Scaling-up lexical variationist research in pluricentric languages with type- and token-level vector semantics

Stefano De Pascale[1], Weiwei Zhang[1], Kris Heylen[1]

[1] *KU Leuven*

stefano.depascale@kuleuven.be, weiwei.zhang@kuleuven.be, kris.heylen@kuleuven.be

In the field of lexical lectometry, i.e. the aggregate-level analysis of lexical variation, solid empirical evidence has been provided about the status of varieties in pluricentric languages such as Dutch (Geeraerts, Grondelaers, & Speelman 1999) and English (Ruette, Ehret, & Szmrecsanyi 2016). The focus of these studies is to measure the distances between national varieties by looking at how they use near-synonyms differently for expressing a given concept. However, at present we lack the same quantitative coverage on how not words, but senses might be sociolinguistically distributed as well. Importantly, in order to establish the interchangeability of near-synonyms in a lexical sociolinguistic variable, one has to control precisely for these variety-specific senses. Moreover, the aggregate perspective inherent in a corpus-based lectometric inquiry urges us to explore computational-semantic techniques in order to deal with corpora whose size hinders manual sense disambiguation.

Type-based distributional semantics as embodied in vector space models (VSMs) has proven to be a successful method for the retrieval of near-synonyms in large corpora. In addition, Peirsman, Geeraerts, & Speelman (2010) showed that constructing type-based vectors on different regiolectal corpora turned out to be very useful for estimating the degree of regional polysemy between Belgian-Dutch and Netherlandic-Dutch. However, such a type-based solution is far from ideal: since all senses of a word are lumped together into one vector representation, we have no direct access to the contextual subtleties that cause the regiolectal polysemy. In addition, operating at the word level, these type-based models are not helpful for removing the tokens that express the variety-specific senses, an important requirement for our lectometric calculations.

Our paper reports on methodological research aiming at better semantic control in the lectometric use of VSMs. We therefore introduce token-based VSMs to disambiguate senses of lexical variants (Heylen, Speelman, & Geeraerts, 2012). This type of VSMs identifies different usage tokens of a word in a corpus, with token clusters revealing the senses of the word. By superimposing the token clouds of the lexical variants in a variable, one can distinguish which meanings are shared by near-synonyms and determine the 'semantic envelope of variation'.

By making use of two regiolectally-balanced corpora of Dutch and Chinese, we aim to show, with a sample of synsets taken from the WordNets of the two languages, the importance of semantic control on the composition of lexical variables. In general, the comparison and fine-tuning of these procedures are meant to contribute to the scaling up of lexical lectometric analyses.

References: Geeraerts, D., Grondelaers, S., & Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat*. Amsterdam: P.J. Meertens-Instituut. Heylen, K., Speelman, D., & Geeraerts, D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. In M. Butt, S. Carpendale, G. Penn, J. Prokic, & M. Cysouw (Eds.), *Proceedings of the EACL-2012 joint workshop of LINGVIS & UNCLH* (pp. 16–24). Avignon, France: ACL. Peirsman, Y., Geeraerts, D., & Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, *16*(4), 469–491. Ruette, T., Ehret, K., & Szmrecsanyi, B. (2016). A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models. *International Journal of Corpus Linguistics*, *21*(1), 48–79.