Does the data deluge redefine the dictionary?

Dirk Geeraerts

University of Leuven, Belgium

Arguably, one of the most inspring but at the same time also most challenging contemporary perspectives for lexicography is the availability of massive amounts of digital corpus data: how can the current data tsunami be channeled for lexicographical purposes, and does the availability of big data change our idea of what a lexicographical description should look like? For the traditional descriptive purposes of the standard-language dictionary, the mass of currently available text data constitutes an opportunity: more raw information holds a promise of better descriptions – provided there are efficient computational tools for tracing meaning changes in the vastness of the data, for selecting the most relevant items for description, or for identifying the most illustrative quotations need further development. But the data explosion should not be envisaged only from the point of view of traditional descriptive practices. Treating all the new data with a classical amount of attention has become a practical impossibility, but at the same time, that difficulty might also seem to carry its own solution: the magnitude of the data makes it easier – at least in principle – to analyze tendencies in the vocabulary that transcend the level of the individual word.

Drawing on the lexical variation research that we have been doing in the Quantitative Lexicology and Variationist Linguistics group, I will explore that latter option. Does it make sense for (academic) dictionaries of the contemporary language to go beyond item-focused descriptions and include an analysis of underlying trends in the lexicon? What kind of questions could be addressed, and what type of sociolectometrical techniques would be necessary to answer them? Which kind of data would be required, and what kind of computational support could be enlisted?