructed by respecting the statistical dependencies between words. Here, we show that human lexical systems are such codes, with word length primarily determined by the average amount of information a word conveys in context. The exact forms of the frequency–length relationship (3, 4) and the distribution of word lengths (5) have been quantitatively evaluated previously. In contrast, information content offers an empirically supported and rationally motived alternative to Zipf's frequency–length relationship.

A lexicon that assigns word lengths based on information content differs from Zipf's theory in two key ways. First, such a lexicon would not be the most concise one possible as it would not shorten highly informative words, even if shorter distinctive wordforms were available. Second, unlike Zipf's system, assigning word length based on information content keeps the information rate of communication as constant as possible (6). A tendency for this type of "smoothing out" peaks and dips of informativeness is known as uniform information density and has been observed in choices made during online language production (7–10). Formally, uniform information density holds that language users make choices that keep the number of bits of information communicated per unit of time approximately constant. For instance, more informative syllables are produced with longer durations than less informative syllables, meaning that speech rate is modulated to prevent communicating too many bits in a short period (7). This idea can be generalized to the design of lexical systems (11, 6): the amount of information conveyed by a word should be linearly related to the amount of time it takes to produce—approximately, its length—to convey the same amount of information in each unit of time. A constant information rate can make optimal use of the speech channel by maximizing the amount of information conveyed, without exceeding the channel capacity of speech or our cognitive systems (12, 13). Thus, lexical systems that assign length according to information content can be communicatively more efficient than those that use frequency.

Importantly, the amount of information conveyed by an instance of a word depends on its context. To formalize this, we can consider two random variables, C for contexts and W for words, with a joint distribution $P(C, W)$ given by the natural statistics of language use. The average amount of information conveyed by a particular word w is given by the following (14):

$$-\sum_{c} P(C = c \mid W = w) \log P(W = w \mid C = c). \qquad [1]$$

Intuitively, this measure corresponds to the expected information conveyed by a randomly chosen instance of w from a large corpus. To see this, note that an instance of w will occur in context $C = c$ with probability $P(C = c \mid W = w)$, and will there convey an amount of information given by $-\log P(W = w \mid C = c)$. When estimated from a corpus, this measure is simply the mean negative log probability of tokens of w:

$$-\frac{1}{N}\sum_{i=1}^{N} \log P(W = w \mid C = c_i),$$

where c_i is the context for the ith occurrence of w and N is the total frequency of w in the corpus.

In general, there are many variables that may count as part of the "context" for the purposes of language processing, including discourse context (15–20), local linguistic context (21–26), syntactic context (27, 28), and more global world knowledge (29–31). All these variables likely influence the probability of any word w, but they are not easily quantified. We chose to approximate $P(W \mid C)$ by using a standard quantitative probabilistic model, an N-gram model, which treats C as consisting only of the local linguistic context containing the previous N − 1 words. This simplification allows Eq. **1** to be estimated cross-linguistically from large corpora, and is an approximation to the true information content of a word that has been used extensively in psycholinguistics. In addition, there are several possible ways to measure word length. Here, we primarily use orthographic length because it is readily available from corpora and tends to be highly correlated with both phonetic length and production time. However,

Author contributions: S.T.P., H.T., and E.G. designed research; S.T.P. and H.T. performed research; S.T.P., H.T., and E.G. analyzed data; and S.T.P., H.T., and E.G. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: piantado@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1012551108/-/DCSupplemental.

we also present results in English, German, and Dutch measuring length in number of phonemes and number of syllables.

With these measures, we tested whether average information content or frequency is a better predictor of word length by computing the information conveyed by each word in Czech, Dutch, English, French, German, Italian, Polish, Portuguese, Romanian, Spanish, and Swedish. In each language, we computed Spearman rank correlations between (*i*) information content and length and (*ii*) frequency and length. This measure allowed us to test correlations without making assumptions about the parametric form of the relationship.

