Modeling fine-grained sociolinguistic variation: The promises and pitfalls of Twitter corpora and neural word embeddings

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy

Modeling fine-grained sociolinguistic variation: The promises and pitfalls of Twitter corpora and neural word embeddings

Running head: Modeling fine-grained sociolinguistic variation

Filip Miletic,[1,2] Anne Przewozny-Desriaux[1] & Ludovic Tanguy[1]

[1]CLLE, CNRS & University of Toulouse / [2]IMS, University of Stuttgart

Abstract:

This paper examines the use of recent data sources and computational methods to study fine-grained sociolinguistic phenomena. We deploy a custom-built corpus of tweets (Miletic et al. 2020) and neural word embeddings to investigate the use of contact-induced semantic shifts in Quebec English. Drawing on an analysis of 40 lexical items, we show that our approach is beneficial in facilitating manual inspection of vast amounts of data and establishing fine-grained patterns of language variation. While it is affected by a range of noise-related issues, which we describe in detail, a coarse-grained annotation provides an efficient way of circumventing them. We use the results filtered in this way to conduct a quantitative analysis of sociolinguistic constraints on contact-induced semantic shifts, further confirming the utility of our approach.

Keywords: semantic shifts, language contact, Twitter corpora, word embeddings, BERT, Quebec English

1. Introduction

In this paper, we deploy a novel corpus-based approach to the study of a complex type of language variation. Our focus is on contact-induced semantic shifts in Quebec English, i.e., preexisting English words used with a meaning typical of a similar French word. Consider the following example taken from a tweet posted by a speaker from Montreal:

(1) I really want to go to an art museum or an art **exposition**.

Here, the word *exposition* refers to what is usually known as an *art exhibition*. This meaning is not conventionally used in English; it is instead associated with the homographous French word *exposition*.

This phenomenon is explained by the local sociolinguistic context: Quebec is the only predominantly French-speaking Canadian province. As of 2021, 74.8% of its inhabitants – close to 6.3 million people – report that their mother tongue is French. Ten times fewer Quebecers – 7.6% of the population, or just under 640,000 individuals – are native speakers of English (Statistics Canada 2022). This constitutes an ongoing situation of language contact between the varieties of English and French spoken in

Quebec. As we more extensively discuss below, its effects on the Quebec English lexicon are documented in the sociolinguistic and corpus linguistic literature, but these reports are often qualitative or anecdotal. We aim to complement them with a more comprehensive empirical description of contact-induced semantic shifts. From a methodological standpoint, we draw both on well-established principles of variationist sociolinguistics (Labov 1972; Tagliamonte 2006) and on new types of corpora and computational tools.

Our analyses rely on a particular type of linguistic data – a large, custom-built corpus of tweets – as well as a recent computational approach to modeling lexical semantic phenomena – neural word embeddings. Large-scale computational studies of language variation increasingly rely on both of these methodological choices (Nguyen 2021; Tahmasebi et al. 2021). Although they are routinely presented as promising alternatives to smaller-scale and often manual corpus analyses, they come with their own challenges, and their potential for linguistic description is not yet clear (Boleda 2020: 218). In applying them to a highly specific linguistic phenomenon, we aim to provide a more thorough understanding of their contributions, recurrent practical challenges, and potential solutions.

More specifically, we analyze patterns of regional variation in our corpus of tweets; drawing on underlying demographic distinctions, we aim to isolate

instances of contact-induced semantic shifts and further characterize their use. For a given lexical item, our computational system produces computational representations of its occurrences in the corpus and splits them into groups based on semantic similarity, prioritizing those that are expected to reflect the influence of French. Despite extensive data filtering and a carefully adapted implementation of a recent neural language model, our system highlights not only contact-induced semantic shifts, but also a range of noise-related phenomena. While this means that it cannot provide reliable results in an unsupervised manner, we show that a coarse manual annotation – conducted on the automatically identified clusters of tweets rather than individual occurrences – provides an efficient way of eliminating false positives. Complementing an earlier analysis of the technical impact of these issues (Miletic et al. 2021), we provide a more extensive, qualitative description of recurrent problems in the data. We further use the filtered quantitative information in a case study exploring the sociolinguistic constraints on contact-induced semantic shifts.

The remainder of this article is organized as follows. We first introduce a summary of related work (Section 2) and a more detailed description of the data and methods we deployed (Section 3). We then present the key results of our analysis (Section 4) and conclude with a discussion and main takeaways (Section 5).

2. Theoretical and methodological background

In this section, we contextualize our work with respect to research on semantic shifts in Quebec English, which represent our descriptive focus. We further discuss the basis for our key methodological choices, addressing the use of Twitter corpora and neural word embeddings.

2.1. Semantic shifts in Quebec English: the need for corpus studies

The use of English in Quebec is influenced by ongoing contact with French, particularly on the lexical level. Evidence for this claim comes from sociolinguistic studies (e.g. Boberg 2005; 2010; 2012; Boberg & Hotton 2015; Chambers & Heisler 1999; McArthur 1989; Rouaud 2019), as well as corpus-based research conducted mainly on newspaper texts (e.g. Fee 1991; 2008; Grant-Russell 1999; Miletic 2019). The reported contact-related phenomena include loanwords (e.g. *dépanneur* 'corner store'), loan translations (e.g. *all-dressed* '(pizza) with all the toppings', cf. Fr. *toute garnie*), and semantic shifts (e.g. *animator* 'group leader', cf. Fr. *animateur*) (Boberg 2012: 501). The importance of these phenomena has been contested due to their overall rarity in recorded speech (Poplack et al. 2006), but they may carry high symbolic value (Boberg 2012: 495). Moreover, large-scale dialect surveys have investigated regional lexical preferences more

systematically, reaffirming that “Montreal appears to be the most lexically distinct region in Canada” (Boberg 2005: 36). This is largely due to French lexical influence, with similar patterns reported elsewhere in Quebec (Boberg 2010: 170–188; Boberg & Hotton 2015).

These observations, however, are mainly based on studies of French loanwords; descriptions of other types of contact-related lexical influence are considerably more limited. This is particularly the case for the previously mentioned issue of semantic shifts, on which we focus in this paper. We understand this phenomenon as the presence of a sense in a preexisting English word that is explained by the presence of the equivalent sense in a formally and/or semantically similar French word. We know from McArthur’s (1989) written survey that the acceptability of semantic shifts varies across lexical items; cases such as *animator* ‘group leader’ and *collectivity* ‘people as a whole, community’ enjoy broader support than *library* ‘bookstore’ and *demand* ‘to ask for something’ (McArthur 1989: 42).[1] Corpus-based research has further described a range of linguistic mechanisms at play, including effects on the word’s connotation (e.g. *functionary* shifting from negative to neutral connotation) and degree of formality (e.g. *ameliorate* shifting from formal to neutral register) (Fee

[1] The examples are related to the French lexical items *animateur*, *collectivité*, *librairie*, and *demander*, respectively. Their meanings are affected by language contact to variable degrees; the most fine-grained distinction is related to *demand* being used in very general contexts, without the notion of authority with which it is usually associated in English.

2008: 181). There is also some evidence that the use of semantic shifts is subject to regional variation (Fee 1991; Miletic 2019) and affected by the speakers' degree of bilingualism (McArthur 1989).

Most of these studies rely on traditional sociolinguistic methods, which involve recording the speech production of carefully sampled speakers using face-to-face interviews (Labov 1972; Tagliamonte 2006) or written questionnaires (Dollinger 2015). However, this provides insufficient data for a systematic analysis of lexical semantic phenomena in spontaneous communication. For instance, Rouaud (2019: 245) describes the use of *campaign* with the sense of the French lexical item *campagne* 'countryside', which occurs only once in her sociolinguistic interviews, precluding generalizations. Another line of work discussed above overcomes this issue of data sparsity by analyzing larger collections of written documents, but it still faces challenges inherent in manual semantic analyses of corpus data. This is a crucial but highly time-consuming process, aggravated by the fact that human annotators may struggle to determine the attested meaning of a lexical item (Miletic 2019). In summary, the existing studies provide dozens of valuable examples of semantic shifts from diverse data sources, but given the methodological challenges outlined above, they are generally qualitative and often anecdotal in nature. As a result, we do not have reliable quantitative estimates of the impact of specific factors on the use of semantic shifts, or of their diffusion in the local speech community. Other

types of corpora and analytical approaches may therefore be better suited to this issue.

