Detecting Contact-Induced Semantic Shifts: What Can Embedding-Based Methods Do in Practice?

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy

Detecting Contact-Induced Semantic Shifts: What Can Embedding-Based Methods Do in Practice?

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

Abstract

This study investigates the applicability of semantic change detection methods in descriptively oriented linguistic research. It specifically focuses on contact-induced semantic shifts in Quebec English. We contrast synchronic data from different regions in order to identify the meanings that are specific to Quebec and potentially related to language contact. Type-level embeddings are used to detect new semantic shifts, and token-level embeddings to isolate regionally specific occurrences. We introduce a new 80-item test set and conduct both quantitative and qualitative evaluations.

We demonstrate that diachronic word embedding methods can be applied to contact-induced semantic shifts observed in synchrony, obtaining results comparable to the state of the art on similar tasks in diachrony. However, we show that encouraging evaluation results do not translate to practical value in detecting new semantic shifts. Finally, our application of token-level embeddings accelerates manual data exploration and provides an efficient way of scaling up sociolinguistic analyses.

1 Introduction

A growing body of work has investigated semantic change detection, mainly relying on word embedding methods to model differences in meaning over time. This has been presented as a promising new approach for descriptive linguistics, providing the means to test theoretical hypotheses and discover new cases of semantic change on a large scale (Tahmasebi et al., 2021). However, it has also been pointed out that existing work has largely focused on generic research questions and datasets, using them as a training ground for proof-of-concept studies (Boleda, 2020). As a result, many reported instances of semantic change are either long established (e.g. *gay* shifting from 'happy' to 'homosexual') or are of otherwise limited importance (e.g. proper names reflecting sociohistorical context).

But how useful are semantic change detection methods when applied to a precisely defined descriptive issue? We use them to investigate contact-induced semantic shifts, focusing specifically on Quebec English. Semantic influence of French is documented in the sociolinguistic literature on Quebec (e.g. *deception* 'disappointment', cf. Fr. *déception*), but mostly from a qualitative standpoint, with the importance of the phenomenon still under debate. This reflects known methodological challenges in studying lexical variation, which is precisely what our approach aims to overcome.

We follow common sociolinguistic practice and use synchronic data, contrasting Quebec with predominantly English-speaking Canadian regions. Drawing on a large Twitter corpus, we implement a two-pronged approach: we use type-level embeddings to detect new semantic shifts, and token-level embeddings to isolate regionally specific occurrences of target lexical items. We conduct standard quantitative evaluation and extensive qualitative analyses, focusing on the descriptive relevance of the results and challenges arising from empirically occurring data.

We provide the following contributions:

- we demonstrate that diachronic word embedding methods can be applied to contact-induced semantic shifts observed in synchrony. We validate these approaches using a new 80-item test set, obtaining results comparable to the state of the art on similar semantic change detection tasks in diachrony;
- we show that encouraging evaluation results do not translate to practical value in detecting new semantic shifts, mainly due to the rarity of the phenomenon and noise in the results;
- we illustrate the utility of token-level embeddings in identifying regionally specific occurrences. Our approach accelerates manual data exploration and provides an efficient way of scaling up sociolinguistic analyses.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10852–10865
November 7–11, 2021.

This paper is organized as follows. We first provide a brief overview of related work (§ 2) and introduce the corpus and the test set (§ 3). We then present two experiments, the first using type-level embeddings to detect new cases of semantic change (§ 4), and the second using token-level embeddings to investigate the individual senses of target lexical items (§ 5). We finally outline key conclusions and directions for future work (§ 6).

2 Related work

Semantic change detection methods developed in recent years have often relied on vector space models (Kutuzov et al., 2018), where meaning is represented as a vector reflecting the target word's cooccurrence patterns. The general approach consists in training models specific to different time periods and then quantifying the distance between a word's vectors in different models. The same principle can be applied to semantic differences across regions, text types, and so on.

Specific implementations include creating separate count-based models while maintaining the same linguistic contexts across time periods (Gulordava and Baroni, 2011) or training neural word embeddings in a way that that captures meaning change while ensuring comparability between vector spaces (Dubossarsky et al., 2019; Hamilton et al., 2016b; Kim et al., 2014). Semantic change is usually quantified using the cosine distance between a word's vectors in different models, but alternative measures have also been proposed (Hamilton et al., 2016a). More recently, pretrained neural networks have been used to obtain contextual embeddings and detect change by computing average representations for different time periods or measuring differences in clustering patterns (Giulianelli et al., 2020; Martinc et al., 2020a,b). Similar work has been done on synchronic data, but it is considerably more limited. It has focused on issues such as semantic variation across online communities of practice (Del Tredici and Fernández, 2017), text types (Fišer and Ljubešić, 2018), large dialect regions (Kulkarni et al., 2016), and different languages (Uban et al., 2019).