Results

The solid and striped bars in Fig. 1 show correlations in the 11 languages between orthographic length and frequency, and orthographic length and information content, as measured by two-gram, three-gram, and four-gram models. Because of the size and form of the Google dataset, these *N*-gram models were not smoothed (although see *SI Text*). Statistical significance was assessed by using *Z*-tests on bootstrapped estimates of the difference between within-language correlations. In the two-gram model, information content is more strongly correlated with length than frequency across all 11 languages ($P < 0.01$, $Z > 2.58$ for each language). The three-gram models show similar patterns, showing significant effects in the predicted direction for 10 of the 11 languages ($P < 0.001$, $Z > 3.30$), with the exception of Polish, in which the trend is not significant ($P > 0.47$, $Z = 0.71$). The four-gram results show effects in the predicted direction for seven of 11 languages; all differences are significant ($P < 0.001$, $|Z| > 4.91$) except for Swedish ($P = 0.29$, $Z = 1.05$). The decreasing consistency of results for higher *N*-gram sizes likely results from increased estimation error caused in part by our use of Google *N*-gram counts, as information content in large *N*-gram models is more difficult to estimate than frequency, and estimation error deflates correlations (Table S1). In general, we take three-gram data—with 10 of 11 languages showing the predicted result—as the most representative finding, because the three-gram results have the highest overall correlations.

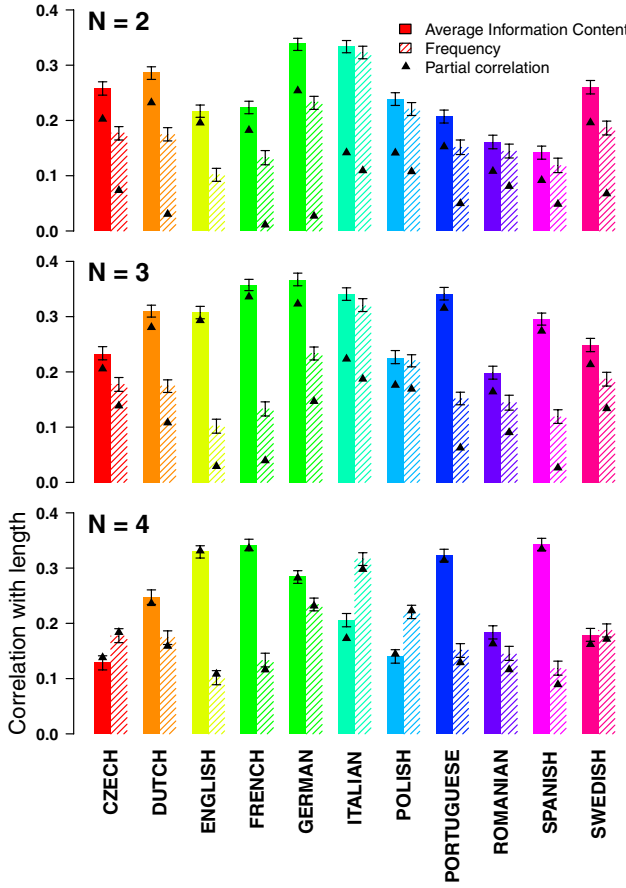

Fig. 1. Correlations between information content and word length (solid) and between frequency (negative log unigram probability) and word length (striped) for two-gram, three-gram, and four-gram models. Error bars show bootstrapped 95% confidence intervals. Triangles show partial correlations (frequency and length partialing out information content; information content and length partialing out frequency).

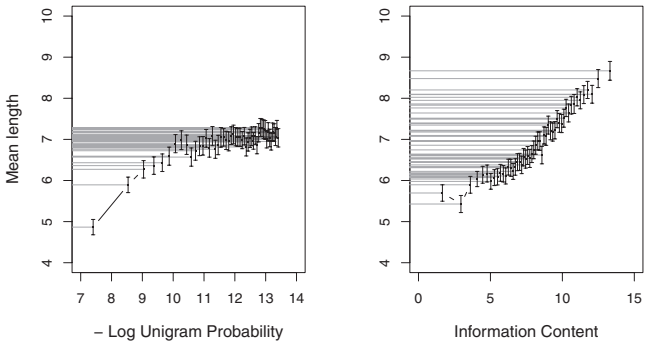

Fig. 2. Relationship between frequency (negative log unigram probability) and length, and information content and length. Error bars represent SEs and each bin represents 2% of the lexicon.

Because frequency and information content are not independent, we also computed how well each variable predicts length controlling for effects of the other. The triangles in Fig. 1 show partial correlations: frequency and length, partialing out information content; and information content and length, partialing out frequency. In several languages, the partial correlation of frequency is close to zero, meaning that the effect of frequency is largely a result of the fact that it is correlated with information content. In most languages, the partial correlation of length with information content, controlling for frequency, is larger even than the absolute correlation between length and frequency.

