Methodological Issues in Using Word Embeddings in a Sociolinguistic Perspective: The Case of Contact-Induced Semantic Variation Across Canadian Twitter Corpora

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy

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

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy

CLLE, CNRS & University of Toulouse (France)

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

Introduction

Semantic variation in Quebec English

- Quebec English (QcE) is a regional variety of Canadian English spoken by a **minority** of Quebecers.
- A possible consequence of **contact** with Quebec French: English words used with meanings typical of French cognates.
- Existing descriptions (Fee, 1991; 2008; Boberg, 2012; Rouaud, 2019) do not explore the **extent** or the precise **status** of this phenomenon.

An interdisciplinary approach

- Aim: systematically identify **semantic variants** specific to QcE and investigate speaker-level **factors** driving this variation.
- Data collection and analysis draw on **variationist sociolinguistics** (Labov, 1972; Tagliamonte, 2012).
- **Word embeddings** are used to computationally detect synchronic semantic variation (e.g. Del Tredici & Fernández, 2017).

Methodology

Data

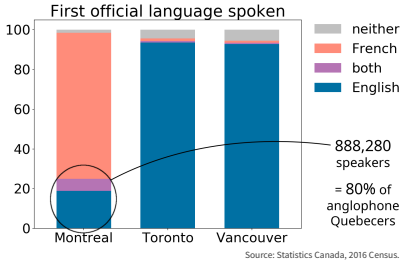
- Corpus of **tweets** published from 2015 onward
- **Search API** to identify users ⇒ timeline **crawl**
- User-level **filtering**: location, language, near-duplicate exclusion

Subcorpus	Users	Tweets	Tokens
Montreal	47k	13m	236m
Toronto	41k	15m	277m
Vancouver	38k	14m	270m
Total	126k	42m	783m

Analysis

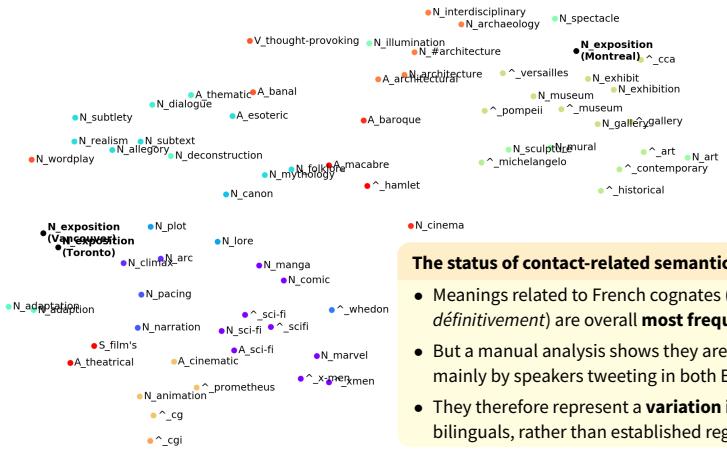
- A word embedding model was trained for each subcorpus using **word2vec** (Mikolov et al., 2013).
- Models were aligned using **Orthogonal Procrustes** as in diachronic studies (Hamilton et al., 2016).
- Prominent divergences in Montreal were detected using **cosine-distances** between a word's vectors.

$$diff(w) = \frac{cos\text{-}dist(w_M, w_T) + cos\text{-}dist(w_M, w_V)}{cos\text{-}dist(w_M, w_T) + cos\text{-}dist(w_M, w_V) + cos\text{-}dist(w_T, w_V)}$$