Evaluation of these methods is challenging and is often limited to a qualitative analysis of a restricted number of examples. Quantitative evaluations have involved background information from lexicographic sources (Basile and McGillivray, 2018) or synthetic corpora (Shoemark et al., 2019). While their focus is on the top semantic change candidates output by a model, an alternative approach consists in using manually annotated datasets containing (usually several dozen) words whose meaning is either stable or subject to change (Basile et al., 2020; Del Tredici et al., 2019; Gulordava and Baroni, 2011; Schlechtweg et al., 2019, 2020). The words are associated with binary labels or semantic change scores, and the models are evaluated on a ranking or binary classification task.

Descriptively oriented studies have used these methods to test theoretical hypotheses, as in the evaluation of two competing laws of semantic change by Xu and Kemp (2015), or to facilitate exploratory analyses by domain experts. For instance, Rodda et al. (2017) used vector space models for Ancient Greek to detect semantic shift candidates and then analyzed their nearest neighbors to establish meaning change. De Pascale (2019) studied regional lexical variation in Dutch using token-level models to identify cases where competing words were used with equivalent meanings.

Computational work on sociolinguistic issues uses social media posts or other large corpora to detect lexical innovations and dialect regions (Donoso and Sánchez, 2017; Eisenstein, 2018; Grieve et al., 2018) or to examine the impact of sociodemographic factors (Bamman et al., 2014; Blodgett et al., 2016; Nguyen et al., 2013). Language contact has been addressed in terms of codeswitching (Bullock et al., 2019; Guzman et al., 2016), language choice (Eleta and Golbeck, 2014; Nguyen et al., 2015), and semantic integration of loanwords (Deng et al., 2019; Serigos, 2017; Takamura et al., 2017). Related work on Canadian English is limited to corpus construction (Cook and Brinton, 2017; Miletic et al., 2020) and descriptive analyses of online communication (Tagliamonte and Denis, 2008). We are unaware of any computational studies of regional variation in Canadian English, or of contact-induced semantic shifts in inherited (i.e. non-borrowed) lexical items.

3 Data for contact-induced semantic shifts in Quebec English

The presence of contact-induced semantic shifts in Quebec English has been reported in a series of sociolinguistic surveys (Boberg, 2012; McArthur, 1989; Rouaud, 2019) and manual corpus analyses (Fee, 1991, 2008; Grant, 2010; Josselin, 2001). They have described around 60 English words at-

tested in Quebec with meanings typical of formally or semantically similar French words. This phenomenon involves the introduction of a new sense into a word's polysemic structure through innovative meaning change (Koch, 2016) driven by semantic interference in bilinguals (cf. Romaine, 1995).

While this process can be observed over time, we adopt a synchronic perspective. Our assumption is that contact-induced semantic shifts correspond to words whose sense distributions are different in Quebec, a majority French-speaking province, compared to predominantly English-speaking Canadian regions. This is a typical comparative sociolinguistic approach (Tagliamonte, 2002): we contrast speech communities to identify linguistic behaviors which mirror the differences in their composition.

In this section, we introduce the data needed to implement our approach: a large corpus of tweets allowing us to analyze different Canadian regions, and a new test set used to tune our models and validate their performance.

3.1 Corpus

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 word embeddings, and it contains information on the regional origin and authors of individual tweets. This structure allows us to identify usage patterns that differentiate Montreal from other parts of Canada. It also allows us to distinguish contact phenomena from unrelated types of variation, such as regional trends that are not driven by 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 by identifying users from the three cities and then crawling their Twitter profiles. The tweets were filtered for geographic origin and language, and near-duplicates posted by individual users were automatically removed. In addition to these 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 the most 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 final corpus is presented in Table 1.

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

Table 1: Corpus structure

3.2 Test set

We introduce a new test set allowing for a systematic evaluation of semantic shift detection in the context of English–French language contact. Similarly to recent shared tasks on diachronic data (Basile et al., 2020; Schlechtweg et al., 2020), we formulate the task as a binary classification problem. The items in the test set are accordingly labeled as semantic shifts or stable words. We include 80 items; this is comparable to the diachronic test sets of which we are aware, containing between 18 and 100 items.

