LETTER

Information content and word frequency in natural language: Word length matters

For centuries, scientists have attempted to uncover commonalities that underlie the structure of human languages (1). In a recent issue of PNAS, Piantadosi et al. (2) reported an exciting finding with respect to one unique type of language universal. The authors empirically demonstrated that word length strongly correlated with information content across 11 distinct natural languages. This finding is remarkable in that it directly contravenes Zipf's principle of least effort, which holds that language optimizes communication by spontaneously truncating words as a function of their relative frequency (e.g., television → TV). Simply stated, more frequent words tend to be shorter, and this length–frequency relationship serves an adaptive function of communicative parsimony.

Piantadosi et al. (2) derived evidence for their hypothesis by mining a massive trillion-word corpus of cross-linguistic language samples from the world wide web. Piantadosi et al. (2) pose an alternative hypothesis to Zipf's widely held view, namely that the information content of a word is a stronger predictor of word length than its relative frequency of occurrence within a given language. For language researchers, the significance of this finding cannot be overstated. In a *Nature News* press release (3), linguist Roger Levy remarked that these results "may now supply the largest leap forward in 75 years in understanding how the evolution of words is governed by the efficiency with which they can be used to communicate" (3).

Information content is one aspect of a word. Two other aspects of word meaning pose obstacles for a clean interpretation of the length–content hypothesis advanced by Piantadosi and colleagues. These include the syntactic dimension of grammatical class (e.g., nouns vs. verbs) and the semantic dimension of word meaning (i.e., word concreteness). With respect to grammatical class, verbs tend to be longer than nouns across many natural languages (4). With respect to word concreteness, word length provides a cue for the sensory salience of a word. That is, longer words tend to denote more abstract ideas. Consider, for example, the concrete noun, *friend*. It is possible to inflect friend and transform its root into an abstract idea (e.g., *friendliness*) (5). This morphological transformation is also evident in other languages (e.g., the German root schade can be inflected to form an abstract concept such as *schadenfreude*). As such, word length can provide at least statistically reliable markers for distinctions such grammatical class and concreteness. Therefore, one difficulty with interpreting these findings regards construct validity of the dependent measure (information content). That is, it is unclear whether verbs (e.g., eat) and abstract nouns (e.g., honesty) convey more information content than their concrete noun counterparts (e.g., dog) and whether this trend holds across unrelated languages. Thus, there remains considerable ambiguity as to the nature of what constitutes the information content of a word and how this might be reliably measured. Despite such ambiguity, Piantadosi et al. (2) have potentially unveiled an intriguing domain for cross-linguistic research. Perhaps even more importantly, the authors have paved the way for a paradigmatic shift in how linguists view the relation between sound and meaning in natural languages.

***Jamie Reilly*[a,1] *and Jacob Kean*[b]**

[a]*Department of Speech, Language, and Hearing Sciences, University of Florida, Gainesville, FL 32611-7420; and* [b]*Department of Physical Medicine and Rehabilitation, Indiana University School of Medicine, Indianapolis, IN 46202-3082*

1. Hauser MD, Chomsky N, Fitch WT (2002) The faculty of language: What is it, who has it, and how did it evolve? *Science* 298:1569–1579.
2. Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proc Natl Acad Sci USA* 108:3526–3529.
3. Ball P (2011) How words get the message across. *Nature News*, 10.1038/news.2011.40. Available at http://www.nature.com/news/2011/110124/full/news.2011.40.html.
4. Langenmayr A, Gözütok M, Gust J (2001) Remembering more nouns than verbs in lists of foreign-language words as an indicator of syntactic phonetic symbolism. *Percept Mot Skills* 93:843–850.
5. Reilly J, Kean J (2007) Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cogn Sci* 31:1–12.

Author contributions: J.R. and J.K. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: jjreilly@phhp.ufl.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1103035108