The use of vector models in aggregate-level studies of semasiological variation

DIRK SPEELMAN
KRIS HEYLEN
DIRK GEERAERTS
KU Leuven

Aggregate-level studies of linguistic variation typically adopt an onomasiological perspective on linguistic variation. Recently, however, a numer of corpus-based techniques have been developed in the distributional semantic framework to detect semantic changes in large corpora (Sagi et al. 2011, Cook & Hirst 2011, Gulordava & Baroni 2011). In this paper, we will adopt one such technique to the corpus-based, aggregate-level investigation of semasiological regional and register variation in Dutch. More specifically, we will discuss **token-based vector space models**, in which (a random subset of) the tokens of a target word are represented as a 'token cloud in vector space'. In such token clouds, the co-ordinates of the tokens are a proxy for the meaning/usage of the target word in that token, and the distances between the tokens are a proxy for how different the meaning/usage of the target word is in these tokens.

In order to compare the use of a target word in two varieties, we superimpose its token clouds from both varieties. In the paper we will discuss measures to quantify to which extent superimposed clouds exhibit non-overlapping regions. Such regions signal possible *differences in the (number of) senses in both varieties*. In figure 1 we see a visualization of superimposed clouds of Belgian (triangles) and Netherlandic (circles) tokens of *monitor*. In this example, manual inspection has revealed that the upper part of the figure (where there are both Belgian and Netherlandic tokens) contains tokens with the sense SCREEN OF COMPUTER OR OTHER ELECTRONIC DEVICE, which exists in both Belgium and the Netherlands, whereas the lower part of the figure (where there are only Belgian tokens) contains tokens with a sense of monitor that doesn't exist in the Netherlands (YOUTH LEADER).

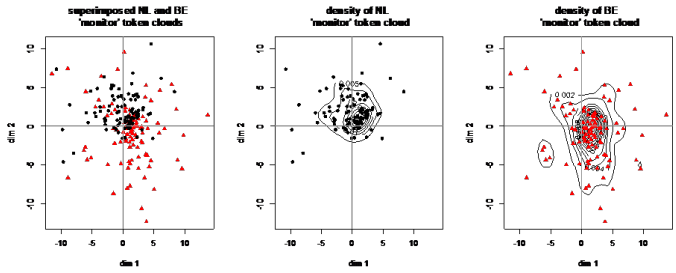

Figure 1: Non-metric MDS representation of Belgian and Netherlandic 'monitor' token clouds