Using vector-space models to visualize the semantic distribution of argument structure constructions

Florent Perek

Universität Basel & Princeton University

Most contemporary theories of verb argument structure emphasize the importance of lexical semantic factors. Yet, the meaning of words can be an elusive notion that proves hard to operationalize. The research presented in this talk explores to what extent computational techniques and data visualization methods can provide a more automatic and objective way to deal with the meaning of lexical items in corpus-based studies of argument structure.

My approach is based on vector-space models as a means of representing word meanings. Drawing on the observation that words occurring in similar contexts tend to have similar meanings, vector-space models represent the meaning of a word as a vector in a multidimensional space recording the occurrence of other words in its surrounding context in a vast text corpus. Mathematical measures of similarity between vectors can be used to derive the degree of semantic relatedness between two words. Vector-space models are widely used in NLP applications (in particular for information retrieval), and while linguists are traditionally skeptical about their accuracy and linguistic relevance, recent research has revealed that vector-space-based measures of semantic relatedness correlate positively with a range of behavioral data, and that distributional information may even well be a necessary component of how humans acquire semantic representations (cf. Lund et al. 1995, Andrews et al. 2009, *inter alia*).

I will show how vector-based semantics can be used to visualize the semantic distribution of argument structure constructions. The method consists in computing all pairwise distances between the semantic vectors of the verbs in the distribution of a construction, which are then fed to a multidimensional scaling algorithm that positions the verbs in a 2-dimensional space. This space can then be plotted to visualize the semantic domain of the construction and observe how verbs in that domain are related to each other. To illustrate the potential of this method, I will present a case study examining the diachronic development of the intensifying construction "V *the hell out of* NP" (e.g., *They scared/annoyed/beat/intimidated the hell out of me*) and its variants, drawing on data from the Corpus of Historical American English (COHA). The first attestations of this construction date back from the 1930s, but its distribution was originally much more limited than in present-day American English. In my case study, I plotted the semantic distribution of the construction in four successive 20-year periods in order to visualize its semantic evolution. The comparison of the plots reveals that the 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 essentially driven by type frequency (Barðdal 2008, Bybee 2010, Zeschel 2012). The semantic plots also show that frequent but isolated verbs are less likely to attract new members than clusters of infrequent verbs, which confirms the idea that token frequency is not a particularly strong source of productivity, and also that entrenched exemplars can still serve as the basis for analogical extensions of a limited scope (Barðdal 2008, Zeschel 2010).

References

Andrews, M., Vigliocco, G. & D. Vinson (2009). Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review* 116(3), 463-498.

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.

Lund, K., Burgess, C. & R. Atchley (1995). Semantic and associative priming in a high-dimensional semantic space. *Cognitive Science Proceedings (LEA)*, 660-665.

Zeschel, A. (2010). Exemplars and analogy: semantic extension in constructional networks. In Glynn, D. & K. Fischer (eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches*. Berlin: Mouton de Gruyter.

Zeschel, A. (2012). *Incipient productivity. A construction-based approach to linguistic creativity*. Berlin/New York: de Gruyter.