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

To cite this version:
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. New Ways of Analyzing Variation (NWAY 49), Oct 2021, Austin, TX, United States. hal-03445857

HAL Id: hal-03445857
https://hal.science/hal-03445857
Submitted on 24 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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
Ctö, CNRS & University of Toulouse (France)
{filip.miletic, anne.przewozny, ludovic.tanguy}@univ-tlse2.fr

Method

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 Conservatifs 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?

Computational models of lexical semantic variation

DATA

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

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Users</th>
<th>Tweets</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Montreal</td>
<td>55 k</td>
<td>193 m</td>
<td>11 m</td>
</tr>
<tr>
<td>Toronto</td>
<td>51 k</td>
<td>223 m</td>
<td>13 m</td>
</tr>
<tr>
<td>Vancouver</td>
<td>48 k</td>
<td>223 m</td>
<td>11 m</td>
</tr>
<tr>
<td>Total</td>
<td>154 k</td>
<td>629 m</td>
<td>35 m</td>
</tr>
</tbody>
</table>

- 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.

Variationist sociolinguistic survey

PROTOCOL

- PAC-LVLT 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.

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.
- Both regional specificity and association with bilingualism vary across different items, suggesting differences in diffusion.

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