Predicting collocates: Task effects, chunk frequency, and association measures

Mittwoch,
04.03.2020
16:30–17:00
ESA1 O 121

Kyla McConnell, Alice Blumenthal-Dramé
University of Freiburg
kyla.mcconnell@anglistik.uni-freiburg.de, alice.blumenthal@anglistik.uni-freiburg.de

Comprehenders may track distributional statistics online and use them to inform predictions (Kuperberg & Jaeger 2016). Consider the modifier-noun collocation *vast majority*: where *vast* appears, there is a strong likelihood that *majority* will follow. Several metrics have been put forward to quantify association strength between words. But can these metrics reflect prediction in online processing? And how is this relationship affected by experimental task?

In a self-paced reading study, we assessed the reading times for the second word in collocated modifier-noun bigrams like *vast majority*. Six of the most common corpus linguistic association scores – MI, MI3, Dice coefficient, T-score, Z-score and log-likelihood – were pitted against predictors of lexical processing cost that are widely established in the psycholinguistic and cognitive linguistic communities: log-transformed forward/backward transition probability and bigram frequency (Evert 2009). 123 native speakers of English read 91 critical sentences and 157 filler sentences. Critical sentences contained one modifier-noun bigram embedded in a neutral sentence head and followed by a three-word spillover region (i.e. Connor was informed about the *great deal* on designer jeans.) Association scores for each bigram were extracted from the British National Corpus.

Reading times to the noun were analyzed with mixed-effects models with one association score per model. Results showed that none of the six traditional corpus linguistic metrics patterned significantly with log-transformed reading times to the noun at the Bonferroni-corrected significance level of 0.005 in the expected direction. However, log backward transition probability and log bigram frequency prove to be realistic predictors of reading times. The significance of bigram frequency provides support for the idea of chunk-level activation, as suggested by usage-based approaches (Arnon & Snider 2010). Concurrently, backward transition probability suggests activation of the bigram's component words, though it suggests backwards integration rather than prediction.

These two metrics were additionally compared across two task conditions: In the control block, comprehension questions had a multiple-choice format. In the task block, questions appeared in a typed free response format. The multiple-choice condition elicited faster overall reading times and the effects of the two metrics were stronger at the critical word. In the typed condition, the effect was weaker but longer lasting across the spillover region. It thus seems possible that during slower reading, low-level information tied to individual words is not dismissed as quickly, thereby exerting stronger effects on neighboring words. We argue that insufficient attention to task effects may have partially obscured the cognitive correlates of association scores in similar research.

References: Arnon, I. & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62(1), 67–82. Evert, S. (2009). Corpora and Collocations. In A. Lüdeling & M. Kytö (Eds.), Corpus Linguistics: An International Handbook (Vol. 2). Berlin, New York: Mouton de Gruyter. Kuperberg, G. R. & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? Language, Cognition and Neuroscience, 31(1), 32–59.