2.2. Twitter-based corpora for language variation

The development of social media has enabled large-scale analyses of language variation relying on publicly available posts from these websites. This is particularly true of Twitter, a social network created in 2006, where users can post 280-character messages known as tweets.[2] In addition to their linguistic content, tweets carry metadata such as the automatically identified language of the message and the user's geographic location at the time of posting (if the user chooses to include it). Each user has a profile page, where they can provide further information, such as a description and an additional, profile-level location. Users can interact with the content posted by others by liking it or replying to it, as well as reproduce it by retweeting it. They can form ties with other users by following them, and as a result regularly see their posts in their own timelines. The public availability of vast amounts of linguistic data, coupled with geographic information and the possibility to analyze patterns of interaction, is a key reason driving the use of Twitter-based corpora in studies of language variation.

[2] The maximum length was 140 characters until November 2017.

This frequently consists in analyzing regional patterns of language use reflected by geotagged posts, often in a bottom-up manner. As an example, Grieve et al. (2018) examine significant rises in frequency over time to semi-automatically identify lexical innovations in American Twitter (e.g. *amirite* 'am I right?'). They then map the spread of these items across the United States, observing five distinct geographic patterns of diffusion. Other related studies have similarly focused on identifying regionally specific lexical items (Donoso & Sánchez 2017; Shoemark et al. 2017) or interpreting language variation in terms of demographic factors (Bamman et al. 2014; Jones 2015).

The sheer amount of data available on Twitter is routinely presented as a key advantage compared to traditional sociolinguistic studies. However, this is counterbalanced by issues such as a lack of reliable demographic information – a mainstay of sociolinguistic research – as well as sources of bias inherent to the platform, affecting key information such as user location (Pavalanathan & Eisenstein 2015). Despite this, the patterns of regional variation derived from Twitter-based corpora broadly coincide with those obtained using traditional dialect surveys (Grieve et al. 2019). This suggests that Twitter corpora are a promising source of information in studies of language variation, especially where large amounts of data are required.

2.3. Vector space models for lexical semantic variation

As previously suggested, the persistent challenges in systematic analyses of lexical semantic variation – in sociolinguistic studies in general (Durkin 2012) and in those on Quebec English in particular – can be attributed not only to the large amount of data required for meaningful quantitative estimates of lexical semantic phenomena, but also to practical difficulties in investigating those phenomena at scale. It may be possible to overcome these issues by automatizing key aspects of semantic analyses. One way to do so consists in using vector space models (VSMs), computational tools that represent a word's meaning as a vector, essentially a list of numbers reflecting the word's cooccurrence statistics in a corpus (Turney & Pantel 2010). These models are rooted in the principles of distributional semantics, and most importantly the assumption that words appearing in similar linguistic contexts have a similar meaning (Firth 1957; Harris 1954). Importantly, VSMs provide the ability to measure the distance between two vectors, which is taken to reflect the difference in meaning of the words represented by them; this in turn facilitates systematic corpus-based studies of lexical semantics (Boleda 2020). While different implementations exist, we focus on a recent approach based on pretrained deep neural networks, the foremost among them being BERT (Devlin et al. 2019). Although originally designed for more complex NLP tasks, they also provide an efficient way of obtaining vector representations of word meanings, also known as word embeddings. They produce a contextually-informed vector

for each occurrence of a given word, allowing for analyses of phenomena such as polysemy.

Methods such as these constitute the cornerstone of recent computational approaches to semantic change. Starting from a diachronic corpus, different VSMs have been used to quantify the change in meaning of all words in a corpus (or any subset of them) over time (see Tahmasebi et al. 2021 for a detailed overview). This includes representations produced by BERT and related approaches, which generally estimate semantic change by analyzing differences in sense usage across time (Giulianelli et al. 2020; Martinc et al. 2020; Montariol et al. 2021). A variety of VSMs have been used to analyze meaning variation in different types of synchronic data (Del Tredici & Fernández 2017; Schlechtweg et al. 2019) and from a contact-linguistic perspective (Takamura et al. 2017; Uban et al. 2019).

However, most of these studies focus on computational issues. Existing descriptive applications include assessing longstanding hypotheses on semantic change (Xu & Kemp 2015) and facilitating corpus analyses by domain experts (De Pascale 2019; Rodda et al. 2017). This type of work is indicative of the descriptive potential of these methods, but their usability in linguistic research is yet to be fully demonstrated (Boleda 2020: 218). And while recent quantitative evaluations have identified the best-performing computational setups (e.g. Schlechtweg et al. 2020), practical issues like the

impact of noise in the data have not been extensively addressed. Echoing this situation, our previous work has shown that standard semantic change models can be applied to contact-induced semantic shifts in Quebec English, but that the resulting corpus-based analyses nevertheless entail important practical challenges (Miletic et al. 2021). In this paper, we examine the issues we encountered in more detail and discuss potential ways to overcome them.

3. Data and method

Our approach relies on contrasting synchronic data from different Canadian regions, under the assumption that linguistic behaviors that are specific to Quebec, but absent from areas where the use of French is limited, are likely to reflect the influence of language contact. This is inspired by the comparative sociolinguistic approach (Tagliamonte 2002), where differences in speech across communities are taken to reflect the sociodemographic differences in their composition. This section presents the main data and methodological choices on which we relied: a corpus of tweets posted in Canada; a curated list of contact-induced semantic shifts; an implementation of a neural word embedding model; and a manual annotation procedure. A high-level overview of our method is presented in Figure 1.

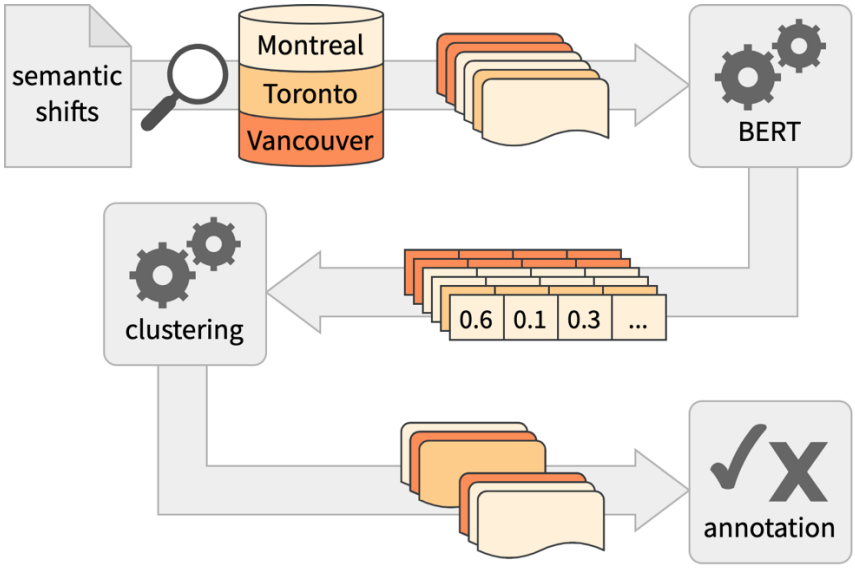

Figure 1. Method overview. For each semantic shift, examples are extracted from a regionally stratified corpus, modeled in the form of vector representations, clustered based on semantic similarity, and annotated for contact influence.

3.1. A corpus of tweets

We use a previously created corpus of Canadian English tweets published by users from Montreal, Toronto, and Vancouver (Miletic et al. 2020). Unlike other corpora of Canadian English, it is both sufficiently large for data-intensive methods such as VSMs, and it contains information on the regional origin and authors of individual tweets. The three cities are roughly comparable in terms of broad demographic properties which may impact sociolinguistic dynamics, such as the size and diversity of the population. A

key differentiator is the proportion of French speakers: they constitute a strong majority in Montreal, and a fraction of the population in the other two, predominantly English-speaking, cities. We therefore expect linguistic patterns specific to Montreal to be related to the influence of French. However, we cannot be certain of that, so we also rely on the structure of our corpus to distinguish contact phenomena from unrelated types of variation, such as regional trends that do not originate in Montreal (which can be identified using the two control regions) and strong idiolectal preferences (which we can isolate by tracing individual speakers).

