Good-enough for all intensive purposes: Eggcorns and noisy channel processing

Gwendolyn Rehrig (glrehrig@ucdavis.edu) and Fernanda Ferreira (UC Davis)

Linguistic communication occurs over a noisy channel that can distort the signal (Gibson et al., 2013; Levy, 2008); these distortions may be rational such that the distorted interpretation is more plausible to the receiver than the original message (Ferreira, 2003; Ferreira & Lowder, 2016). Psycholinguistics has largely overlooked a form of naturalistic data that may inform language processing models: eggcorns. Eggcorns are misperceptions of a source word or phrase (e.g., *up and coming* → *up incoming*; Liberman, 2003) that become codified in the lexicon—as evidenced by their repeated usage without self-correction—suggesting eggcorns do not register as errors to the speaker. Although eggcorns do not constitute sentences by themselves, they can be multiword sequences (18% of eggcorns in our dataset), and often occur in sentential contexts that help accommodate the mistaken forms. We posit that eggcorns may be ’good-enough’ representations that approximate the source phrase signal well, can substitute for the source phrase in conversation (eggcorns are usually detected in written form), and may arise from rational language processing (Fig. 1). The current corpus study analyzes the characteristics of attested eggcorns in the context of noisy channel and good-enough language processing.

Method. We scraped 632 unique entries from The Eggcorn Database (Waigl, 2005). Syllables in each source and eggcorn were counted, and the difference in syllable counts was computed. Levenshtein distance between IPA transcriptions of the source phrase and resulting eggcorn approximated phonological similarity. Each pair was automatically transcribed to IPA using the ‘eng_to_ipa‘ package in Python. Semantic relatedness and frequency were obtained from ConceptNet 5 (Speer et al., 2017) and COCA (Davies, 2008-), respectively. To assess whether frequent words form eggcorns, the difference in log frequency between the eggcorn and its source phrase was calculated. Pairs with a Levenshtein distance of 0 (misspellings; N = 146) and pairs for which the source and eggcorn were not both present in either ConceptNet (N = 17) or COCA (N = 73) were excluded. The remaining 396 pairs were analyzed.

Results. The number of syllables were equal in the majority of the pairs (93%; N = 370); few of the eggcorns either added (4%, N = 17) or deleted (2%, N = 9) a syllable. Levenshtein distance for most source-eggcorn pairs (58%, N = 229) was 1; an additional 31% (N = 121) had a Levenshtein distance of 2 (Fig. 2). Semantic relatedness between source and eggcorn was low (M = 0.23, SD = 0.29), though 8% were synonymous (relatedness = 1), and the difference in log frequency was negative on average ($M = {}^{-}0.76$, SD = 3.24), indicating eggcorns were less frequent than their corresponding source. We conducted an ordered probit regression using Levenshtein distance as the dependent variable to characterize the relationship between phonological similarity, semantic relatedness, and the change in log frequency from source phrase to eggcorn. Larger Levenshtein distance was associated with greater relatedness (β = 1.42, t = 3.22, p = .001) and negative changes in frequency such that eggcorns were less frequent than sources ($\beta = {}^{-}0.10$, $t = {}^{-}2.52$, p = 0.01), and there was a marginal interaction between relatedness and frequency change (β = 0.20, t = 1.95, p = 0.05). The results suggest eggcorns tend to closely match the source phrase in sound, but may compromise sound similarity to better fit the context.

Eggcorns overwhelmingly matched the sound of the source signal at the expense of both frequency and semantic similarity. However, when phonological similarity was low, semantic relatedness was higher, suggesting a trade-off when the closest sound match does not fit the context well. We suggest that speech segmentation processes optimize first for similarity to the source signal and second for fit with the surrounding context. These processes operate in a good-enough fashion that is faithful to the input signal most of the time, but occasionally can deviate from the input in principled ways. We suggest that psycholinguists should take eggcorns seriously as naturalistic data points that can inform theories of language processing.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): One billion words, 1990-2019. https://www.english-corpora.org/coca/

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203.

Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good-enough language processing. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 217–247). Academic Press.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243.

