Info/information theory: Speakers actively choose shorter word in predictable contexts

Kyle Mahowald & Evelina Fedorenko (Massachusetts Institute of Technology), Steven Piantadosi (University of Rochester), & Edward Gibson (Massachusetts Institute of Technology)

kylemaho@mit.edu

Uniform information density; Word length; Corpus analysis; Behavioral study; Efficiency; Surprisal

Piantadosi, Tily & Gibson (2011, henceforth PTG) demonstrated that the average information conveyed by a word is a significant predictor of its length. However, PTG did not examine length effects within part of speech categories nor attempt to control for meaning. Here, we show that language users actively choose short forms of synonym pairs (e.g., math/mathematics, exam/examination) in predictive contexts, a result consistent with uniform-information density (UID) accounts of language production (Genzel & Charniak 2002, van Son & Pols 2003, Aylett & Turk 2004, Jaeger 2006, Frank & Jaeger 2008, and Jaeger 2010). The use of near-synonyms that vary in length ensures that the observed effects hold for content words of the same part of speech and meaning. This work extends previous work on UID by showing that information rate can be manipulated not just through phonetic reduction (as in Bell, et al. 2003), syntactic factors (such as *that* omission, as in Levy & Jaeger, 2007), and choice of contractions (Frank & Jaeger 2008), but through active selection of noun word forms.

In a corpus study, we first used the data from PTG (a three-gram model from the Google corpus) to obtain average surprisal estimates for 22 long/short word pairs. Replicating PTG with this paired sample, the mean surprisal for long forms (9.21) was significantly higher than the mean surprisal for short forms (6.90) ($P = .004$ by Wilcoxon signed rank test). Of the 22 pairs, 18 showed higher average surprisal for the long form than for its short counterpart. A linear regression revealed that this difference held even while controlling for frequency: an intercept of 1.45 ($t = 2.76$, $P = 0.01$) indicated that, when there is no difference in frequency between the forms, the mean surprisal of long forms is 1.45 higher than of short forms.

To test whether participants actively choose short forms in predictive contexts, we presented participants with forced-choice sentence completions in which they had to choose between the short and long form of a word pair (exam/examination) based on which sounded more natural. The manipulation of interest was whether the context provided by the sentence was predictive of the missing final word (supportive-context condition) or was non-predictive (neutral-context condition), as in the sample item below. The order of the answer choices (i.e., whether the short form or long form was listed first) was balanced across participants and items.

(1) **supportive-context:** Susan was very bad at algebra, so she hated...
1. math 2. mathematics
neutral-context: Susan introduced herself to me as someone who loved...
1. math 2. mathematics

In supportive-context sentences, the short form was chosen significantly more often (67%) than in neutral-context sentences (56%). The effect was significant by a mixed-effect logistic regression with both item and participant slopes and intercepts ($P < .01$).

These results indicate that speakers use content words to manipulate information rate, choosing words that optimize communicative efficiency. Moreover, these results suggest that the correlation between word length and informativeness is likely influenced by language production phenomena, where users actively prefer to convey meanings with short forms when the meanings are contextually predictable. We thus conclude that information-theoretic considerations are part of a speaker's knowledge and likely a causal factor in language change.

References

Aylett M, Turk A (2004). The smooth signal redundancy hypothesis. Lang Speech 47:31–56.
Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. The Journal of the Acoustical Society of America, 113, 1001.
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.
Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. Proc. of the 40th Annual Meeting on Assoc. for Comp. Ling. (pp. 199–206).
Jaeger TF (2006) Redundancy and syntactic reduction in spontaneous speech. PhD thesis (Stanford University, Stanford, CA). Jaeger TF (2010) Redundancy and reduction: Speakers manage syntactic information density. Cognit Psychol 61:23–62. 11.
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.
Piantadosi S, Tily H, Gibson E (2011). Word lengths are optimized for efficient communication. Proc Natl Acad Sci USA 108:3526–3529.
van Son R, Pols L (2003). How efficient is speech? Proc Inst Phonetic Sci 25:171–184.