The data were collected from January to November 2019. We initially used Twitter's Search API to look up tweets tagged as written in English and associated with the geographic area of one of the target cities. The users identified in this way were narrowed down to those whose free-text profile location strictly corresponded to one of the three target cities. We then crawled their profiles, collecting up to 3,200 most recent tweets per user; this allowed us to increase the amount of data and obtain basic sociolinguistic information. For instance, we stored the distribution of language tags in the users' tweet production and subsequently used it as a rough estimate of their degree of bilingualism. Finally, we only retained the tweets tagged by Twitter as written in English, and we automatically removed near-duplicates posted by individual users. Exploratory analyses

have shown that the retained data are both specific to the target cities and comparable across them (for more details, see Miletic et al. 2020).[3]

In addition to the preprocessing decisions applied to the original corpus, we introduced additional filtering for the experiments presented in this paper. First, we removed the content posted before 2016 in order to limit the likelihood of picking up diachronic effects; the tweets in the original corpus date back to 2006. In determining the cut-off point, our aim was to find a reasonable trade-off between a reduction in time span and the remaining amount of data. We then excluded all users with fewer than 10 tweets in the corpus. A maximum of 1,000 tweets per user were retained, with random subsampling performed where this was exceeded. This ensured that the corpus was not dominated by very few highly active individuals: an average of 229 tweets were retained per user, with the top 1% of users accounting for 4% of tweets. The corpus was tokenized and POS tagged using *twokenize* (Gimpel et al. 2011; Owoputi et al. 2013) and lemmatized using the NLTK WordNet lemmatizer (Bird et al. 2009). The structure of the final corpus is presented in Table 1.

Table 1. Corpus structure

[3] In accordance with Twitter's developer terms, the corpus is released in the form of tweet IDs, which can be used with off-the-shelf software to collect the underlying data: http://redac.univ-tlse2.fr/corpus/canen.html

Subcorpus	Users	Tweets	Tokens
Montreal	54,726	11,318,184	193,228,246
Toronto	51,245	12,465,659	222,508,471
Vancouver	47,697	11,381,080	213,200,523
Total	153,668	35,164,923	628,937,240

3.2. A set of semantic shifts in Quebec English

In order to examine a manageable number of semantic shifts in greater detail, we started from a previously established list of lexical items whose meaning is subject to contact-related influence in Quebec English, and which are attested with that meaning in our data. We specifically used an 80-item test set constructed for a technical evaluation of semantic change models on our data (Miletic et al. 2021).[4] Since we investigated the models' ability to distinguish words that do or do not undergo change, half of the items correspond to contact-induced semantic shifts, and the remaining half to stable words; given the scope of this paper, we focus on the former.

In identifying the semantic shifts, we relied on descriptions provided in the literature on Quebec English (Boberg 2012; Fee 1991; 2008; Grant 2010a; Josselin 2001; McArthur 1989; Rouaud 2019) as well as initial manual

[4] The full test set is available at http://redac.univ-tlse2.fr/misc/canenTestset.html

exploration of the Twitter corpus. In order to ensure the availability of sufficient data for a meaningful quantitative interpretation, we only retained the items occurring at least 100 times in each subcorpus. A concordance-based analysis was then used to determine if the items presented at least one contact-related occurrence in the Montreal subcorpus; those that did not were excluded. In establishing the potential contact-related use, the existing descriptions and corpus-based observations were complemented with lexicographic evidence.[5] This process resulted in a final list of 40 semantic shifts retained after all filtering steps; their mean frequency in the entire corpus is 5,268 (min = 345, max = 97,188). The items are summarized in Table 2.

Table 2. Set of semantic shifts with posited contact-related senses and sociolinguistic sources on which we principally relied to identify them. Items without provided sources were identified through manual corpus exploration.

Item	**Shifted sense**	**Sources**	**Item**	**Shifted sense**	**Sources**
affirmation	claim, statement	F_1	*hesitate*	deliberate btw. options	—
ambiance	atmosphere, vibe	—	*laureate*	winner	—
animator	activity/team leader	B F_1 G_1 M	*local* (n)	room, site, premises	M
availability	(pl.) available times	—	*manifestation*	protest, demonstration	G_1
boutique	shop, store	—	*merit*	deserve, be worthy of	—

[5] Canadian Oxford Dictionary (Barber 2004); Dictionary of Canadianisms on Historical Principles (Dollinger & Fee 2017); Trésor de la langue française informatisé (Dendien & Pierrel 2003); Usito (Cajolet-Laganière et al. 2014).

chalet	summer cottage	B F_2 G_1	*militant*	activist, campaigner	G_1
circulation	traffic	F_1	*nomination*	appointment to a role	G1
coordinate	(pl.) contact details	G_1	*occasion*	chance, opportunity	—
deceive	disappoint	F_2 M R	*pass* (v)	stop by	B
deception	disappointment	—	*permit*	driver's license	—
definitively	definitely, certainly	—	*population*	people, general public	F_1 F_2 G_1 R
deputy	member of parliament	R	*portable*	cell phone, laptop	G_2
dossier	issue, portfolio	F_2 G_1 R	*proposition*	suggestion, proposal	—
entourage	family circle, friends	—	*prudent*	careful	—
exchange (v)	talk with someone	—	*remark* (v)	notice	M
exploration	study of sth. little known	—	*reparation*	repairs (of a device)	—
exposition	art exhibition	F_1	*resume*	summarize	M R
formation	training, course	B M R	*souvenir*	memory	M R
formidable	great, terrific	—	*terrace*	patio, eating area	F_2 R
grave (adj)	highly important	—	*trio*	sandwich-fries-drink	B

Sources: **B** = Boberg (2012); **F_1** = Fee (1991); **F_2** = Fee (2008); **G_1** = Grant (2010a); **G_2** = Grant (2010b); **M** = McArthur (1989); **R** = Rouaud (2019).

The meaning of these items, as attested in a range of occurrences from the corpus of tweets, was then computationally modeled. Our aim was to automatically identify occurrences used in similar contexts and quantify sense distributions, focusing on both regional and user-level patterns.

3.3. Neural word embeddings

For each of the 40 lexical items from Section 3.2, we first produced word embeddings for their individual occurrences; each corresponded to a slightly different vector, informed by its immediate linguistic context. We then used

these representations to automatically group the occurrences into clusters, which were expected to reflect similar contexts (and thereby similar uses of the target lexical item). This allows for a more efficient subsequent analysis of the full range of uses exhibited by a lexical item: for instance, the fact that similar occurrences are grouped together means that it is not necessary to disambiguate them one at a time.

Word embeddings were produced using the previously discussed BERT model, and specifically the Hugging Face implementation (Wolf et al. 2020) of *bert-base-uncased*, a 12-layer, 768-dimension version pretrained on generic English data.[6] Models such as this are often adapted to specific applications using fine-tuning, i.e., partial retraining on a specific task. Similarly to some existing work (e.g. Giulianelli et al. 2020), we did not perform fine-tuning given the assumption that word senses are reflected by differences in immediate linguistic context, which the pretrained model should be able to capture.

For each analyzed lexical item, we extracted the tweets in which it appears in all three regional subcorpora. In order to limit processing and memory requirements, we retained no more than 1,000 total occurrences per word,

[6] A version of the model specifically adapted to tweet processing (BERTweet; Nguyen et al. 2020) was released after our experiments were implemented, but is relevant for future work using the same type of data.

and used a random sample for more frequent items. We fed each tweet, in its raw text form, as a single sequence into BERT, which then produced context-informed vectors for each token in the tweet. The model outputs multiple vector representations per token, each corresponding to a different hidden layer in the neural network architecture. Similarly to other recent studies (e.g. Laicher et al. 2021), we averaged over the last four hidden layers to obtain a single token-level vector. BERT's tokenizer splits some words into subparts with separate representations; when this occurred, we averaged over the subparts to produce a single vector. The embedding computation procedure is schematically presented in Figure 2.

