Token-based Vector Space Models as semantic control in sociolinguistic research on lexical variation

Stefano De Pascale, Stefania Marzo & Dirk Speelman
(KU Leuven)

Meaning, for decades virtually absent from mainstream sociolinguistic theory, has taken center stage in "third wave" sociolinguistics, which has brought into view how speakers' expression of social identity is conveyed by sociolinguistic variation in local linguistic practices (Eckert, 2012). This new focus has been accompanied by a methodological shift, with in-depth ethnographic observational studies on the *meaning of sociolinguistic variation* replacing large-scale quantitative surveys, considered inadequate for the new research program.

However, the goal of this paper is to show that "first wave"-style, i.e. variationist research is reconcilable with an interest in uncovering the *sociolinguistic variation of meaning*. From a theoretical point of view, we want to show that a firmer grip on the types of lexico-semantic variation must start from a theory of semantics that embraces an experiental and flexible view on linguistic categorization (Geeraerts, Grondelaers, & Bakema, 1994). Methodologically, this usage-based view on meaning gets translated in the use of large corpora and quantitative techniques. This paper reports exactly on the methodological advances that have made possible the "first wave", variationist study of meaning.

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. These words have then been used in lexical sociolinguistic variables (e.g.: *tube* and *television* for the concept TELEVISION; Ruette, Geeraerts, Peirsman, & Speelman [2014] and Ruette, Ehret, & Szmrecsanyi [2016]). However, a limitation of type-based VSMs is that all senses of a word are lumped together in one vector representation, making it harder to control for polysemy and subtle contextual distinctions.

To address this shortcoming, first, we introduce token-based VSMs in lexical variation research, in order to disambiguate different senses of lexical variants (Heylen, Speelman, & Geeraerts, 2012). Such VSMs identify different meaning/usage tokens of a word in a corpus that are afterwards represented as token clouds in a multidimensional space, with token clusters of semantically similar tokens revealing the senses of the word. Second, by superimposing the token clouds of the lexical items, one can distinguish the overlapping areas of tokens of the different near-synonyms and so determine the 'semantic envelope of variation'. The remainder of the study will show how the calculation of the overlapping area can be carried out, using a set of cluster overlap indices evaluated in Speelman & Heylen (2015).

The finetuning of token-based VSMs targeted by the present study not only contributes to the scaling up of lexical variationist research, and confirms the potential of "first wave" studies in lexico-semantic variation. At the same time, token-based VSMs comply with the need of detailed analysis argued by "third wave" sociolinguistics, by allowing the possibility of zooming in on the behavior of individual tokens in order to determine more subtle contextual distinctions.

References

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. Annual Review of Anthropology, 41, 87–100.

Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). The Structure of Lexical Variation. Berlin: De Gruyter Mouton.

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: Visualization of Language Patters and Uncovering Language History from Multilingual Resources (pp. 16–24). Avignon, France: Association for Computational Linguistics.

Ruette, T., Peirsman, Y., Speelman, D., Geeraerts, D. (2014). Semantic weighting mechanisms in scalable lexical sociolectometry. In: Szmrecsanyi B., Waelchli B. (Eds.) *Aggregating Dialectology, Typology, and Register. Analysis Linguistic Variation in Text and Speech.* Berlin: de Gruyter, 205-230.

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.

Speelman D., Heylen K. (2017). From dialectometry to semantics. In: Wieling M., Kroon M., Van Noord G., Bouma G. (Eds). *From Semantics to Dialectometry. Festschrift in honor of John Nerbonne*, UK: College Publications, 325-334