# Checking the adequacy of second-order vector space models of meaning

Mariana Montes, Dirk Geeraerts, Dirk Speelman, Kris Heylen
KU Leuven
mariana.montes@kuleuven.be, dirk.geeraerts@kuleuven.be, dirk.speelman@kuleuven.be,
kris.heylen@kuleuven.be

Keywords: distributional semantics, English, sense disambiguation, vector space models

**Research question** – In the Nephological Semantics research project of which the present study is a part, automatic semantic analysis takes the form of second order vector space models (Heylen, Speelman, & Geeraerts, 2012; Heylen, Wielfaert, Speelman, & Geeraerts, 2015). While type-level models collapse the variation of each lexical item into a unique vector, token-level (or second-order) models build on the type-level vectors to represent individual tokens, and thus model semantic variation within each type.

The question then arises: how adequate are such models for discriminating senses, compared to a manual, concordance-based lexical analysis?

**Relevance** – The question addressed in this study is important for the methodology of Cognitive Linguistics, both from a fundamental and a practical point of view. From a fundamental perspective, the investigation is a contribution to the long-standing debate about converging evidence in Cognitive Linguistics: what are the limits and complementarity of various data types (in the case at hand, alternative corpus-based approaches) in semantics?

Practically speaking, corpora constitute an immensely valuable source of language material for usage-based linguistic analyses. In the case of semantics, large-scale annotation is still an expensive, time- and energy-consuming enterprise that relies mainly on manual work. Thus, checking the reliability of automatic sense disambiguation techniques is highly relevant for any attempt to scale up corpus-based semantic analyses of large corpora.

**Method and materials** – The case study presented here involves the item *church* in the Brown corpus. As the vector space models may vary along different parameters (association measures used for weighting, span of context words, etc.), the study considers a wide range of alternative models. These are based on (weighted) frequency matrices based on the co-occurrence frequency of the token's context words with other types (i.e. merging the type-level vectors of the context words to represent a token). The resulting matrices are then rendered, through multidimensional scaling, as 2D-scatterplots. In this visualization, each point represents an occurrence, while its position relative to the other points, the similarity to the other occurrences.

These models are assessed by visual inspection, and by means of the diagnostics described in Speelman & Heylen (2016), against two disambiguated versions of the data: on the one hand, one based on the Semcor semantically tagged section of the Brown corpus, on the other hand, a manual concordance-based version that takes into account specifically cognitive linguistic features of semantic structure (such as multidimensionality and overlapping of senses).

**Results and discussion** – The case study shows that the adequacy of the token level vector space models depends largely on the granularity of the semantic analysis. Broad distinctions like that between an institutional and a material reading of *church* can be approximated, but finer metonymical, facet-like distinctions (like that between the church building as architectural object or as place of worship) rely on specific lexical cues that are not sufficiently picked up and weighted by the present models.

Further steps in the research may therefore involve the questions, first, through which techniques such cues could be detected in the corpus, and second, whether the models could be improved by enriching them with externally available information, like lexical associations or other experimental data.

## References

Heylen, K., Speelman, D., & Geeraerts, D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. In *Proceedings of the eacl 2012 Joint Workshop of LINGVIS & UNCLH* (pp. 16–24). Avignon.

Heylen, K., Wielfaert, T., Speelman, D., & Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, *157*, 153–172. https://doi.org/10.1016/j.lingua.2014.12.001

Speelman, D., & Heylen, K. (2016). From dialectometry to semantics. In M. Wieling, G. Bouma, & G. van Noord (Eds.), *From Semantics to Dialectometry (Festschrift John Nerbonne)* (pp. 1–12). Groningen: University of Groningen.