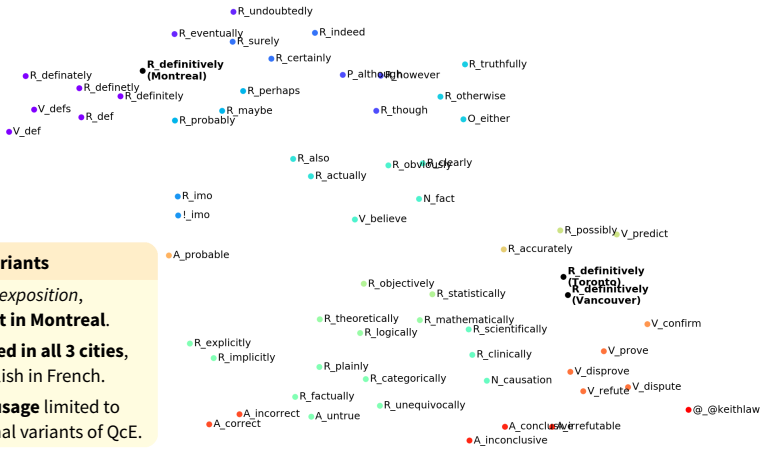


Key examples

A previously described case: *exposition*



A newly identified case: *definitively*



The status of contact-related semantic variants

- Meanings related to French cognates (cf. *exposition*, *définitivement*) are overall **most frequent in Montreal**.
- But a manual analysis shows they are **used in all 3 cities**, mainly by speakers tweeting in both English in French.
- They therefore represent a **variation in usage** limited to bilinguals, rather than established regional variants of QcE.

*I really want to go to an art museum or an art **exposition** :(*

*Canada's centennial year saw Montreal host the 1967 International and Universal **Exposition***

*On parle de notre **exposition** Brown's Hill! // An article about our **exhibition** Brown's Hill*

*Three straight scenes of clunky dialogue filling in for **exposition**. Yup, it's a Schwarzenegger film!*

*A brilliant **exposition** of dietary fiber & the wonders it can perform for human health.*

*Pouring coffee beans in the water tank... I **definitively** need coffee!!*

*That's most **definitively** a 10*

*Im getting tattooed right now and it's **definitively** the one that hurt the most*

*Worse, the research community has performed multiple trials & studies that all **definitively** show no connection between vaccination & autism*

*I think we can **definitively** say Carey Price is permanently broken.*

Methodological issues

Results of limited interest

- **Cultural factors:** *unsupervised* refers to machine learning due to Montreal's IT industry; *chum* denotes a species of salmon in Vancouver due to its geography.
- **Local referents:** *plateau* refers to the borough of Plateau-Mont-Royal in Montreal.
- **Prolific users:** *waffle* 'make waffles' (rather than 'be undecided' or 'speak at length') is prominent in Montreal due to a single Twitter account.
- **French items in codeswitched tweets:** *pour* 'for' affects the Montreal representation of the English verb *pour*.

General issues

- The variation we study is **subtler** than long-term semantic change: conventional and contact-related meanings coexist ⇒ **polysemy**.
- Regional regularities do not suffice to explain this variation: we need to identify and describe **speakers** with similar behaviors.
- Uncertainty over the relationship between computational results and **real-life sociolinguistic behaviors**.

Ongoing work

Preprocessing

- Word-level language ID
- Topic modelling

Word embedding models

- Contextual word embeddings ⇒ polysemy, user clusters

Sociolinguistic fieldwork

- **Cohort study** based on a group of native speakers reflecting linguistic profiles from the Twitter corpus.
- Aim: examine the **status** and **representations** of the variants detected using word embeddings.
- The results will **inform future studies** in computational sociolinguistics.

Conclusions

- This study has brought to light important descriptive observations, particularly the role of **bilingualism** and the importance of **polysemy** in contact-induced semantic variation.
- **Methodological issues** obfuscate relevant results in our models, but they are being addressed through ongoing work.
- The need for our interdisciplinary approach is already clear: **word embedding models** are needed to detect semantic variation, and fine-grained **sociolinguistic analysis** to clarify its nature.

References

Boberg, Charles. (2012). English as a minority language in Quebec. *World Englishes* 31 (4): 493–502.
Del Tredici, Marco, and Raquel Fernández. (2017). Semantic variation in online communities of practice. In *Proceedings of ICCS*.
Fee, Margery. (1991). Frenglish in Quebec English newspapers. *Fifteenth Annual Meeting of the APLA*, 12–23.
Fee, Margery. (2008). French borrowing in Quebec English. *Anglistik* 19: 173–188.
Hamilton, William L., Jure Leskovec, and Dan Jurafsky. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of ACL*.
Labov, William. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.
Rouaud, Julie. (2019). *Lexical and Phonological Integration of French Loanwords into Varieties of Canadian English since the Seventeenth Century*. PhD thesis, University of Toulouse.
Tagliamonte, Sali. (2012). *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.