We first identify a set of positive examples, starting from a list of candidates based on the literature on Quebec English (Boberg, 2012; Fee, 1991, 2008; Grant, 2010; Josselin, 2001; McArthur, 1989; Rouaud, 2019). Any items with fewer than 100 occurrences per subcorpus are excluded. A concordance-based analysis is then used to determine if the items present at least one contact-related occurrence in the Montreal subcorpus; those that do not are also excluded. This leaves us with a list of 40 semantic shifts, whose mean frequency in the entire corpus is 5,268 (min = 345, max = 97,188).

We then select 40 stable words, aiming to limit the probability of using items with formally similar French equivalents, as they are more likely to be involved in contact-induced semantic change. We therefore start from a list of 3,231 English words

of Anglo-Saxon origin,[1] around half of which meet the frequency threshold (100 per subcorpus). For each of the 40 semantic shifts, we identify a word in the list which is of the same part of speech and is the closest to it in terms of frequency measured on the whole corpus. Using a sample of occurrences, we then verify that the words are not affected by meaning variation across subcorpora or other issues which could bias subsequent analyses (homography, use in proper names etc.). If necessary, we replace them with the word which is the next closest in frequency. Sample items are presented in Table 2; the entire test set is publicly available.[2]

Note that most recent diachronic test sets were created through costly annotation campaigns, whereas we rely on expert judgment, similarly to some existing work (e.g. McGillivray et al., 2019; Perrone et al., 2019). This is a viable approach because the phenomenon under study is more specific than general semantic change over time, and its existence can be reliably established based on the sociolinguistic literature, lexicographic sources, and observation of corpus data. In addition, our test set only uses binary labels, so our decisions are comparatively straightforward.

4 Detecting semantic shift candidates

We first examine contact-induced semantic shifts from an exploratory perspective. Our aim is to determine which words, within the entire vocabulary, have the most different overall meanings in Montreal compared to Toronto and Vancouver. This section introduces the experimental setup, presents the results of an evaluation used to find the best performing model, and examines the results it obtains on the detection of semantic shift candidates.

4.1 Experimental setup

We implement an approach based on type-level word embeddings, which multiple shared tasks on diachronic data have shown to be robust and to strongly outperform more recent token-level approaches (Basile et al., 2020; Schlechtweg et al., 2020). Given the specifics of our work — synchronic variation rather than diachronic change, cross-linguistic influence — we experiment with several parameters which may influence model performance depending on datasets and tasks.

[1] https://en.wikipedia.org/wiki/List_of_English_words_of_Anglo-Saxon_origin

[2] http://redac.univ-tlse2.fr/corpora/canen.html

Sem. shift	Fr. meaning	Freq.	Stable word
formidable	'terrific'	1.48	damp
circulation	'traffic'	2.12	campfire
deceive	'disappoint'	2.98	cram
souvenir	'memory'	3.11	hassle
resume	'summarize'	4.91	arise

Table 2: Sample semantic shifts, with frequency per million words and corresponding stable words in the test set (same POS and closest in frequency)

Meaning representations. We use `word2vec` (Mikolov et al., 2013) to train vector space models using the skip-gram architecture with negative sampling (SGNS). We experiment with different vector dimensions (100, 300) and context window sizes (2, 5, 10), and we set the minimum word frequency to 100 in each subcorpus. We use the default values for other parameters, and set the negative sampling rate to 5, the subsampling rate to 10^{-3}, and the number of iterations to 5. In order to control for word embedding instability (Pierrejean and Tanguy, 2018), we train three models for each configuration.

Comparing regional subcorpora. We experiment with two approaches used to compare meaning representations from different subcorpora.

In the first approach, we train individual word embedding models for each regional subcorpus. Since the resulting models do not share the same vector space, they need to be aligned so that they can be directly compared. We do so using Orthogonal Procrustes (OP) in the implementation first introduced by Hamilton et al. (2016b), which corresponds to finding the optimal rotational alignment for each pair of matrices. As suggested by Schlechtweg et al. (2019), we mean-center the matrix columns before alignment.

