Distributional characterization of constructional meaning

Despite their descriptive adequacy, constructional approaches to language are still challenging for corpus linguistics. A common problem is that the tokens of a particular construction may be difficult to identify automatically in a corpus. While some constructions can be identified on the basis of lexical material (e.g. *way* in the *way*-construction) or part-of-speech tags, these strategies do not work for Argument Structure Constructions (hereafter ASCs), defined by Goldberg (1995) as independent form-meaning pairs that associate a set of argument roles and their syntactic realization with a basic clausal meaning. For instance, the caused-motion construction superficially consists of the same set of constituents that is found in a transitive clause with a locative adjunct:

(1) a. He swept the dirt under the carpet.
    b. He saw the dirt under the carpet.

Moreover, ASCs convey their own independent meaning to the clause, which eludes corpus-based analyses because only formal cues are available in most corpora; there is no direct way to access the meaning of words, let alone that of larger units. The main goal of this paper is to design and test ways to derive constructional meaning from corpus data.

Drawing on recent corpus findings about the interaction of verbs and constructions (Stefanowitsch and Gries 2005, Gries and Stefanowitsch 2006), we suggest that the meaning of ASCs can be approximated by the distribution of the verbs that occur with it, a prediction actually supported by Goldberg (1995). This predicts that differences in constructional meaning should be reflected by differences in verbal distribution; in other words, the more similar the distribution of two distinct constructions, the closer they are semantically. If successful, this technique could be ultimately used to objectively determine which ASCs are available for a given syntactic pattern, in other words, to resolve constructional homonymy.

In order to implement this hypothesis, we used several indices quantifying the similarity between distributions: the cosine distance metric of vector space and our own indices based on type and token frequencies. We derived a construction sample from the ICE-GB corpus by manually identifying several constructions defined in Goldberg (1995) and Goldberg & Jackendoff (2004), namely the ditransitive, the conative, and the family of resultative constructions. We used the indices to quantify the similarity between the verbal distributions of these constructions and checked whether they adequately reflect semantic differences between them.

Our results show that the basic strategy is successful at distinguishing between the meaning of different constructions, since constructions with very different meanings have very divergent distributions, and conversely, the distributions of constructions that share some meaning, like the ditransitive and the caused motion constructions that both feature a transfer component, are identified as more closely related (but not identical). Since those two constructions assign different semantic roles, the latter statement suggest that the arguments of each construction play a significant role in further contrasting semantically close patterns, which indicates a way in which our method could be refined.

References

Goldberg, A. E. and Jackendoff, R. (2004). The English resultative as a family of constructions. *Language*, 80(3):532–568.

Gries, S. T. and Stefanowitsch, A. (2006). Cluster analysis and the identification of collexeme classes. In Newman, J. and Rice, S., editors, *Empirical and Experimental Methods in Cognitive/Functional Research*. CSLI Publications, Stanford, California.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Stefanowitsch, A. and S. T. Gries (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1 (1), 1-43.