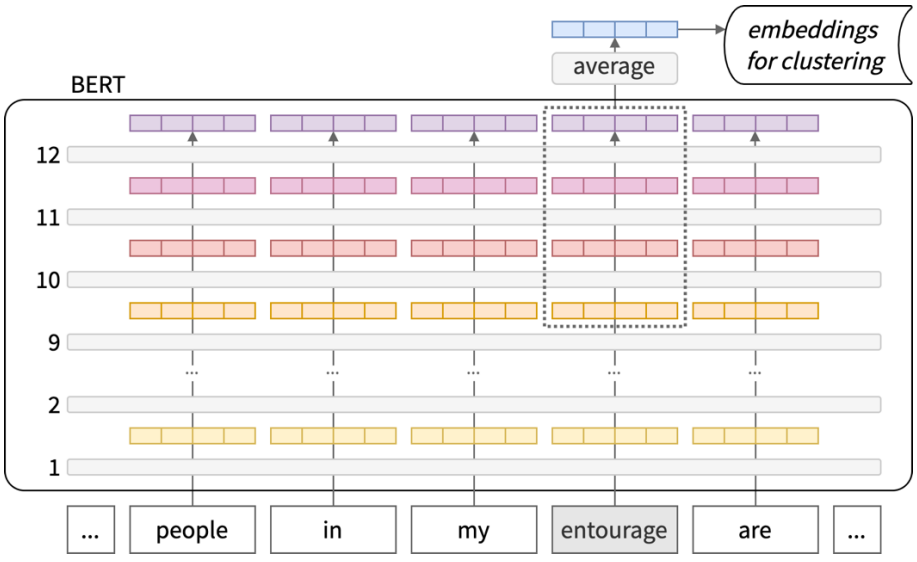

Figure 2. Embedding computation for an occurrence of *entourage*. Each computational layer (grey bars) outputs contextualized embeddings for each

token in the sequence. We average over the last four embeddings for the target token.[7]

3.4. Clustering and annotating the uses of a lexical item

Similar uses of a lexical item were automatically identified by clustering its token-level vectors using affinity propagation, an algorithm which performed well in other semantic change studies (e.g. Martinc et al. 2020). It does not require that the number of clusters be specified beforehand, and it produces clusters of variable size, making it well-suited for studying sense distributions. We used the *scikit-learn* implementation (Pedregosa et al. 2011) with default parameters, which is based on the negative squared Euclidean distance. Data from the three regional subcorpora were clustered at the same time, meaning that a single cluster may contain tweets from all three cities. This allows us to examine how each cluster's occurrences are distributed across the regions.

In analyzing the output of the analysis, we considered the clusters containing at least five tweets, and retained them if more than half of the tweets were from the Montreal subcorpus. This is because of the focus on the uses which are clearly more frequent in Montreal than elsewhere, but

[7] Figure inspired by Jay Alammar's illustrations available at http://jalammar.github.io/illustrated-bert/

which may occasionally appear in other regions. Up to 10 such clusters were retained for each lexical item, starting with those with the highest proportion of Montreal tweets. The data for the 40 retained lexical items were then manually annotated at the level of clusters. We used binary labels, and established if a cluster presented a contact-related sense based on the majority usage in it.

More specifically, a target item's use in a cluster was annotated as contact-related if it was regionally specific to Montreal and potentially explained by the influence of a formally and/or semantically related French word. This determination relied on the same evidence used to select the target set of semantic shifts, i.e., previous sociolinguistic studies and lexicographic sources (see Section 3.2). Recurrent phenomena that were not annotated as contact-related included a range of noise-related issues; these will be discussed in more detail below. A 15-word sample (91 clusters) was annotated by two annotators in order to test the reliability of the general procedure, obtaining a reasonably high inter-annotator agreement (Cohen's kappa coefficient of 0.55).

On average, 8 clusters per word (min = 3, max = 10) were retained for annotation. The mean number of tweets per cluster stands at 13 (averaged over the means for individual lexical items; min = 8, max = 20). As shown by the examples discussed below, the clusters are largely homogeneous;

although some are occasionally difficult to interpret, this is overall rare. Our analysis also allows for a degree of uncertainty, as the annotation targets the predominant use in a cluster. The utility of this approach is confirmed by the fact that it led to the identification of at least one contact-related cluster for each of the 40 target items. From a practical standpoint, using cluster-level annotations was an order of magnitude faster than analyzing individual tweets. This is due to the lower number of required decisions and the comparative ease in determining the meaning of a larger number of similar examples appearing together.

4. Results

This section discusses the results derived from the annotated data. It first presents a general overview of cluster types across lexical items; it then illustrates a range of true and false positives observed in the data; and it concludes with a case study examining the link of contact-induced semantic shifts with bilingualism.

4.1. An overview of regionally specific clusters

A global overview of the analysis (Figure 3) outlines the distribution of annotated tweets for the 40 target lexical items based on the annotated usage

types. Regionally specific clusters may capture the effect of language contact, but this is not systematic: contact-related use was observed in at least one cluster for each item, but it was prevalent in a minority of cases. The proportion of contact-related tweets ranges from seemingly clear-cut cases such as *definitively* ‘definitely’ (100%) and *terrace* ‘outside eating area; patio’ (90%) on one end of the spectrum, to *animator* ‘group leader’ (6%) and *reparation* ‘repairs’ (4%) on the other; the median value is 26%. In other words, regional variation appears to be helpful in guiding the analysis of contact-induced phenomena, but it is not sufficient in isolating linguistic behaviors of interest.

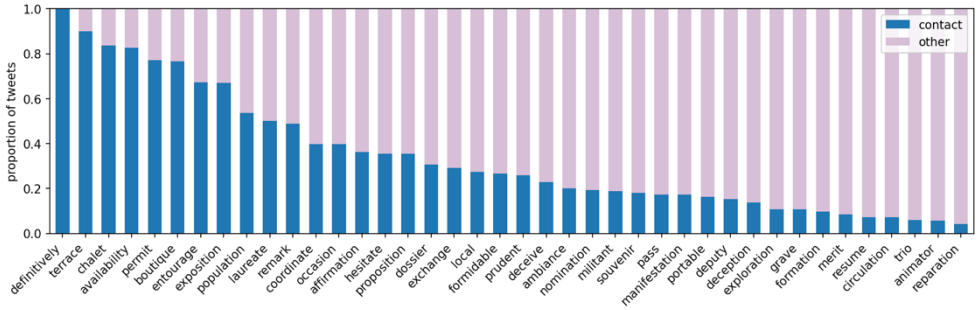

Figure 3. Distribution of annotated tweets from Montreal-specific clusters across the binary labels. For example, *terrace* had the contact-related meaning ‘patio’ in 90% of tweets (blue bar), and other meanings in 10% of tweets (pink bar)

Furthermore, a wide variety of linguistic phenomena is subsumed under the binary distinction in the overview plot. We now provide a more detailed,

qualitative analysis of both contact-related and non-contact-related uses, which can respectively be viewed as true and false positives output by our computational system.

4.2. Types of variation captured by the analysis

This section discusses examples of tweets extracted from our corpus using the clustering analysis described above. Sample clusters of tweets are presented in the keyword-in-context format, for ease of reading as well as to illustrate the effect of this approach on manual perusal of corpus data, as observed during the manual annotation. Each sample cluster contains three representative tweets published in Montreal and occurring in a single original cluster output by our analysis. Further information on the size and regional composition of the clusters is also provided. In order to protect the privacy of tweet authors, we only reproduce textual content without any metadata. For the same reason, usernames, hashtags, URLs, and names of individuals are redacted from the tweets, except for widely known public figures or if necessary to interpret the meaning of the tweet.

4.2.1. True positives

We begin by examining positive contributions of our computational system, focusing on the help it provided in distinguishing between conventional and contact-related uses based on documented patterns from the corpus. This

was beneficial across different semantic mechanisms and degrees of granularity of contact-related influence.

