# The Status and Representation of Contact-Induced Semantic Shifts in Quebec English: From Twitter Users to Sociolinguistic Informants

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy

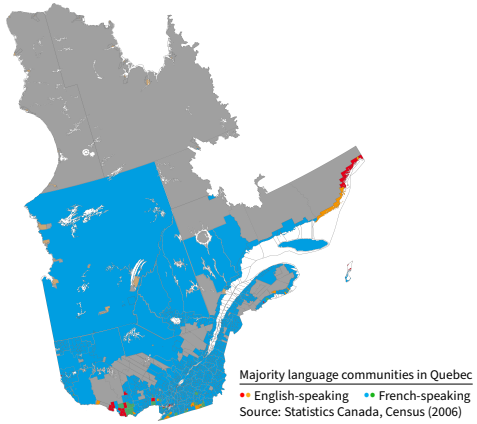
# The status and representation of contact-induced semantic shifts in Quebec English: from Twitter users to sociolinguistic informants

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy
CLLE, CNRS & University of Toulouse (France)
{filip.miletic, anne.przewozny, ludovic.tanguy}@univ-tlse2.fr

## Object of study

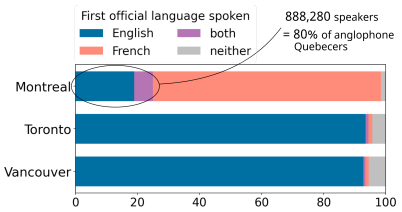
- Contact-induced **semantic shifts in Quebec English**, e.g. *deception* 'disappointment' (cf. Fr. *déception*).

  *Big **deception**... you were not present in the Pride Parade in Montreal today. [...] I keep waiting for a breakthrough but Conservatives keep disappointing.*
- There are **dozens of described examples** (Boberg, 2012; Fee, 1991, 2008; Rouaud, 2019), but most are anecdotal.
- How **widespread** is this phenomenon?
  Which **factors** condition its use?
  What **representations** are associated with it?




## Method

- We operationalize contact-induced semantic shifts as **regional semasiological variation**, and study them comprehensively using an **interdisciplinary approach**.
- We use **computational semantic models** to systematically identify target linguistic patterns in a large Twitter corpus.
- We then implement a **variationist survey** to see how these patterns are reflected by real-life sociolinguistic behaviors.




## Computational models of lexical semantic variation

### Data

- A custom-built corpus of **tweets** (Miletic et al., 2020) published **from 2016** onward.

| Subcorpus | Users | Tweets | Tokens |
|---|---|---|---|
| Montreal | 55 k | 11 m | 193 m |
| Toronto | 51 k | 13 m | 223 m |
| Vancouver | 48 k | 11 m | 213 m |
| Total | 154 k | 35 m | 629 m |

- Basic **sociolinguistic information**: stated location; degree of bilingualism based on languages in tweets.
- The data are used for analyses based on different **word embedding** models (Miletic et al., 2021).

### Type-level analysis

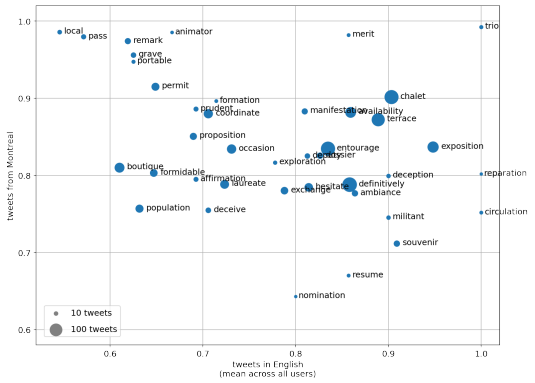
- Identifies the **words** within the whole vocabulary with the **most different meanings in Montreal**.
- Uses **word2vec** (Mikolov et al., 2013) to compare meanings across regions ⇒ ≈**20 new cases**.

### Token-level analysis

- Identifies the contact-induced **senses** of a word.
- Uses **BERT** (Devlin et al., 2019) to analyze 40 target items, producing **clusters** of semantically similar occurrences ⇒ **annotation** for the presence of contact influence.

