There and back again. Re-introducing distributional semantics into linguistics.

Yves Peirsman

Stanford University & University of Leuven

The distributional hypothesis in lexical semantics claims that two words with similar meanings occur in similar contexts (Firth 1957). For example, both bike and bicycle often appear in the construction ride a, and tend to occur together with words like cyclist or wheel. This means it is often possible to identify synonyms automatically, simply on the basis of a large corpus. Although this distributional hypothesis finds its origin in theoretical linguistics, today distributional methods are mostly restricted to more technical disciplines like Natural Language Processing (Pado and Lapata 2007, Turney and Pantel 2010, Baroni and Lenci 2010).

At the same time, research in corpus linguistics is becoming more and more data-driven. Corpus-driven research in lexical semantics, however, appears to focus more on difference than similarity in meaning (see e.g., Gries and Stefanowitsch 2004, Janda and Solovyev 2009, Divjak 2010). Yet, this difference-oriented research typically requires a set of (near-)synonymous words or constructions as its input. Instead of choosing these manually, distributional methods can often identify this data automatically. They can therefore help scale present research in corpus-driven lexical semantics (Geeraerts 2010). In this presentation, we will show how distributional methods can be applied further to address typically linguistic research questions. We will focus on lexical variation as a case study.

Distributional models are usually constructed on the basis of a corpus representative of one language variety (like British English) or genre (like newspaper articles). This means they are of relatively limited use to the study of language variation. However, if we build a distributional model on the basis of two corpora from different language varieties, we can automatically identify synonyms across these varieties. In our case study, we focus on the standard varieties of Dutch used in Belgium and the Netherlands, and the standard varieties of German used in Austria and Germany. We show how it is possible to automatically identify Netherlandic Dutch synonyms to Belgian words, like the Netherlandic Dutch synonym toetje for Belgian dessert "dessert", and German German synonyms to Austrian German words. Next we will illustrate the opposite perspective. Instead of identifying words with the same meaning in two corpora, we will present a method to find words that have a different meaning in two language varieties. For example, tas in Netherlandic Dutch can only mean "bag", while in Belgium it can also mean "cup (of coffee)". This type of words can be identified through their difference in contextual distribution between the two corpora.

The introduction of distributional methods in more theoretical corpus linguistics has a number of advantages. For example, automatically identified sets of synonyms can serve as the input for later studies in the vein of Janda and Solovyev (2009) and Divjak (2010), and help scale these approaches to a larger part of the lexicon. Similarly, together with an automatic method for identifying differences in meaning, they can prove useful in areas like lexicography or applications like terminology extraction.