A clear-cut distinction. Perhaps the prototypical mechanism underlying contact-induced semantic shifts involves using an English lexical item to denote a referent conventionally designated by a formally similar French lexical item. One such example is *manifestation*, which is generally used to signify 'a display of the existence of something', but is also attested in Quebec English with the sense of 'protest, demonstration', typical of the homographous French lexical item *manifestation*. This sense is absent from the Canadian Oxford Dictionary (COD), but it is anecdotally reported by Grant (2010a: 187). It is also recorded in the Oxford English Dictionary (OED), but the most recent example (1978) comes from New Brunswick, an officially bilingual Canadian province bordering Quebec, and contains metalinguistic commentary supporting a link with language contact. Our corpus-based analysis, partly presented in Table 3, provides empirical evidence of the ongoing use of both senses; the French-related one is strongly regionally specific to Montreal. The first cluster in the table corresponds to the contact-related sense of 'demonstration', which can be interpreted based on spatio-temporal references as well as near-synonyms (e.g., *walk*); the second cluster corresponds to the conventional sense of 'existence', reflected by the medical topic.

Table 3. Sample clusters for *manifestation*. Contact-related use (top): 9 out of 9 tweets (100%) from Montreal; conventional use (bottom): 8 out of 13 tweets (62%) from Montreal.

Montreal's	manifestation	in progress at the immigration office
Pro refugee	manifestation	at the Big O . <user> <user>
Quebec's history . This walk is the biggest	manifestation	for this week . And 52 more
locating colonial trauma in genetic	manifestation	of health issues , all the funding goes into
ask yourself how they might affect clinical	manifestation	of disease . <hashtag>
is an emergency problem it's local	manifestation	for a systemic issue

A subtler distinction. A more nuanced type of contact-related semantic influence is illustrated by the case of *entourage*. In English, this lexical item generally has the highly specific sense of 'retinue, people surrounding an important person', whereas in French it is also used with the more general sense of 'family circle' or 'group of friends'. This sense is not recorded in either the COD or the OED, and we are unaware of any sociolinguistic descriptions of it. However, our corpus data point to the existence of both senses in Quebec English (Table 4). They additionally illustrate the fact that the contact-related sense involves a generalization of the preexisting English sense. This implies a partial overlap and hence a less striking difference between the two, which is often only apparent from contextual cues. The first cluster in the table captures the 'family circle' sense; this is suggested by the possessives associated with *entourage*, which point to the speaker or their interlocutor. The second cluster corresponds to the 'retinue' sense, as

shown by third person references as well as background knowledge on the people whose entourage is being discussed (e.g., sports players and politicians). Fine-grained distinctions such as these underscore the complexity of the phenomenon under study and the need for linguistic expertise; the fact that they can be clearly established based on the output of our computational system confirms its value in guiding empirically grounded analyses.

Table 4. Sample clusters for *entourage*. Contact-related use (top): 12 out of 13 tweets (92%) from Montreal; conventional use (bottom): 17 out of 21 tweets (81%) from Montreal.

learn it on your own and have people in your	entourage	help you practice . Personally , I tried taking
have more than 30 people (coworkers) in my	entourage	who are new fan of Harry but they don't know
hateful shit . But when people from your	entourage	excuse someone’s homophobic actions ; it
@MariaSharapova where are her	entourage	? Managers ...? Don't they check for her ?
Damage control brought to you by POTUS	entourage	. He's been golfing and hamming it up with
I really feel like somebody from his	entourage	told him " oh yes , best we have ever seen "

4.2.2. False positives

We have so far focused on the informativeness of our semi-automated analysis in understanding often fine-grained patterns of contact-related language use, but this process was complicated by different types of false positives. This section provides a detailed analysis of the most frequent patterns that we encountered. We distinguish between the following types of

locally specific word usage which do not constitute contact-induced semantic shifts:

- *cultural effects*, where word usage is related to the local cultural context of Montreal;
- the use of common nouns as *proper names* denoting locally specific referents;
- *French homographs* of English words, attested in codeswitched tweets;
- *structural patterns*, such as the position of the target item across tweets, which accidentally affect model performance.

For each category that we discuss, sample clusters of tweets are provided in order to illustrate the contrast between the contact-related use – the target of our analysis – and the noise that we identified along the way.

Cultural effects. The regionally specific character of some clusters output by our analysis is not related to the use of the target lexical items with a French-related sense, but rather to the local cultural context of Montreal. Take for example *formation*, whose English senses include 'the action or process of forming' and 'arrangement or disposition'. In our data, it is also attested with the sense of 'course, training program', typical of the French homograph *formation*. This sense is absent from the COD and the OED, but its existence is noted in the sociolinguistic literature (McArthur 1989: 57;

Boberg 2012: 497). As shown in Table 5, our computational system enabled us to identify corpus-based evidence of its use (top cluster). But it also accidentally highlighted its involvement in topical variation (bottom cluster), related to the widespread use of *formation* 'disposition of players in a sports team'. The regional specificity of this cluster is not explained by language contact; the one apparent characteristic shared by these occurrences is discussion of local sports teams. This is evidenced by the names mentioned in the tweets: at the time of posting, Dominic Oduro and Donny Toia were players for the Montreal Impact soccer team; Mauro Biello was manager of the same team; and Carey Price and Al Montoya were goaltenders for the Montreal Canadiens hockey team. This is also a striking illustration of a more general practical challenge in analyzing Twitter data: detailed background information on a variety of topics – in this case, sports – is often required to fully grasp the local significance of the attested language use.

Table 5. Sample clusters for *formation*. Contact-related use (top): 5 out of 7 tweets (71%) from Montreal; cultural effects (bottom): 9 out of 14 tweets (64%) from Montreal.

, the company I work at gave us a quick	formation	and that was it !
Little sister is taking on a	formation	to be a game tester and I am going to univ to
realize that ??? Well i have internet on my	formation	but its a shitty internet .. so i probably not be
Oduro isn't strong enough defensively for this	formation	. Toia not strong enough offensively , but
room for only 1 deep ball retriever . What	formation	should Biello use with a full lineup ? 2/2

The goalie formation is solid . even if Price gets injured , Montoya

A different but related effect was observed in the case of *animator*. This lexical item is generally used with the sense of 'creator of animated films', whereas the formally similar French equivalent *animateur* also includes the sense 'group leader; organizer; facilitator'. The use of *animator* with the former senses is attested in the sociolinguistic literature (McArthur 1989: 53; Grant 2010a: 187; Boberg 2012: 497). It is also noted in dictionaries, including the OED, and is marked as being related to French (COD) and specific to Quebec and New Brunswick (Dictionary of Canadianisms on Historical Principles). Our system found some occurrences of the contact-related use, but these were limited to a single cluster; the remaining eight regionally specific clusters reflected the conventional sense (Table 6). This is likely due to Montreal's status as a global center for the animation industry, which would explain a larger number of conversations on this topic compared to the two control cities, making this use more prominent for our computational system.

Table 6. Sample clusters for *animator*. Contact-related use (top): 4 out of 6 tweets (67%) from Montreal; cultural effects (bottom): 10 out of 14 tweets (71%) from Montreal.

Mr. <name> , Spiritual and Community Animator , is organizing a Costa Rica trip for … <url>
Ms. <name> , Spiritual Animator , has been busy with the annual #PoppyDrive

created by <name> , Spiritual Community	Animator	. Firefighter <name> , Fire House Station 54
for an animation tech director , technical	animator	, and a VFX artist . I don't know what any of
SO many different jobs in animation , not just	animator	, and they're all essential to every single prod .
I'm a graphic and web designer turned 3d	animator	an motion designer . What I'd really like to do

Proper names. The regional specificity of some clusters is explained by the target lexical item being used as a proper name, generally to denote a regionally specific referent. Take for example *deception*, which in English refers to 'the action of misleading someone', but whose French homograph also means 'disappointment'. This use is not recorded in the OED or the COD, nor is it described in the sociolinguistic literature we reviewed.[8] Potential influence of French is, however, attested in our data (Table 7). But in addition to those reflecting the contact-related sense, some clusters specific to Montreal involve occurrences of the target lexical item in the phrase *Deception Bay*, the name of a song by the Montreal band Milk & Bone. This is explained by the origin of the band in question and is entirely unrelated to language contact.