### Patterns of variation

- For most lexical items, contact-related senses are used by speakers who **tweet in French more often** than those who use the same item with a conventional sense.
- Semantic shifts likely represent **variations in usage** associated with **bilingualism** rather than established regional variants.
- But both regional specificity and association with bilingualism vary across different items, suggesting **differences in diffusion**.




## Variationist sociolinguistic survey

### Protocol

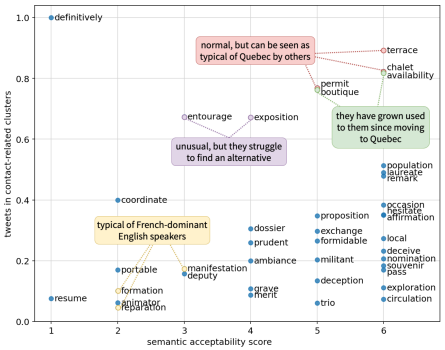
- **PAC-LVTI** protocol (Przewozny et al., 2020): a standard variationist sociolinguistic **interview** and a detailed thematic **questionnaire**.
- A new **protocol extension** for semantic variation (Bailey & Durham, 2020; Dollinger, 2017; Robinson, 2010) using 40 **tweets**, each with an item in a contact sense:
  - **Read** the tweet out loud ⇒ phonological information;
  - Rate **acceptability** from 1 to 6 ⇒ semantic information;
  - Give a **synonym** in this context ⇒ interpretation check;
  - Feel free to **comment** ⇒ representations.

### Participants

- 942 Montreal users from the corpus having used at least one target item with a contact-related sense ⇒ **40 target users** chosen based on idiomaticity.
- Recruitment **through Twitter**, including an explanation of the participants' presence in the initial corpus.
- Data collection is **ongoing**.

### A case study

- 32 y. o. **monolingual English** speaker, ≈10 years in Montreal.
- All lexical items are **phonologically integrated** into English. Acceptability **ratings vary**, with rich qualitative comments.
- The ratings are **not correlated** with computational variation scores, but general patterns (bilingualism, interaction) **mirror Twitter**.




## Conclusion

### Summary

- Our **computational method** for semantic shift analysis entailed the creation of a Twitter corpus and of multiple semantic models.
- This approach identified **previously undescribed** examples and provided insight into their **status and diffusion**.
- Our **sociolinguistic survey** uses a custom interview task building on the computational analyses to further investigate lexical semantic variation.
- Initial results show that this is crucial in establishing **sociolinguistic profiles** and eliciting **representations**.
- The link between Twitter and interview data is complex, which points to **different dimensions** of variation.

### Ongoing work

- **Further sociolinguistic interviews** will provide additional information, contributing to clearer conclusions.
- A direct **comparison of sociolinguistic and Twitter data** will provide descriptive insight as well as allow for systematic evaluation of computational methods.

## References

Bailey, L. R., & Durham, M. (2020). A cheeky investigation: Tracking the semantic change of *cheeky* from monkeys to wines: Can social media spread linguistic change? *English Today*, 1–10.

Boberg, C. (2012). English as a minority language in Quebec. *World Englishes*, 31(4), 493–502.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Dollinger, S. (2017). TAKE UP #9 as a semantic isogloss on the Canada-US border. *World Englishes*, 36(1), 80–103.

Fee, M. (2008). French borrowing in Quebec English. *Anglistik*, 19(2), 173–188.

Fee, M. (1991). Frenglish in Quebec English newspapers. In *Papers of the Fifteenth Annual Meeting of the APLA*, 12–23.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR 2013*.

Miletic, F., Przewozny-Desriaux, A., & Tanguy, L. (2020). Collecting tweets to investigate regional variation in Canadian English. In *Proceedings of LREC*.

Miletic, F., Przewozny-Desriaux, A., & Tanguy, L. (2021). Detecting contact-induced semantic shifts: What can embedding-based methods do in practice? In *Proceedings of EMNLP*.

Przewozny, A., Viollain, C., & Navarro, S. (2020). *The corpus phonology of English: Multifocal analyses of variation*. Edinburgh University Press.

Robinson, J. A. (2010). Awesome insights into semantic variation. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in Cognitive Sociolinguistics* (pp. 85–110). De Gruyter Mouton.

Rouaud, J. (2019). *Lexical and phonological integration of French loanwords into varieties of Canadian English since the seventeenth century*. Doctoral dissertation, Université Toulouse - Jean Jaurès.