In the second approach, we train a single word embedding model using the entire corpus. Target words are tagged so as to be specific to the subcorpus in which they appear, while context words are the same across the subcorpora. As a result, the meaning representations are specific to each region, but they share the same vector space and are directly comparable. This corresponds to the Temporal Referencing method, which was introduced in diachronic studies with the aim of limiting noise in model alignment (Dubossarsky et al., 2019). For clarity, we will refer to this method as Spatial Referencing (SR) given the focus of our work.

The OP models are trained using `gensim` (Řehůřek and Sojka, 2010), and the SR models using `word2vecf` (Levy and Goldberg, 2014).

Measuring differences in meaning. Following common practice, we use cosine distance (CD) to measure semantic differences. We compute the CD for a word w in subcorpora a and b as follows:

$$CD(w_a, w_b) = \frac{\sum_{i=1}^{n} CD(\vec{w}_{a_i}, \vec{w}_{b_i})}{n}$$

for $n = 3$ runs of the SGNS model, where $\vec{w}_{a_i}$ is the word's vector corresponding to the subcorpus a in the i^{th} run. We then compute the variation score:

$$var(w) = \frac{CD(w_m, w_t) + CD(w_m, w_v)}{2} - CD(w_t, w_v)$$

where w_m, w_t, and w_v are the word's representations in the Montreal, Toronto, and Vancouver subcorpora. The score prioritizes the words whose meaning in Montreal differs from Toronto and Vancouver, but remains similar in those two cities.

4.2 Finding the best performing model

We begin by evaluating the overall performance of model configurations on the previously introduced test set. This allows us to tune the models as well as to validate their performance relative to the results reported on other similar tasks.

Since we focus on the general patterns captured by the models, we use a simple classification based on the median variation score: we compute the score for the 80 words in the test set and consider that the 40 words with the highest score represent semantic shifts, whereas the others are stable.[3] The best performing configuration (Orthogonal Procrustes, 100-dimensional vectors, window size of 5) obtains an accuracy score of 0.8. This is an improvement of 0.175 points compared to the worst result, which confirms the interest of dataset-specific model tuning on this task. Key results are outlined in Table 3; accuracy for all tested models is provided in Appendix A.

Comprehensive evaluation is beyond the scope of this paper, but we briefly overview the main takeaways. (i) The 100-dimension models strongly outperform the 300-dimension models, in line with Pražák et al.'s (2020) results in diachrony.

[3]This evaluation method reflects the split of positive and negative items in the test set. Although this may introduce a bias, our approach allows for a simple and efficient comparison of different sets of hyperparameters before subsequent qualitative analyses of the best performing model.

Parameters		Accuracy		
		mean	**min**	**max**
Dim	100	**0.738**	**0.700**	**0.800**
	300	0.675	0.625	0.750
Win	2	**0.713**	**0.675**	0.750
	5	**0.713**	0.650	**0.800**
	10	0.694	0.625	0.775
Type	OP	0.700	**0.650**	**0.800**
	SR	**0.713**	0.625	0.775

Table 3: Accuracy across model configurations using different parameters. Dim: vector dimensions; win: window size; OP: Orthogonal Procrustes; SR: Spatial Referencing.

(ii) Smaller window sizes perform somewhat better, in line with synchronic work on German (Schlechtweg et al., 2019). (iii) The alignments are roughly comparable, with OP obtaining the best individual result and SR the higher mean score.

The highest accuracy score we obtain is comparable to the state of the art on other semantic change detection tasks. In addition to indicating the best individual configuration, this validates the general setup of our experiment, confirming that a clear regional regional distinction is present in our data in relation to the semantic influence of French.

4.3 Deploying the model

We now turn to the discovery of semantic shift candidates from the whole vocabulary using the best performing model configuration. Since our aim is to examine its performance on empirically occurring data for potentially ongoing language change, we manually analyze the contexts in which the candidates appear rather than using external information or synthetic corpora.

We calculate the variation score for the whole vocabulary (open classes only, i.e. nouns, verbs, adjectives, and adverbs), and select the 50 words with the highest score; they are presented in Appendix B. We verify for each word (i) whether it presents a regionally specific use in the Montreal subcorpus, and (ii) whether this use can be explained by the influence of French, and specifically the presence of an equivalent sense in a formally or semantically similar French word, as established by lexicographic sources.[4] This allows us to compute

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

precision; we do not compute recall as it would require a manual analysis of the entire vocabulary.

