Supporting Information

Piantadosi et al. 10.1073/pnas.1012551108

SI Text

We additionally evaluated results by using several different smoothing methods on the BNC data set. Table S1 shows Spearman correlations between length and information content as measured by three *N*-gram smoothing methods.

Because, generally, Kneser–Ney is a better method than Witten–Bell, which is in turn better than unsmoothed counts (1), these results demonstrate that better smoothing methods show higher correlations between information content and length. This is expected as good smoothing methods provide less noisy estimates of information content. Note that the one-gram model is a frequency measure, and its correlations are considerably less sensitive to the smoothing method, likely because unigram counts are less sparse. In fact, the one-gram model with the best smoothing method has lower correlations than the two-, three-, and four-gram models with the worst smoothing method. In general, these results indicate that we have likely underestimated the true correlation between length and information content because we did not use sophisticated smoothing methods on the Google data: our main analyses are conservative with respect to our hypothesis of higher correlations for information content.

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

Table S1. Spearman correlations between length and information content as measured by three *N*-gram smoothing methods

	Model			
Method	One-gram	Two-gram	Three-gram	Four-gram
Kneser–Ney	0.121	0.161	0.168	0.170
Witten–Bell	0.118	0.160	0.162	0.163
Unsmoothed	0.118	0.144	0.148	0.150