

**Bigram attraction reflects prosodic boundaries:
Evidence from some old and one new collocation measure(s)**

Recent work in psycholinguistics has examined how language users possess "rich memory representations" of prior linguistic experiences (Bybee, 2010). These representations are considered to reflect the distributional regularities with which different words co-occur. A co-occurrence of particular words, or collocates, that is attested more frequently than would be expected by chance is said to exhibit greater attraction and thus greater cognitive entrenchment as a unit-like chunk, or collocation (Manning & Schütze, 1999). Recent corpus work has quantified how attracted to or repelled from each other collocates are via various association statistics. Researchers have also begun to search for comprehension- and production-based effects which provide psycholinguistic evidence that higher attraction or "stronger" collocations are indeed mentally represented as unit-like. Specifically, psycholinguistic studies have shown that **syntactically** earlier collocates prime later ones (Durrant and Doherty, 2010; Ellis et al., 2009), while corpus studies have shown that strong collocations exhibit patterns of phonological reduction characteristic of prefabricated wholes (Gregory et al., 1999).

However, to my knowledge no research has examined how prosodic boundaries interact with collocations to evidence their unitary nature. Specifically, it has been repeatedly shown that, in English, prosodic constituents do not split words (e.g., Shukla et al., 2006). Thus, if high-strength collocations are stored more or less like wholes analogous to how individual words are stored as wholes, then prosodic boundaries should be less likely to split them while being more likely to split low-strength collocations. I test this hypothesis here, drawing on data from the Santa Barbara Corpus of Spoken American English (SBC). The SBC is unique in that it is transcribed into intonation units (IU), a prosodic constituent that spans on average four words and is usually characterized by a single, coherent intonational contour (Chafe, 1994). I employ several popular measures of word attraction to quantify strength of adjacent two-word collocations (bigrams) both within and across IU boundaries. These measures include mutual information, the t-score, and log likelihood. In addition, I use Delta P; though rare, this measure is more psychologically realistic to the extent that it encodes directionality of association. The data show that, regardless of the association measure used, there are significant differences between the median bigram attractions within and between IUs (mutual information: med_within=3.28, med_between=1.70, $p < .001$; the t-score: med_within=9.21 med_between=2.63 $p < .001$; log likelihood: med_within=16.00, med_between=3.24, $p < .001$). In addition, the new measure of Delta P brings out the difference between within- and between-IU bigrams well: med_within=.076, med_between=.023, $p < .001$. Thus, IU boundary placement evidences the more entrenched nature of bigrams whose collocates are more highly attracted to each other in that such boundaries show a dispreference for splitting high-strength collocations, while instead being placed so as to split low-strength ones.

These findings demonstrate how the modeling of chunking and rich memory with corpora, using both established and new, more psychologically motivated association measures, can take prosody into account to provide for better resolution of these phenomena]. And since prosodic cues must play a central role in distributionally-based chunking of language (i.e., in children's acquisition), such an inclusion of prosody in the analysis of chunking is an imperative .

REFERENCES

- Bybee, J. (2010). *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.
- Chafe, W. L. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: The University of Chicago Press.
- Durrant, P. & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2), 125-155.
- Ellis, N. C., Frey, E. & Jalkanen, I. (2009). The psychological reality of collocation and semantic prosody. In U. Romer & R. Schulze (Eds.), *Exploring the Lexis-Grammar Interface* (pp. 89-114). Philadelphia: John Benjamins.
- Gregory, M.L., Raymond, W.D., Bell, A., Fosler-Lussier, E. & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *Proceedings of the Chicago Linguistic Society*. Chicago, Illinois.
- Manning, C.D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.
- Shukla, M., Nespor, M. & Mehler, J. (2006). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, 54, 1-32.