Cognitive sociolinguistics meets Culturomics: A longitudinal distributional semantic analysis of lexical variation in immigration discourse

KRIS HEYLEN
DIRK SPEELMAN
KU Leuven

Within Cognitive Linguistics, Cognitive Sociolinguistics (CS: Kristiansen & Dirven 2008, Geeraerts, Kristiansen, Peirsman 2010) emphasises the socio-cultural grounding of language and the need for a usage-based, quantitative methodology to study linguistic phenomena. Studies typically investigate syntactic and lexical variation through a multivariate statistical analysis of a sizeable, yet manually coded dataset which is culled from socio-culturally stratified corpora. Culturomics, on the other hand, was introduced by Michel et al. (2011) as a way to investigate cultural trends quantitatively in textual 'Big Data' without manual coding. Both approaches are thus interested in the relation between culture and language but they reverse explanans and explanandum and make different methodological choices: In general, CS prefers 'deep' semantic variables that are coded manually in a corpus sample, whereas Culturomics relies on surface phenomena that are easily retrievable in the (large) corpus as a whole. In this paper, we explore how a culturomic approach (longitudinal analysis of cultural trends in Big Data) can be integrated into the Cognitive Sociolinguistic research programme by going beyond the analysis of mere surface patterns and using distributional semantic modelling (Turney and Pantel 2010 for an overview) to find higher level semantic patterns in large, stratified corpora. As a case study, we look at immigration discourse in six Dutch-language Belgian newspapers from 1999 to 2013. We collected all occurrences of lexemes referring to the concept IMMIGRANT in Dutch (*allochtoon, vreemdeling, (im)migrant, buitenlander, nieuwe Belg: n = 180K*) and use distributional semantic modelling to detect shifting patterns in the contextual semantics associated with each of the words. We can see for example that *allochtoon specialises* over time for contexts related to equal job opportunities, whereas migrant is used to discuss voting rights. These distributionally identified semantic clusters are then included together with sociolinguistic variables in a time series analysis of the variation in IMMIGRANT designations.