Methodological issues in using word embeddings in a sociolinguistic perspective: the case of contact-induced semantic variation across Canadian Twitter corpora

Filip Miletic[1], Anne Przewozny-Desriaux[1], Ludovic Tanguy[1]

[1] *CLLE: CNRS & University of Toulouse*

{filip.miletic, anne.przewozny, ludovic.tanguy}@univ-tlse2.fr

We present an interdisciplinary approach to lexical semantic variation in Quebec English, a regional variety characterized by contact with French. We base data collection and analysis on variationist sociolinguistic research (Labov, 1972) and follow recent computational studies in applying word embeddings to synchronic semantic variation (Fišer & Ljubešić, 2019). We use a corpus of 42M tweets (728M tokens) from Montreal, Toronto and Vancouver, collected between January and April 2019, with phenomena specific to Montreal expected to reflect the influence of French. For each subcorpus, an embeddings model was trained using word2vec (Mikolov et al., 2013). Similarly to diachronic studies (Hamilton et al., 2016), we aligned the models and computed cosine-distances between each word's vectors to detect divergences in Montreal.

Our method correctly identified contact-induced meanings including *exposition* 'exhibition' and *terrace* 'restaurant patio', which reflect previous sociolinguistic studies (Boberg, 2012), as well as similar new cases such as *definitively* 'definitely': the unconventional meanings are all likely related to French cognates (*exposition*, *terrasse*, and *définitivement*). Crucially, an analysis of users' language choices on Twitter shows that, unlike established regional variants, the contact-induced meanings represent a variation in usage largely limited to bilingual speakers.

However, other results are of limited interest. In addition to the sporadic impact of prolific users, certain meanings relate to cultural factors, such as Montreal's thriving IT sector (*unsupervised* referring to machine learning) or Vancouver's proximity to the Pacific Ocean (*chum* denoting a species of salmon). A local referent is at play in *plateau*, which in Montreal refers to the borough of Plateau-Mont-Royal. Our method also identifies French items which are homographous with unrelated English words and occur in code-switched tweets (*pour* 'for').

Our ongoing work focuses on addressing these issues through word-level language identification and control of topical usage variation. More generally, the successfully identified examples confirm the need for our approach at the intersection of natural language processing and sociolinguistics: word embeddings trained on geotagged data are instrumental in detecting semantic shifts, while fine-grained variationist sociolinguistic analysis is necessary to uncover precise usage patterns. While this analysis currently relies on Twitter metadata, our methodology also includes sociolinguistic fieldwork based on a subset of the indexed users. Our aim is to investigate the precise status of the computationally identified linguistic variants and their relationship with real-life sociolinguistic behaviors. This will in turn help inform data collection, analysis and evaluation in future computational studies of sociolinguistic phenomena.

References: Boberg, C. (2012). English as a Minority Language in Quebec. World Englishes 31(4), 493–502. Fišer, D. & N. Ljubešić (2019). Distributional modelling for semantic shift detection. Int. J. of Lexicography 32(2), 163–183. Hamilton, W. L., J. Leskovec & D. Jurafsky (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proc. of ACL. Labov, W. (1972). Sociolinguistic Patterns. Philadelphia: UPP. Mikolov, T., K. Chen, G. Corrado & J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. Proc. of Workshop at ICLR.