Using token-based semantic vector spaces for corpus-linguistic analyses

This paper showcases a recent corpus-linguistic method that is typically associated with “big data” approaches, namely token-based semantic vector space modeling (Schütze 1998, Heylen et al. 2012). I will discuss how this approach can provide the input for in-depth qualitative studies of grammatical phenomena, as for example a contrast between two near-synonymous constructions. I will illustrate this approach by contrasting pairs of English constructions.

Semantic vector space models are routinely used in computational linguistics, where they are applied to problems such as word sense disambiguation or information retrieval (Turney and Pantel 2010). The technique has been adopted in a number of corpus-based studies (e.g. Sagi et al. 2011, Jenset 2013, Perek 2016, amongst others), but it remains a relatively underused technique. Its core idea is captured by the slogan You shall know a word by the company it keeps (Firth 1957: 11), which reflects the hypothesis that the meaning of a word is related to its distribution in actual language use. Semantic vector spaces analyze the meaning of a given word in terms of other words that occur frequently in close proximity to that word. For instance, the
noun toast frequently occurs in close proximity of nouns such as tea, cheese, and coffee. A statistically processed frequency list of all collocates of toast in a given corpus is called a semantic vector. Semantic analysis enters the picture when semantic vectors of several words are compared. Most current applications of semantic vector space models analyze word types, thus averaging collocate frequencies over many occurrences of the same word. Our approach builds on the general logic of such type-based semantic vector spaces, but we adopt a more specific proposal from Heylen et al. (2012), which operates at the level of word tokens, thus capturing meaning differences between individual occurrences of the same word. The primary unit of data in such an approach is the concordance line, that is, a key word with a context window of several words to the left and several words to the right.

As an illustration of the approach, I will first discuss contrasts between different English modal auxiliaries. By contrasting concordance lines of synonymous expressions such as may and might (or must vs. have to) in a token-based semantic vector space that is based on the British National Corpus, we can visualize both semantic differences and areas of semantic overlap between the two. Furthermore, we can identify concordance lines in which may and might are synonymous, and others in which the two expressions convey very different meanings. A second illustration will focus on pairs of constructions that involve grammaticalized forms and their lexical counterparts. It will be shown that these can be distinguished on the basis of their collocational properties, and I will argue that this kind of comparison can be applied in theoretical arguments concerning the unidirectionality of grammaticalization (Jäger and Rosenbach 2008).

References


Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. Linguistics 54/1, 149-188.


Everyone is cordially invited.