Information content and frequency exhibit qualitatively different relationships with length. Fig. 2 shows the mean length for (binned) frequency and information content in English, which is typical of the languages here with large effect sizes. The spread in the gray lines illustrate that length is more predictable from information content than from frequency. Frequency does not predict word length substantially for lower-frequency words: words in low unigram probability bins have approximately the same average length. In contrast, length varies as a function of information over almost its entire range of values, with the exception of just the few words with the very lowest information content. This pattern is observed in 11 of 11 languages and indicates that information content is not a good predictor of length for the 5% to 20% least informative (and typically also most frequent) words. This is potentially caused by the fact that, in text on the Internet, many long words occur in highly predictable collocations such as "all rights reserved." These long words are highly predictable, conveying little information, and thus increase the mean length of the least informative words.

To ensure that the results were not driven by artifacts of the Google dataset, we replicated this study in English by using the British National Corpus (BNC) (32), state-of-the-art *N*-gram smoothing techniques, and separate training corpora for $P(W = w|C = c)$ and $P(C = c|W = w)$. This was not possible in all languages because large enough corpora are not available.[†] In

[†]Europarl (33), for instance, contains only approximately 50 million words per language—approximately one 2,500th the size of the non-English Google dataset. Bootstrapping revealed this to be too small to yield reliable estimates of information content.

these data, the correlation between frequency (negative log unigram probability) and length was 0.121. The correlation for information content was 0.161 with the use of two-gram models and 0.168 with the use of three-gram models. Both *N*-gram correlations were significantly higher than the frequency correlation ($P < 0.001$, $Z > 9.49$). Interestingly, this analysis reveals no partial effects of frequency: the correlation of frequency partialing out information content was negative (−0.009 for two-gram and −0.009 for thee-gram models), but not significantly so ($P > 0.13$, $Z < -1.48$ for each comparison). The correlations for two-gram and three-gram models partialing out frequency were 0.106 and 0.117, respectively, both of which were significantly greater than zero ($P < 0.001$, $Z > 16.90$). These results show qualitatively similar results to the Google data. The numerical pattern of correlations differs somewhat from the Google data, likely because the BNC contains only 100 million words, only one 10,000th the size of the Google dataset for English.

In English, German, and Dutch, detailed phonetic information is available from the CELEX database (34). For these languages, we also computed the correlations measuring length in number of phones and also number of syllables. This resulted in two comparisons between information content and frequency for each language and each *N*-gram model. For two-gram and three-gram models, eight of eight English and Dutch correlations all are significant in the predicted direction ($P < 0.001$, $Z > 9.19$ for each comparison), with information content a better predictor of length than frequency. For German, phonetic length and syllable length trend in the right direction for two-gram models, but only the latter is statistically significant ($P < 0.001$, $Z = 5.22$). For three-gram models, German is significant in the wrong direction for length measured in phones ($P < 0.001$, $Z = -4.42$), but not significant for length measured in syllables ($P = 0.29$, $Z = 1.04$). Like the orthographic results, the four-gram results are more mixed, with two of two significant effects in the predicted direction for English ($P < 0.001$, $Z > 12.43$), zero of two significant effects in the predicted direction for German ($P < 0.001$, $Z > -5.49$), and marginally significant effects each way for Dutch. In general, these result show similar patterns to the orthographic measure of length used here earlier. Indeed, in several of the other languages studied (e.g., Spanish, Italian, Portuguese), there is a close relationship between the orthography and phonetics, meaning that these results likely generalize to phonetic length as well.

Discussion

Our results indicate that information content is a considerably more important predictor of word length than frequency. These results hold across languages with varying degrees of morphological inflection, indicating that the results are generally true regardless of how languages use morphemes to encode information. We expect that our basic theory—that information content predicts word length—should generalize to languages with a more free word order, although it is not clear that *N*-gram contexts could be used to estimate information content in such languages.

One likely mechanism for how the lexicon comes to reflect predictability is that information content is known to influence the amount of time speakers take to pronounce a word: words and phones are given shorter pronunciations in contexts in which they are highly predictable or convey less information (7, 13, 27, 28, 35–38). If these production patterns are lexicalized, word length will come to depend on average informativeness.