We find that only one candidate clearly corresponds to a contact-induced semantic shift, which translates to a precision score of 0.02. The positive example is the noun *exposition* ($var = 0.22$, ranked 26th); it usually refers to an art exhibition in the Montreal subcorpus, and to narrative structure in the two other subcorpora. The meanings are respectively illustrated by examples 1 and 2:

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

2. I found the first 2 episodes a little slow, but it does pick up once the **exposition** is done with

The contact-related sense is typical of the French homograph *exposition* 'exhibition'. This use has been previously described (Fee, 1991), and it is included in our test set.

The contrast between the accuracy on the test set and the precision on the discovery task is striking. Figure 1 shows that the problem lies in the fact that the lexical items of interest, like those in the test set, are not the ones with the most extreme variation scores. They tend to have higher variation scores than stable words, showing that the model does capture meaningful trends. But relevant results are ultimately obscured by other types of variation.

False positives include proper names denoting local referents (*plateau* referring to Plateau-Mont-Royal, a Montreal neighborhood); topical variation (*detached* limited to real estate in Toronto and Vancouver, which have notoriously tight housing markets); French homographs in code-switched tweets (*pour* 'for'); misspellings indicative of an imperfect command of English (*trough* 'through'). While some of these issues have been reported in previous diachronic studies (e.g. referential effects in Del Tredici et al., 2019), our results underscore that they are highly widespread even when model configurations are carefully tuned. It is tempting to say that they could easily be avoided using basic data filtering, such as the exclusion of the words attested in French corpora or the use of additional frequency thresholds. But things are more complicated than that: for instance, homography also affects many longstanding borrowings (*bureau*) and targeted semantic shifts (*exposition*); higher relative frequency in Montreal may reflect noise as well as increased use of a word that has undergone semantic change. It is also not viable to keep extending the list of top

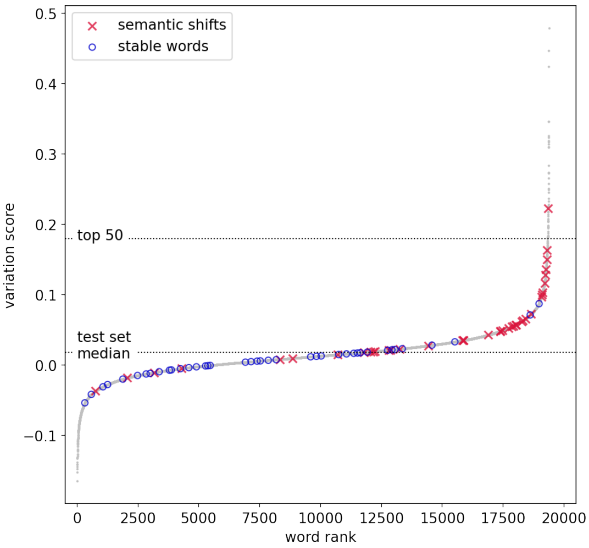

Figure 1: Variation scores for the whole vocabulary, with the position of semantic shifts and stable words from the test set. Horizontal lines indicate the cutoff score for the top 50 candidates and the test set median.

candidates: there are on average 78 words in the list between each of the top 10 positive examples from the test set; this increases to 476 if all 40 positive examples are considered.

5 Characterizing semantic shifts in context

A key limitation of the approach we introduced in the previous section is that it is sense-agnostic: all contexts in which a word appears are grouped into a single vector. Although it can capture meaningful general trends, it is limited in accounting for rare phenomena, which may be troublesome when analyzing language change. This process is not abrupt; rather, it involves a period of coexistence of a new linguistic behavior with the old one (Weinreich et al., 1968), which for semantic change implies gradual modifications to the polysemic structure of the word (Traugott, 2017). We therefore expect contact-related senses to appear in a limited number of occurrences per word. We further expect them to be attested in similar immediate contexts, but to be dispersed throughout the corpus.

