Token-based distributional semantics for grammatical alternation research

Syntactic research has been increasingly interested in the **study of alternating constructions**, i.e. forms that are largely considered to be mutually interchangeable (e.g.: the dative alternation (Szmrecsanyi et al. 2016)). Corpus-based analyses of such grammatical alternations typically involve meticulous annotation of high-order properties of the context in which the grammatical form appears (e.g.: animacy of the theme etc.). Yet, taking stock of grammatical alternation research Gries (2019: 78) cannot help but notice that "one aspect of the context seems to be crucially underutilized when it comes to modeling speakers' choices: the **lexical context**."

The integration of lexical context in alternation research can be operationalized in various ways. In this paper we explore the use of **token-based distributional-semantic modelling** (Schütze 1998; Hilpert and Saavedra 2020). This type of analysis constitutes a corpus-driven method for modelling the meaning of individual corpus occurrences of a certain variant (e.g.: the prepositional variant in the dative alternation). The so-called semantic vector of such a corpus token is derived indirectly by modelling precisely the meaning of the lexical context surrounding that grammatical token. The modelled tokens can then be represented as token clouds in a multidimensional vector space, with clusters of tokens revealing the **polysemy of the grammatical forms**. Token-based vector semantics has proven a promising method for the study of lexical-semantic phenomena (Author1, 2019; Heylen et al., 2015). The novelty of our contribution is the extrapolation of this technique from the purely lexical domain to that of morphosyntax.

Concretely, we choose as case study the **transitive-prepositional alternation in Dutch**, exemplified by a construction such as *naar een boek zoeken* vs. *een boek zoeken* 'to search (for) a book'. This alternation was investigated in depth in Author2 (2019). The large manually annotated dataset underlying his study comprises 117697 tokens of different verbs and prepositions participating in this alternation and offers an important point of comparison to evaluate our token-based distributional semantic take on the issue.

With this case study, we will illustrate the several **benefits** that follow from our approach. First, in contrast to the traditional top-down identification of high-order predictors, a token-based distributional analysis can now be used to identify those features in a bottom-up way. Second, by superimposing the token clouds of each of the grammatical variants, one can distinguish regions of contextual overlap (i.e. where the variants are interchangeable) from token regions in which the forms cannot alternate. The semi-automatic identification of overlapping token clouds contributes to scaling up grammatical alternation research, by providing methods for dealing with corpora whose size exceeds manual analysis. Third, as the window span of the lexical context is a tunable parameter in our token-based models we can compare the local lexical context, only encompassing the relevant syntactic slots of the variants, to that of the broader lexical context, which might include other, topically related lexical items. As the focus of most grammatical alternation research goes to the former type of context words, it has to be verified what other semantic information such broader bag-of-words context can contribute.

Bibliography

Gries, Stefan Th. 2019. The Discriminatory Power of Lexical Context for Alternations: An Information-theoretic Exploration. *Journal of Research Design and Statistics in Linguistics and Communication Science* 5(1–2). 78–106.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman and Dirk Geeraerts. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157. 153–172. (26 October, 2015).

Hilpert, Martin and David Correia Saavedra. 2020. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 16(2). De Gruyter Mouton. 393–424.

Schütze, Hinrich. 1998. Automatic Word Sense Discrimination. *Computational Linguistics* 24(1). Cambridge, MA, USA: MIT Press. 97–123.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller and Melanie Röthlisberger. 2016. Around the world in three alternations. *English World-Wide. A Journal of Varieties of English* 37(2). John Benjamins. 109–137.