Table 7. Sample clusters for *deception*. Contact-related use (top): 9 out of 10 tweets (90%) from Montreal; proper names (bottom): 7 out of 7 tweets (100%) from Montreal.

[8]This pattern is however anecdotally reported for the verb *deceive* 'mislead' and the French equivalent *décevoir* 'disappoint' (McArthur 1989: 17), as well as for the corresponding deverbal adjective *deceived* 'disappointed' (Rouaud 2019: 167).

be my year , but so far it's been nothing but	deceptions	and heartbreaks . That won't stop me from
Great expectations , few	deceptions	and stunning debuts make a unique
I had some very bad	deceptions	lately ... * coughs * The Technomancer ....
	Deception	Bay , the title track from @milknbone's
The new song	Deception	Bay , from Milk & Bone's second album , is
	Deception	Bay on repeat !! Can’t wait for the whole

French homographs in codeswitched tweets. The performance of our computational system is occasionally affected by crosslingual homographs of the target lexical items. They are generally used in a span of French text within a codeswitched tweet where most tokens are in English; this explains why the tweets were tagged as written in English and retained in the corpus. Codeswitching is overall rare in our corpus, but its relative frequency is considerably higher in Montreal, as can be expected given the prevalence of bilingual speakers in the city.

The practical implications are illustrated by the case of *souvenir*. Our analysis focused on the conventional English sense ‘keepsake, memento’ and the potential presence of the more abstract sense ‘memory’, typical of the formally identical French equivalent. The use of the English lexical item with the French-associated sense is not recorded in the COD. It is however attested in the OED, though only as “chiefly literary”, as well as in the sociolinguistic literature (McArthur 1989: 69; Rouaud 2019: 167). It is also present in our data, providing a corpus-based contribution to the existing descriptions (Table 8). But in addition to this positive result, three out of the

ten regionally specific clusters are entirely composed of codeswitched tweets containing the French homograph of the target lexical item. These instances do not correspond to our definition of semantic shifts – we are interested in English, not French, lexical items – and as such represent noise detrimental to our analysis.

Table 8. Sample clusters for *souvenir*. Contact-related use (top): 12 out of 12 tweets (100%) from Montreal; French homographs (bottom): 12 out of 16 tweets (75%) from Montreal.

which I had kept no memories . The only fond	souvenir	which still haunted me , her hand on my jaw .
ago Winter 2003 with my son Raphaël . Great	souvenir	. Let's… <url>
Montreal . We had VIP tickets . An amazing	souvenir	😭 ♡
year ago . Wonderful memories ! Quel beau	souvenir	! 🫛🫛
Francois Fournier a partagé un	souvenir	. 1 h · 7 years ago , i played my third gig with
Old memories . Very old . Oh , les vieux	souvenirs	! <url>

Structural patterns affecting model performance. A final recurrent issue is that of clusters where tweets appear to be grouped together based solely on structural regularities. This was observed in the case of *trio*, which conventionally means ‘a group of three’, but is also used with the sense of its Quebec French homograph, denoting a ‘sandwich-fries-soda special, combo’. This specific use is not attested in the lexicographic sources we consulted, but it is described in the sociolinguistic literature (Boberg 2012: 500). Our analysis provided corpus-based evidence for this use; however,

the second cluster shown in Table 9 illustrates the problem of uninformative results mentioned above. Here, the only characteristic common to the tweets is the fact that the target lexical item occurs at the end of the tweet, whose content is otherwise ambiguous. This is potentially explained by underlying issues with the way in which BERT calculates some vector representations. This may be exacerbated by short sequences of text, where positional information may carry excessive influence.

Table 9. Sample clusters for *trio*. Contact-related use (top): 8 out of 8 tweets (100%) from Montreal; structural patterns (bottom): 7 out of 10 tweets (70%) from Montreal.

I could legit eat 3 Big Mac	trios	right now I'm so hungry 😭
Customer : can i have a number 3	trio	? Cashier who used to be a middle school
I rarely order fries anymore , let alone a	trio	. If I did , however , I usually go fries first . (But
oh you know it man ! This is the new	trio	
got mr beefcake yay now I have the full ossan	trio	
2 Years ago , still the same goofy	trio	<url>

The examples discussed so far illustrate the variety of patterns of contact-related use attested in our data, for which we obtained valuable qualitative evidence thanks to the computational analysis we implemented. However, a range of regionally specific uses were explained by different types of noise rather than, as we had anticipated, language contact. The system in its current implementation cannot reliably identify contact-related use in an

unsupervised way. However, the coarse annotation we conducted facilitates manual corpus analysis, as shown throughout this section; it validates the existence of contact-related uses; and it helps exclude the patterns that are related to noise. We now draw on the resulting data to present a broader, quantitative overview of contact-related language use in the corpus.

4.3. Deploying coarsely annotated data for linguistic description

The structure of the clusters output by our analysis shows that lexical items differ in terms of the diffusion of contact-related usage (how many tweets are related to contact, out of all those retained in the regionally specific clusters) as well as its regional specificity (how many tweets in contact-related clusters come from Montreal). These patterns may be indicative of different degrees and factors of diffusion of semantic shifts within the local speech community.

To explore the descriptive utility of this information, we calculated scores reflecting the two points raised above for each of the 40 manually annotated lexical items: a *diffusion score*, corresponding to the proportion of tweets tagged as contact-related, out of all manually annotated tweets; and a *regionality score*, corresponding to the proportion of tweets posted in Montreal, out of all tweets tagged as contact-related. In order to explore the potential impact of the degree of bilingualism on the use of semantic shifts,

for each lexical item we also calculated a *bilingualism score*, corresponding to the mean proportion of tweets in English (out of tweets in English in French) posted by users who used the contact-related sense in the clusters tagged as such. It ranges from 0 for users tweeting only in French to 1 for users tweeting only in English, with intermediate values indicating a production of tweets in both languages.

We first checked the relationship between the three scores by calculating Spearman's rank correlation coefficient. The diffusion score is uncorrelated with both the regionality score ($\rho = -0.13$, $p = 0.42$) and the bilingualism score ($\rho = 0.02$, $p = 0.90$). However, the regionality and bilingualism scores exhibit a moderate negative correlation ($\rho = -0.53$, $p < 0.001$); this link is explored in more detail in Figure 4.

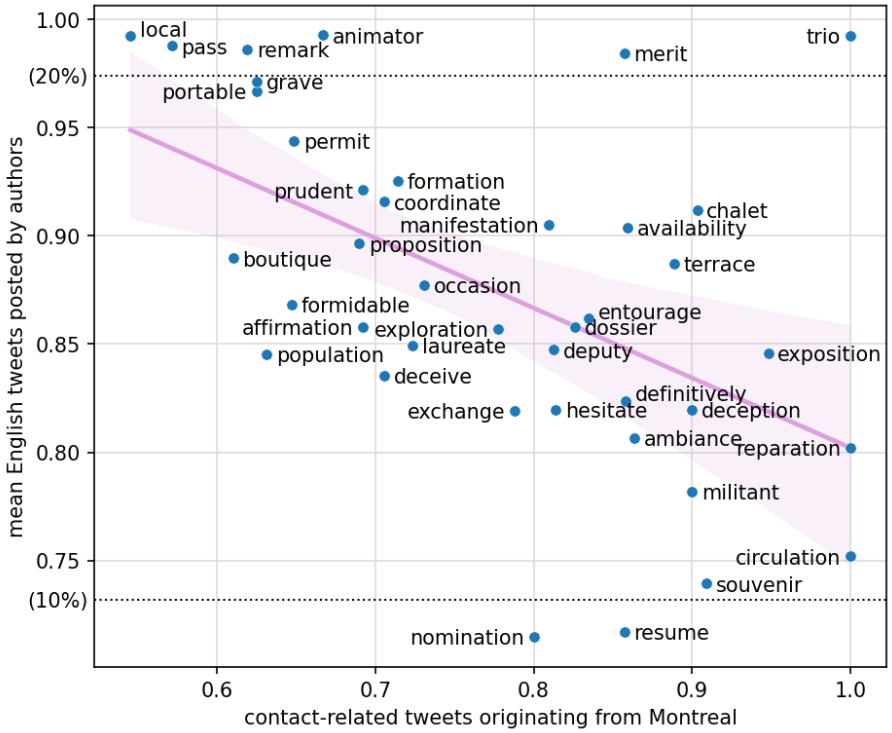

Figure 4. Regression plot showing the relationship between the regionality score (x-axis) and the bilingualism score (y-axis). Dotted lines show the 10th and 20th percentile for the bilingualism score, for all users in the corpus.

