Reply to Reilly and Kean: Clarifications on word length and information content

Citation: Piantadosi, S. T., H. Tily, and E. Gibson. “Reply to Reilly and Kean: Clarifications on word length and information content.” Proceedings of the National Academy of Sciences 108 (2011): E109-E109. ©2011 by the National Academy of Sciences.

Published Version: http://dx.doi.org/10.1073/pnas.1103550108

Publisher: National Academy of Sciences (U.S.)

Permanent Link: http://hdl.handle.net/1721.1/65887

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

LETTER

Reply to Reilly and Kean: Clarifications on word length and information content

First, we disagree with Reilly and Kean (1) that our results on word length (2) contradicted Zipf's principle of least effort. Our findings were in the same spirit, except that we measured effort in a more principled way than Zipf could have (2). Assigning word length by information content is least effort under an assumption of a superlinear relationship between effort and information content (3), and it is optimal under a desire to stay just under the channel capacity of linguistic systems (4).

Reilly and Kean (1) bring up the important issue of how we quantified information content (2). In one sense, negative log probability is an impoverished notion of information, because it ignores other aspects of meaning, such as a word's denotation, connotation, significance, etc. Mathematically, however, it has a precise meaning, quantifying the number of bits of information that it would take an optimal code to convey that a given word occurs. Testing whether the lexicon matches the predictions of a statistically optimal code follows a tradition in cognitive science of rational analysis (5), where human behavior is explained in terms of what would be optimal given the problem to be solved. We argued that lexicons would be communicatively optimal if the average number of bits of information conveyed per unit time according to an optimal code—a word's average negative log probability—was kept constant (2).

Reilly and Kean (1) state that one problem with our analysis is that it was unclear whether verbs and abstract nouns convey more information than concrete nouns. This is not a problem with our theory—it is an empirical prediction. In fact, this prediction is borne out using words in CELEX with unambiguous parts of speech that have concreteness ratings in the Medical Research Council (MRC) psycholinguistic database. Verbs and abstract nouns (defined by a median split on MRC's noun concreteness ratings) have a mean length of 7.24 characters and a mean information content of 8.23. Concrete nouns have a mean length of 5.99 characters and an information content of 7.52. Thus, verbs and abstract nouns are longer and do convey more information than concrete nouns. Both length and information content differences were significant with a Wilcoxon rank sum test ($P < 0.001$). Importantly, our theory made more fine-grained predictions than this, predicting word length across a range of information content values (figure 2 in ref. 2) and not just for this binary distinction (2).

Finally, we see no conflict between information content being a major determinant of word length and morphological processes. Derivational word forms often get shortened to have no apparent morphology, as in "exam" for "examination" and "ad" for "advertisement." If words that are shortened like this tend to convey little information, then the remaining words will be long and morphologically complex, with high-information content. This predicts the abstractness patterns that Reilly and Kelly (1) describe, because information content correlates negatively with concreteness in the MRC database ($R = -0.24$, $P < 0.001$), meaning the words that are not shortened will tend to be abstract.

Steven T. Piantadosi*[1], *Harry Tily, and Edward Gibson
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

1. Reilly J, Kean J (2011) Information content and word frequency in natural language: Word length matters. *Proc Natl Acad Sci USA* 108:E108.
2. Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proc Natl Acad Sci USA* 108:3526–3529.
3. Levy R (2005) *Probabilistic Models of Word Order and Syntactic Discontinuity* (Stanford University Press, Palo Alto, CA).
4. Frank A, Jaeger TF (2008) Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society,* eds Love BC, McRae K, Sloutsky VM (Washington, DC), pp 939–944.
5. Chater N, Oaksford M (1999) Ten years of the rational analysis of cognition. *Trends Cogn Sci* 3:57–65.

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

The authors declare no conflict of interest.

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