Liberman, M. (2003). Egg corns: Folk etymology, malapropism, mondegreen, ??? http://itre.cis.upenn.edu/~myl/languagelog/archives/000018.html

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of AAAI 31*, 4444–4451.

Waigl, C. (2005). The eggcorn database. https://eggcorns.lascribe.net/

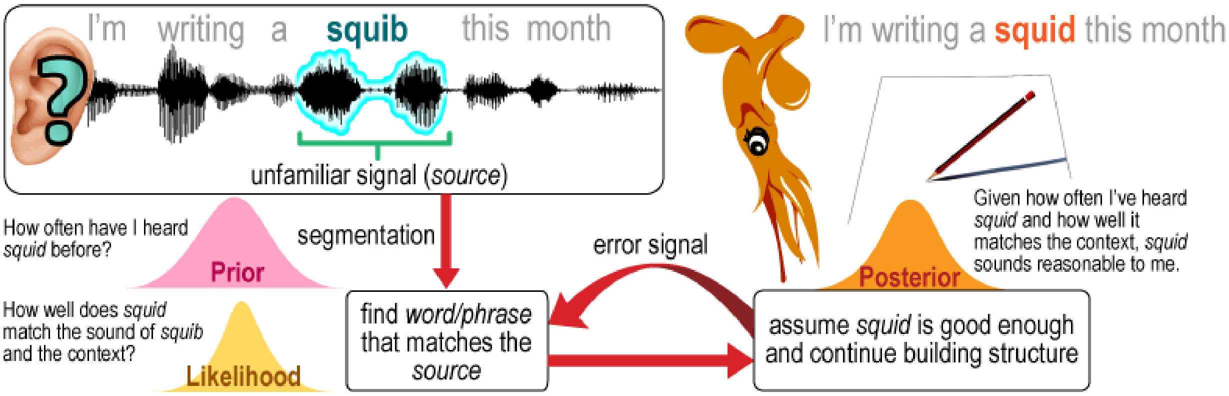
I'm writing a squib this month
unfamiliar signal (*source*)
How often have I heard *squid* before?
Prior
segmentation
How well does *squid* match the sound of *squib* and the context?
Likelihood
find *word/phrase* that matches the *source*
error signal
I'm writing a squid this month
Given how often I've heard *squid* and how well it matches the context, *squid* sounds reasonable to me.
Posterior
assume *squid* is good enough and continue building structure

Figure 1. Schematic illustrating how unfamiliar linguistic signal (*squib*) could be misacquired as an eggcorn (*squid*).

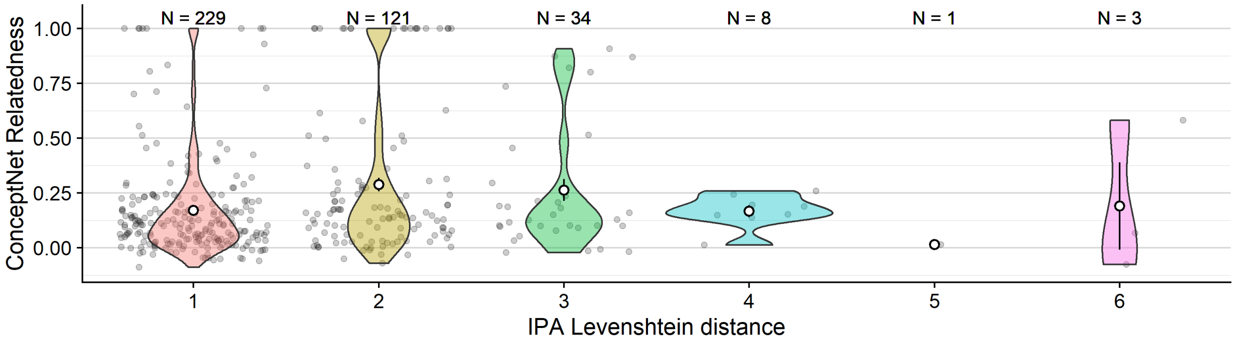

Figure 2. Scatter violin plots showing ConceptNet relatedness values (points, y-axis) plotted against the Levenshtein distance between source and eggcorn IPA transcriptions (violins, x-axis). White points superimposed over each violin plot indicate the mean and standard error.