Our results show significant partial effects of frequency, and it may be that lexicons assign word lengths based on the effective information content for the language processing system. People's expectations about upcoming words are influenced by word frequency, even controlling for information content and other relevant factors (39). This means that a communication system which is optimal for people's probabilistic expectations may include frequency effects. Alternatively, the frequency effects might be illusory: a model of information content that takes into account more than *N*-gram contexts—for instance, global discourse context or world knowledge—may explain all the variance frequency explains.

In general, we take these results to necessitate a revision of Zipf's view of frequency and lexical efficiency. The most communicatively efficient code for meanings is one that shortens the most predictable words—not the most frequent words. Human lexicons are these types of information—theoretically efficient codes for meaning, constructed to respect the interword statistical dependencies of normal language use.

Materials and Methods

In Fig. 1, we approximated the information content of each word by using an unsmoothed *N*-gram model trained on data from Google (40). The size and form of the Google dataset makes standard smoothing techniques infeasible and also likely unnecessary, as we study only the most frequent words and these would not be highly affected by smoothing. To uniformly define lexicons across languages, we extracted the 25,000 most frequent strings in the Google dataset for each language that also occurred in the OpenSubtitles section of the OPUS Corpus (41). This was necessary because the Google dataset was gathered from the internet and contains many frequent strings that are not words (e.g., "aaaaa" and "aaaaaa"). The most frequent words in the dataset were used because information content can be estimated reliably only from a large number of occurrences. For each language, we computed the Spearman correlation between frequency (negative log unigram probability[‡]) and length, information content and length, and the corresponding partial correlations. We present Spearman correlations because length is related nonlinearly to frequency and information content, although Pearson correlations give nearly identical results.

For our replication on the BNC (32), we estimated Eq. **1** by splitting the BNC into two halves. The first half was used to train an *N*-gram model for computing $P(W = w \mid C = c)$ by using the SRILM toolkit (42) with modified Kneser–Ney smoothing (43). The *N*-gram model was evaluated on the second half of the corpus to estimate Eq. **1**. As with the Google data, we evaluated the 25,000 most frequent words that also occurred in the OpenSubtitles section of the OPUS Corpus.

ACKNOWLEDGMENTS. We thank Roger Levy and Florian Jaeger for helpful discussion about this work. This work was supported by a National Science Foundation graduate research fellowship and a Social, Behavioral & Economic Sciences Doctoral Dissertation Research Improvement Grant (to S.T.P.), as well as National Science Foundation Grant 0844472 (to E.G.).

[‡]This transformation puts frequency on the same scale as information content, bits, and corresponds to a measure of the information conveyed by a word if its context is not known.

1. Zipf G (1936) *The Psychobiology of Language* (Routledge, London).
2. Zipf G (1949) *Human Behavior and the Principle of Least Effort* (Addison-Wesley, New York).
3. Sigurd B, Eeg-Olofsson M, van de Weijer J (2004) Word length, sentence length and frequency–Zipf revisited. *Studia Linguistica* 58:37–52.
4. Strauss U, Grzybek P, Altmann G (2006) Word length and word frequency. *Contributions to the Science of Text and Language: Text, Speech and Language Technology, Vol. 31*, ed Grzybek P (Springer, Berlin), pp 277–294.
5. Grzybek P (2006) History and methodology of word length studies. *Contributions to the Science of Text and Language: Text, Speech and Language Technology, Vol. 31*, ed Grzybek P (Springer, Berlin), pp 15–90.
6. Manin D (2006) Experiments on predictability of word in context and information rate in natural language. *J Inform Processes* 6:229–236.
7. Aylett M, Turk A (2004) The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang Speech* 47:31–56.
8. Jaeger TF (2006) Redundancy and syntactic reduction in spontaneous speech. PhD thesis (Stanford University, Stanford, CA).
9. Levy R, Jaeger TF (2007) Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems 19*, eds Schölkopf B, Platt J, Hoffman T (MIT Press, Cambridge, MA), pp 849–856.
10. Jaeger TF (2010) Redundancy and reduction: Speakers manage syntactic information density. *Cognit Psychol* 61:23–62.
11. Piantadosi S, Tily H, Gibson E (2009) The Communicative Lexicon Hypothesis. *The 31st Annual Meeting of the Cognitive Science Society (CogSci09)* (Cognitive Science Society, Austin, TX), pp 2582–2587.

