Vector-space semantic maps: A data-driven approach to the study of syntactic productivity in diachrony

In usage-based linguistics, syntactic productivity (i.e., the property of syntactic constructions to attract new lexical fillers) is largely taken to depend on the degree of semantic fit between a new item and the priorly witnessed distribution (Barðdal 2008, Bybee 2010, Suttle and Goldberg 2011). However, the reliance of this account on lexical semantic factors can make it hard to operationalize and test on naturally occurring examples.

In this paper, I present a data-driven approach to the study of syntactic productivity in diachrony that uses vector-space representations as a proxy to the meaning of words. Vector-space models approximate the meaning of a word by representing it as a vector in a multidimensional space recording its co-occurrence with other words in a vast text corpus. By computing pairwise distances between the semantic vectors of the items occurring in a construction and feeding them to a multidimensional scaling algorithm that positions the words in a 2-dimensional space, it is possible to visualize the semantic domain of the construction and observe how words in that domain are related to each other.

To illustrate the potential of this method, I present a case study examining the development of the intensifying construction "V *the hell out of* NP" (e.g., *They scared/beat/intimidated the hell out of me*) and its variants, drawing on data from the COHA (Davies 2010). Vector-space semantic maps plotted in four successive 20-year periods from the 1930s to the 2000s reveal that more populated regions of the semantic space are more likely to attract new members at the next epoch, which lines up with the finding that syntactic productivity is driven by semantic similarity and type frequency. They also suggest that token frequency is not a particular strong predictor of productivity, since frequent but isolated verbs only sporadically attract new members, which is likely to be explained by analogical extensions of a limited scope.

References

Barðdal, J. (2008). *Productivity: Evidence from Case and Argument Structure in Icelandic*. Amsterdam: John Benjamins.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Davies, M. (2010-). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/.

Suttle, L. and A. Goldberg (2011). The partial productivity of constructions as induction. *Linguistics* 49 (6): 1237–1269.