The plotted results indicate that contact-related semantic shifts which are more regionally specific (i.e., attested in Montreal to a higher extent) are also more directly related to the effects of bilingualism (i.e., a lower proportion of English, and hence a higher proportion of French, tweets). A typical example (bottom right) is the case of *circulation*, attested in the Quebec English data with the sense of 'traffic', which is associated with the corresponding French homograph. All of its tweets from clusters tagged as contact-related come from Montreal; moreover, the mean proportion of

English tweets stands at 0.75 per user. This may appear to be a relatively high value, but it is in fact just above the 10th percentile for all users in the corpus (0.73); at least within this dataset, this is suggestive of a comparatively important influence of bilingualism. Patterns at the other end of the spectrum (upper left) are illustrated by the verb *remark*; we focused on the sense 'notice', with which the French verb *remarquer* is widely used. It is less regionally specific (62% of contact-related tweets posted in Montreal) and less strongly associated with bilingualism (higher mean proportion of English tweets per user, at 0.99). Unlike in the previous example, however, the contact-related sense is attested in dictionaries, but the OED marks it as rare in some syntactic contexts. While it is likely accessible to most English speakers, cross-linguistic influence might facilitate its wider use; this scenario is consistent with our data.

It is also relevant to look at the outliers from the general trend. For instance, in the previously mentioned case of *trio* 'sandwich-fries-soda special, combo' (upper right in the plot above) all contact-related tweets similarly come from Montreal. However, the mean proportion of English tweets is higher, at 0.99 per user. This is indicative of a use which is regionally specific, but is widespread in the local linguistic community, including among monolingual speakers. This is further supported by existing descriptions which have shown it to be typical of the speech of native

English-speaking Quebecers (Boberg 2005: 36; Boberg & Hotton 2015: 307).

These observations indicate that, barring some exceptions, the more regionally specific the contact-related use is, the more strongly it is associated with use of French. It is important to once again note that the manual annotation was conducted on the level of clusters, rather than individual tweets, meaning that some non-contact-related occurrences may have been included in the counts. Moreover, the information on the use of French has the benefit of being empirically grounded in the attested use of languages by individual Twitter users, but it is only a very rough approximation of their linguistic profiles; for instance, there is no reliable way to determine their native language. That said, our analysis identified clear trends regarding the use of semantic shifts based on a large amount of data, further confirming the potential that corpus-based analyses have in understanding the patterns behind complex linguistic behaviors. It also constituted the basis of a face-to-face sociolinguistic survey we conducted in January 2022, the results of which will provide a clearer understanding of the links between corpus-based and in-person analyses of language variation.

5. Discussion and conclusion

We have presented an analysis of a fine-grained sociolinguistic phenomenon – contact-induced semantic shifts in Quebec English – using a large, custom-built corpus of tweets and a recent pretrained language model relying on a deep neural network architecture. This approach has paved the way for a more detailed account of previously reported semantic shifts, contributing extensive empirical evidence where original descriptions often consisted in a single anecdotal mention of a lexical item of interest; our approach was also beneficial in more comprehensively characterizing previously undescribed semantic shifts, initially observed in isolated tweets. The computational tools we used facilitated manual inspection of vast amounts of data, directing our attention to the most relevant subsets of occurrences; they also enabled broad quantitative estimates of the use of semantic shifts, highlighting possible interpretations and informing the design of subsequent studies. The results have also broadly confirmed our high-level assumption that regional variation in synchrony can be used as a proxy for detecting contact-induced phenomena. More generally, the overall setup – data extraction, clustering based on semantic similarity, and analysis of the distribution of occurrences over an explanatory factor – can be generalized to other descriptive issues.

However, we cannot gloss over the fact that our computational system provided actionable results only once it was complemented with extensive

manual analyses. The challenges that we encountered are related to several distinct issues: (i) a strong assumption on regional variation underpinning the methodological design – while some language use specific to Montreal is related to language contact, not all is; (ii) inherent limitations of the methods we used, with BERT occasionally capturing phenomena unrelated to lexical semantics; (iii) inherent limitations of the data we used, with a carefully filtered Twitter corpus representing an improvement on highly generic datasets, but still suffering from the 280-character limit and the limited ability to validate user descriptions, among other issues; (iv) the complexity of the phenomenon under study, which often involves very subtle – but nevertheless perceptible and socially meaningful – differences in language use. Some of the described false positives, such as French codeswitching and referents typical of Montreal, are specific to our corpus; however, they echo the observation that semantic change models capture different types of variation in word usage, also raised in other recent studies (Giulianelli et al. 2020; Hengchen et al. 2021; Tahmasebi et al. 2021).

Despite these challenges, data-intensive computational approaches to lexical semantic phenomena, and to language variation in general, have an important role to play in descriptive linguistic research. They can provide meaningful quantitative accounts of lexical phenomena, including for the whole vocabulary, based on data obtained in an unobtrusive way; this is clearly complementary to traditional sociolinguistic methods. While

methods such as those we implemented still require adaptations to the task at hand as well as some manual analysis, they simplify the tasks required of the linguist. One example of this approach is our analysis based on coarse cluster-level annotations; its relevance is confirmed by the fact that, together with the results presented in Miletic et al. (2021), this line of work represents the first systematic corpus-based analysis of contact-induced semantic shifts in Quebec English. The forthcoming results of a face-to-face sociolinguistic survey that this work enabled will shed further light on the relationship between computational descriptions and real-life sociolinguistic behaviors.

References

Bamman, David, Eisenstein, Jacob & Schnoebelen, Tyler. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2). 135–160. (doi:10.1111/josl.12080)

Barber, Katherine (ed.). 2004. *Canadian Oxford dictionary*. Don Mills: Oxford University Press.

Bird, Steven, Loper, Edward & Klein, Ewan. 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.

Boberg, Charles. 2005. The North American Regional Vocabulary Survey: New variables and methods in the study of North American English. *American Speech* 80(1): 22–60.

Boberg, Charles. 2010. *The English language in Canada: Status, history and comparative analysis*. Cambridge: Cambridge University Press.

Boberg, Charles. 2012. English as a minority language in Quebec. *World Englishes* 31(4): 493–502. (doi:10.1111/j.1467-971X.2012.01776.x)

Boberg, Charles & Hotton, Jenna. 2015. English in the Gaspé region of Quebec. *English World-Wide* 36(3): 277–314. (doi:10.1075/eww.36.3.01bob)

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6: 213–234. (doi:10.1146/annurev-linguistics-011619-030303)
Cajolet-Laganière, Hélène, Martel, Pierre, Masson, Chantal-Édith & Mercier, Louis. 2014. *Usito*. <https://usito.usherbrooke.ca/>
Chambers, J. K. & Heisler, Troy. 1999. Dialect topography of Québec City English. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 44(1): 23–48. (doi:10.1017/S0008413100020703)
De Pascale, Stefano. 2019. *Token-based vector space models as semantic control in lexical lectometry*. PhD dissertation, KU Leuven.
Del Tredici, Marco & Fernández, Raquel. 2017. Semantic variation in online communities of practice. In *IWCS 2017 –12th International Conference on Computational Semantics – long papers*. <https://www.aclweb.org/anthology/W17-6804>
Dendien, Jacques & Pierrel, Jean-Marie. 2003. Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues* 44(2): 11–37.
Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton & Toutanova, Kristina. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. (doi:10.18653/v1/N19-1423)
Dollinger, Stefan. 2015. *The written questionnaire in social dialectology: History, theory, practice*. Amsterdam: John Benjamins. (doi:10.1075/impact.40)
Dollinger, Stefan & Fee, Margery. 2017. *DCHP-2: The dictionary of Canadianisms on historical principles, second edition*. <http://www.dchp.ca/dchp2>
Donoso, Gonzalo & Sánchez, David. 2017. Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 16–25. Valencia, Spain: Association for Computational Linguistics. (doi:10.18653/v1/W17-1202)
Durkin, Philip. 2012. Variation in the lexicon: The 'Cinderella' of sociolinguistics?: Why does variation in word forms and word meanings present such challenges for empirical research? *English Today* 28(4): 3–9. (doi:10.1017/S0266078412000375)
Fee, Margery. 1991. Frenglish in Quebec English newspapers. In *Papers of the Fifteenth Annual Meeting of the Atlantic Provinces Linguistic Association*, 12–23. Atlantic Provinces Linguistic Association.