12. Frank AF, Jaeger TF (2008) Speaking rationally: Uniform information density as an optimal strategy for language production. *The 31st Annual Meeting of the Cognitive Science Society*, eds Love BC, McRae K, Sloutsky VM (Cognitive Science Society, Austin, TX), pp 939–944.
13. van Son R, Pols L (2003) How efficient is speech? *Proc Inst Phonetic Sci* 25:171–184.
14. Cohen Priva U (2008) Using information content to predict phone deletion. *Proceedings of the 27th West Coast Conference on Formal Linguistics*, eds Abner N, Bishop J (Cascadilla Proceedings Project, Somerville, MA), pp 90–98.
15. Altmann G, Steedman M (1988) Interaction with context during human sentence processing. *Cognition* 30:191–238.
16. Altmann G, Garnham A, Dennis Y (1992) Avoiding the garden path: Eye movements in context. *J Mem Lang* 31:685–712.
17. Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* 268:1632–1634.
18. Spivey MJ, Tanenhaus MK (1998) Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *J Exp Psychol Learn Mem Cogn* 24:1521–1543.
19. Kaiser E, Trueswell JC (2004) The role of discourse context in the processing of a flexible word-order language. *Cognition* 94:113–147.
20. Grodner D, Gibson E, Watson D (2005) The influence of contextual contrast on syntactic processing: evidence for strong-interaction in sentence comprehension. *Cognition* 95:275–296.
21. Frazier L (1979) *On Comprehending Sentences: Syntactic Parsing Strategies* (Univ of Connecticut, Storrs, CT).
22. Jurafsky D (1996) A probabilistic model of lexical and syntactic access and disambiguation. *Cogn Sci* 20:137–194.
23. Gibson E (1998) Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68:1–76.
24. McDonald SA, Shillcock RC (2003) Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychol Sci* 14:648–652.
25. Frisson S, Rayner K, Pickering MJ (2005) Effects of contextual predictability and transitional probability on eye movements during reading. *J Exp Psychol Learn Mem Cogn* 31:862–877.
26. Levy R (2008) Expectation-based syntactic comprehension. *Cognition* 106:1126–1177.
27. Gahl S, Garnsey S (2004) Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80:748–775.
28. Tily H, et al. (2009) Syntactic probabilities affect pronunciation variation in spontaneous speech. *Lang Cognition* 1:147–165.
29. Trueswell J, Tanenhaus M, Garnsey S (1994) Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *J Mem Lang* 33:285–318.
30. Altmann GT, Kamide Y (1999) Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73:247–264.
31. Kamide Y, Altmann G, Haywood S (2003) The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *J Mem Lang* 49: 133–156.
32. Clear J (1993) *The British National Corpus* (MIT Press, Cambridge, MA).
33. Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit (MT Summit)* (Asia-Pacific Association for Machine Translation, Aichi, Japan), pp 79–86.
34. Baayen HR, Piepenbrock R, Gulikers L (1995) *The CELEX Lexical Database. (Release 2)* (Univ of Pennsylvania, Philadelphia).
35. Lieberman P (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Lang Speech* 6:172–187.
36. Jespersen O (1922) *Language: Its Nature, Development, and Origin* (Henry Holt, New York).
37. Fowler C, Housum J (1987) Talkers signaling of new and old words in speech and listeners' perception and use of the distinction. *J Mem Lang* 25:489–504.
38. Gregory M, Raymond W, Bell A, Fosler-Lussier E, Jurafsky D (1999) The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistics Society 35 (CLS-99)* (Univ of Chicago, Chicago), pp 151–166.
39. Demberg V, Keller F (2008) Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109:193–210.
40. Brants T, Franz A (2009) *Web 1T 5-Gram, 10 European Languages, Version 1* (Linguistic Data Consortium, Philadelphia).
41. Tiedmann J (2009) News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing*, eds Nicolov N, Bontcheva K, Angelova G, Mitkov R (John Benjamins, Amsterdam), Vol V, pp 237–248.
42. Stolcke A (2002) SRILM - an extensible language modeling toolkit. *Proceedings of the Seventh International Conference on Spoken Language Processing* (ICSLP/Interspeech, Denver), Vol 3, pp 901–904.
43. Chen SF, Goodman J (1999) An empirical study of smoothing techniques for language modeling. *Comput Speech Lang* 13:359–393.