Informativity effects can be probability effects in disguise

Vsevolod Kapatsinski

University of Oregon

Informativity effects on pronunciation reduction show that a particular linguistic unit is reduced not only when it is predictable in the current context, but also when it has usually been predictable in the speaker's experience (Barth, 2019; Cohen Priva, 2017; Seyfarth, 2014; Sóskuthy & Hay, 2017). These effects suggest either that speakers store phonetically detailed tokens of a word in memory (Brown, 2004; Bybee, 2002; Sóskuthy & Hay, 2017) or that speakers rationally take listener needs into account (Jaeger & Buz, 2017; Levshina, 2022).

The present paper argues for caution in interpreting effects of informativity. Informativity of a word is (the inverse of) its average predictability. Consequently, informativity can improve an estimate of a word's probability in context: as proposed by hierarchical / mixed-effects models (Gelman & Hill, 2006; Barth & Kapatsinski, 2018), when an estimate of a word's predictability in context is unreliable – because the context is rare – that estimate can be improved by taking into account the word's predictability across contexts (informativity). Thus, even if the word's pronunciation is determined entirely by its probability in the current context, informativity should significantly predict pronunciation, above and beyond probability in context, insofar as estimates of probability in context are noisy.

We demonstrate this through simulation of durations in a sample of words from the Switchboard corpus of American English conversations (Deshmukh et al., 1998); specifically, words following disfluencies in Harmon & Kapatsinski (2021). This context is useful because it makes it less ambiguous whether forward or backward transitional probabilities should matter (cf., Seyfarth, 2014) – as the speaker is still trying to access the target word, it is unlikely they have preplanned the words that follow. We fit a linear regression model of log word duration as a function of log forward transitional probability (FTP). This model explains 38% of variance in log word duration (N=5060 tokens). We also fit a model with both log FTP and informativity, which explains 43% of the variance. The effect of informativity is highly significant ($t < -18$). The question is whether such a strong effect could emerge from noise in probability estimates.

To address this question, we repeatedly generate durations from a model in which 38% of the variance in log duration is predicted by log FTP as in the corpus but the rest is normally-distributed noise. For each simulation, the estimated FTP of each word in each context is generated by sampling from a binomial distribution with success probability drawn from the real corpus and the number of trials matching context frequency (i.e., preceding word frequency). The binomial sampling process is what creates noise in probability estimates. We compute informativity and fit a regression model in which log duration (really a function of true FTP) is a function of the estimated log FTP and estimated informativity.

Informativity is significant in the expected direction in 100% of these simulations (even though FTP is the only true predictor). Therefore, a significant effect of informativity on top of an effect of probability does not show that there is a true effect of informativity. This result appears to contradict the results of Cohen Priva and Jaeger (2018), who conclude that segment probability effects are unlikely to account for segment informativity effects. However, lexical contexts are far less frequent than segmental contexts, which increases noise in probability estimates and therefore leaves room for informativity to step in.

The effect of informativity observed in the real data is larger than in the simulated data (Figure 1), suggesting that there is a true effect of informativity in the corpus. We recommend performing such simulations to determine whether an observed effect of informativity is really a probability effect in disguise, as about 2/3 of the effect of informativity in the original model is artifactual.

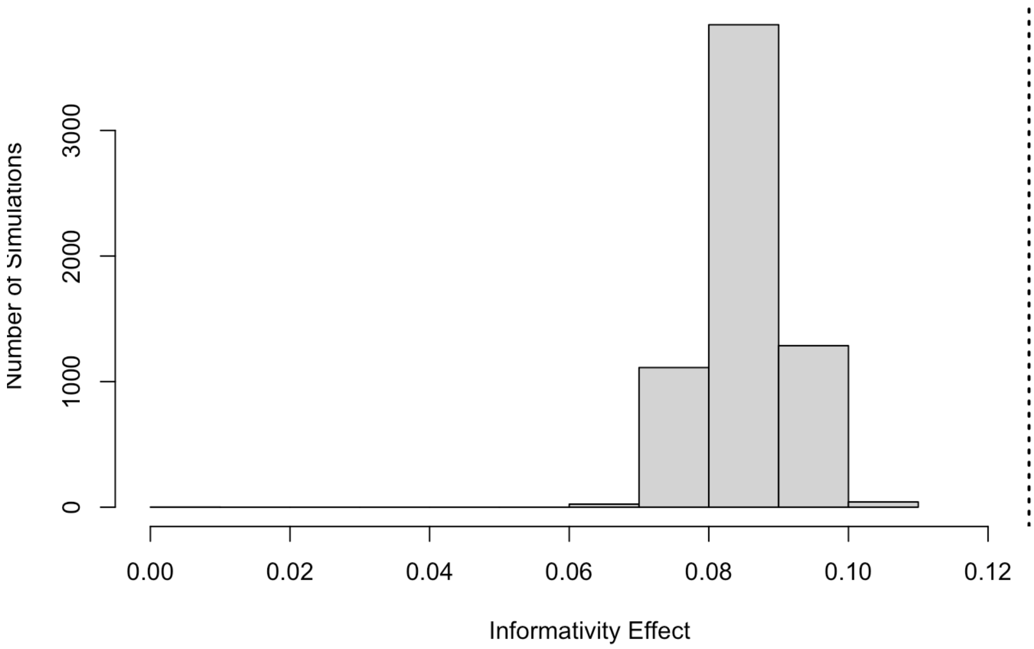

Figure 1. Informativity effects in simulated samples where there is no informativity effect (distribution) vs. the informativity effect in the real data (dashed vertical line).

References:

Barth, D. (2019). Effects of average and specific context probability on reduction of function words BE and HAVE. *Linguistics Vanguard*, *5*(1), 20180055.

Barth, D., & Kapatsinski, V. (2018). Evaluating logistic mixed-effects models of corpus-linguistic data in light of lexical diffusion. *Mixed-effects regression models in linguistics*, 99-116.

Brown, E. L. (2004). *Reduction of syllable-initial /s/ in the Spanish of New Mexico and Southern Colorado: A usage-based approach*. Doctoral Dissertation, University of New Mexico.

Bybee, J. B. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14, 261–290.

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language*, 93, 569-597.

Cohen Priva, U., & Jaeger, T. F. (2018). The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard*, *4*(s2), 20170028.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., & Picone, J. (1998, November). Resegmentation of SWITCHBOARD. In *ICSLP*.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Harmon, Z., & Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological Review*, *128*(6), 1112-1144.

Jaeger, T. F., & Buz, E. (2017). Signal reduction and linguistic encoding. *The handbook of psycholinguistics*, 38-81.

Levshina, N. (2022). *Communicative efficiency*. Cambridge University Press.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140-155.

Sóskuthy, M., & Hay, J. (2017). Changing word usage predicts changing word durations in New Zealand English. *Cognition*, *166*, 298-313.