Fee, Margery. 2008. French borrowing in Quebec English. *Anglistik: International Journal of English Studies* 19(2): 173–188. (doi:10.14288/1.0074544)
Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1–32. Oxford: Blackwell.
Gimpel, Kevin, Schneider, Nathan, O'Connor, Brendan, Das, Dipanjan, Mills, Daniel, Eisenstein, Jacob, Heilman, Michael, Yogatama, Dani, Flanigan, Jeffrey & Smith, Noah A. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 42–47. Portland, Oregon, USA: Association for Computational Linguistics.
Giulianelli, Mario, Del Tredici, Marco & Fernández, Raquel. 2020. Analysing lexical semantic change with contextualised word representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3960–3973. Online: Association for Computational Linguistics. (doi:10.18653/v1/2020.acl-main.365)
Grant, Pamela. 2010a. English usage in contemporary Quebec: Reflections of the local. In *Canadian English: A Linguistic Reader* [Strathy Occasional Papers on Canadian English 6], Elaine Gold & Janice McAlpine (eds), 177–197. Kingston, ON: Queen's University.
Grant, Pamela. 2010b. *Is Quebec English distinct? English usage in contemporary Quebec* [lecture slides]. <https://slideplayer.com/slide/10010/>
Grant-Russell, Pamela. 1999. The influence of French on Quebec English: Motivation for lexical borrowing and integration of loanwords. *LACUS Forum* 26: 473–486.
Grieve, Jack, Montgomery, Chris, Nini, Andrea, Murakami, Akira & Guo, Diansheng. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2: 11. (doi:10.3389/frai.2019.00011)
Harris, Zellig S. 1954. Distributional structure. *Word* 10(2–3): 146–162. (doi:10.1080/00437956.1954.11659520)
Hengchen, Simon, Tahmasebi, Nina, Schlechtweg, Dominik & Dubossarsky, Haim. 2021. Challenges for computational lexical semantic change. In *Computational Approaches to Semantic Change*, Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds), 341–372. Berlin: Language Science Press.
Jones, Taylor. 2015. Toward a description of African American Vernacular English dialect regions using "Black Twitter." *American Speech* 90(4): 403–440. (doi:10.1215/00031283-3442117)
Josselin, Amélie. 2001. *L'emprunt lexical en France et au Canada : le cas particulier des anglicismes et des gallicismes et leur traitement lexicographique*. DEA thesis, Université de Lyon II.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Laicher, Severin, Kurtyigit, Sinan, Schlechtweg, Dominik, Kuhn, Jonas & Schulte im Walde, Sabine. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 192–202. Online: Association for Computational Linguistics.
Martinc, Matej, Montariol, Syrielle, Zosa, Elaine & Pivovarova, Lidia. 2020. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020* (WWW '20), 343–349. New York, NY, USA: Association for Computing Machinery. (doi:10.1145/3366424.3382186)
McArthur, Tom. 1989. *The English language as used in Quebec: A survey* [Strathy Occasional Papers on Canadian English 3]. Kingston, ON: Queen's University.
Miletic, Filip. 2019. Contact-induced lexical variation in Quebec English: An accountable description. In *RJC2019 - 22èmes Rencontres des jeunes chercheurs en Sciences du Langage*, Paris, France. <https://hal.archives-ouvertes.fr/hal-02280295/>
Miletic, Filip, Przewozny-Desriaux, Anne & Tanguy, Ludovic. 2020. Collecting tweets to investigate regional variation in Canadian English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6255–6264. Marseille, France: European Language Resources Association.
Miletic, Filip, Przewozny-Desriaux, Anne & Tanguy, Ludovic. 2021. Detecting contact-induced semantic shifts: What can embedding-based methods do in practice? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10852–10865. Punta Cana, Dominican Republic: Association for Computational Linguistics.
Montariol, Syrielle, Martinc, Matej & Pivovarova, Lidia. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4642–4652. Online: Association for Computational Linguistics. (doi:10.18653/v1/2021.naacl-main.369)
Nguyen, Dong. 2021. Dialect variation on social media. In *Similar Languages, Varieties, and Dialects. A Computational Perspective*, Marcos Zampieri & Preslav Nakov (eds.), 204–218. Cambridge: Cambridge University Press. (doi:10.1017/9781108565080.014)
Nguyen, Dat Quoc, Vu, Thanh & Tuan Nguyen, Anh. 2020. BERTweet: A pre-trained language model for English Tweets. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing: System Demonstrations, 9–14. Online: Association for Computational Linguistics. (doi:10.18653/v1/2020.emnlp-demos.2)
Owoputi, Olutobi, O'Connor, Brendan, Dyer, Chris, Gimpel, Kevin, Schneider, Nathan & Smith, Noah A. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 380–390. Atlanta, Georgia: Association for Computational Linguistics.
Pavalanathan, Umashanthi & Eisenstein, Jacob. 2015. Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2138–2148. Lisbon, Portugal: Association for Computational Linguistics. (doi:10.18653/v1/D15-1256)
Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier & Blondel, Mathieu et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.
Poplack, Shana, Walker, James A. & Malcolmson, Rebecca. 2006. An English 'like no other'?: Language contact and change in Quebec. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 51(2/3): 185–213. (doi:10.1017/S0008413100004060)
Rodda, Martina A., Lenci, Alessandro & Senaldi, Marco S. G. 2017. *Panta rei*: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics* 3(1): 11–24.
Rouaud, Julie. 2019. *Lexical and phonological integration of French loanwords into varieties of Canadian English since the seventeenth century*. PhD dissertation, Université Toulouse – Jean Jaurès.
Schlechtweg, Dominik, Hätty, Anna, Del Tredici, Marco & Schulte im Walde, Sabine. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 732–746. Florence, Italy: Association for Computational Linguistics. (doi:10.18653/v1/P19-1072)
Schlechtweg, Dominik, McGillivray, Barbara, Hengchen, Simon, Dubossarsky, Haim & Tahmasebi, Nina. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23. International Committee for Computational Linguistics.
Shoemark, Philippa, Sur, Debnil, Shrimpton, Luke, Murray, Iain & Goldwater, Sharon. 2017. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*,

1239–1248. Valencia, Spain: Association for Computational Linguistics. (doi:10.18653/v1/E17-1116)
Statistics Canada. 2022. Table 98-10-0218-01. Mother tongue by age: Canada, provinces and territories. <https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810021801>
Tagliamonte, Sali A. 2002. Comparative sociolinguistics. In *The Handbook of Language Variation and Change*, J. K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds), 729–763. Malden: Blackwell.
Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
Tahmasebi, Nina, Borin, Lars & Jatowt, Adam. 2021. Survey of computational approaches to lexical semantic change. In *Computational Approaches to Semantic Change*, Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds), 1–91. Berlin: Language Science Press.
Takamura, Hiroya, Nagata, Ryo & Kawasaki, Yoshifumi. 2017. Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1195–1204. Valencia, Spain: Association for Computational Linguistics. (doi:10.18653/v1/E17-1112)
Turney, Peter D. & Pantel, Patrick. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37: 141–188. (doi:10.1613/jair.2934)
Uban, Ana, Ciobanu, Alina Maria & Dinu, Liviu P. 2019. Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 161–166. Florence, Italy: Association for Computational Linguistics. (doi:10.18653/v1/W19-4720)
Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Delangue, Clement, Moi, Anthony & Cistac, Pierric et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.
Xu, Yang & Kemp, Charles. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2703–2708. Austin, Texas: Cognitive Science Society.