These assumptions, coupled with the previously outlined issues with type-level analyses, suggest that manual inspection of corpus data remains necessary to reliably identify semantic shifts. In order to facilitate this process, we implement a system which allows us to automatically group semantically similar occurrences of a target word and identify the uses that are specific to Montreal. This

(1)	portraits in honour of Janet Werner's upcoming	**exposition**	at the museum . Starting November 10th , every
	Collection " Come to admire Laura Granata's	**exposition**	at #CLDV
	Such a beautiful	**exposition**	!!! #mbam #art #montrealmuseum
(2)	space will be turned into a citizens' area with	**exposition**	space and multipurpose room . #CJAD #polmtl
	is WINDSOR STATION - It's now part of the	**exposition**	events space in Montreal . Its located next to the
	media students are showcasing their work at an	**exposition**	hall in Trois-Rivière .
(3)		**exposition**	d'aquarelles , exhibition of my watercolor works
		Exposition	du World Press Photo 2016 #photo #feedly
		Exposition	en cours - Galerie d'art Stewart Hall

Table 4: Sample clusters for *exposition*

enables us to manually analyze batches of occurrences all at once, rather than examining them one at a time, which streamlines both the discovery of contact-related uses and the exclusion of false positives. In the remainder of this section, we present the experimental setup, a sample analysis of clustered data, and the overall patterns for 40 semantic shifts that were manually annotated using this approach.

5.1 Experimental setup

We produce contextual meaning representations, reflecting differences in the use of individual occurrences. We then group them into clusters in order to identify regionally specific patterns.

Token-level embeddings. We use BERT (Devlin et al., 2019), a Transformer-based language model, to produce token-level embeddings. We use the HuggingFace implementation (Wolf et al., 2020) of `bert-base-uncased`, a 12-layer, 768-dimension version of the model pretrained on English data. No fine-tuning is performed given its computational cost and the assumption that word senses are reflected by differences in context which the pretrained model should be able to capture.

For each analyzed word, we extract the tweets in which it appears in all three subcorpora. In order to limit processing and memory requirements, we retain no more than 1,000 total occurrences per word, and use a random sample for more frequent items. We feed each tweet as a single sequence into BERT, which then produces context-informed embeddings for each token in the tweet. We extract the representations for the target word and average over the last 4 hidden layers to obtain a single contextual embedding, similarly to other recent studies (e.g. Laicher et al., 2021). BERT's tokenizer splits out-of-vocabulary words into subparts with known representations; when this occurs, we average over the subparts to produce a single embedding.

Clustering. We identify a word's similar uses by clustering its contextual embeddings using affinity propagation, which performed well in other semantic change studies (e.g. Martinc et al., 2020b). We use the `scikit-learn` (Pedregosa et al., 2011) implementation with default parameters.

Analyzing regional use. We consider the clusters containing at least five tweets, and retain them if more than half of the tweets were published in Montreal. This is because we are interested in the senses which are clearly more frequent in Montreal than elsewhere, but which may occasionally appear in other regions. We retain up to 10 such clusters per item, starting with those with the highest proportion of Montreal tweets. We then manually annotate the data for the 40 semantic shifts included in the test set. We use binary labels, and establish if a cluster presents a contact-related sense based on the majority usage in it.

More specifically, a target word's use is annotated as contact-related if it fulfills the criterion we previously defined: it must be regionally specific to Montreal and potentially explained by the influence of French. The clusters affected by the issues observed in type-level analyses — referential effects, topical variation, French homographs — are not considered as contact-related. (See Section 4.3 for more details on both the positive and negative criteria.) Neither do we annotate as contact-related the clusters involving structural patterns (e.g. the target word is used with different senses but systematically appears in the tweet-initial or tweet-final position) or those where no reliable determination can be made (e.g. short or ambiguous tweets).

5.2 Exploring clusters of tweets

On average, 8 clusters per word (min = 3, max = 10) were retained for annotation. The mean number of tweets per cluster for a given word stands on average at 13 (min = 8, max = 20).

The sample clusters for *exposition* shown in Table 4 illustrate several types of usage that are frequently grouped together. Cluster 1 contains straightforward examples of contact-induced semantic shifts, which in this case refer to art exhibitions. This is the sense previously illustrated by example 1 in Section 4.3. Cluster 2 shows the effect of this usage on a specific collocational pattern (*exposition space / hall*), indicating further diffusion of the semantic shift. Cluster 3 reflects noise in the results: *exposition* is attested as its French homograph in code-switched tweets (in which most text is in English, explaining why they were tagged as English and retained during corpus creation). Additional sample clusters are provided in Appendix C.

As these examples illustrate, the clusters are largely homogeneous. Although some are occasionally difficult to interpret, e.g. due to the influence of orthographic information on BERT's representations, this is overall rare. A 15-word sample was annotated by two annotators in order to test the reliability of the general procedure, obtaining a reasonably high Cohen's kappa coefficient of 0.55.

The utility of this approach is confirmed by the fact that it enabled us to identify 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.

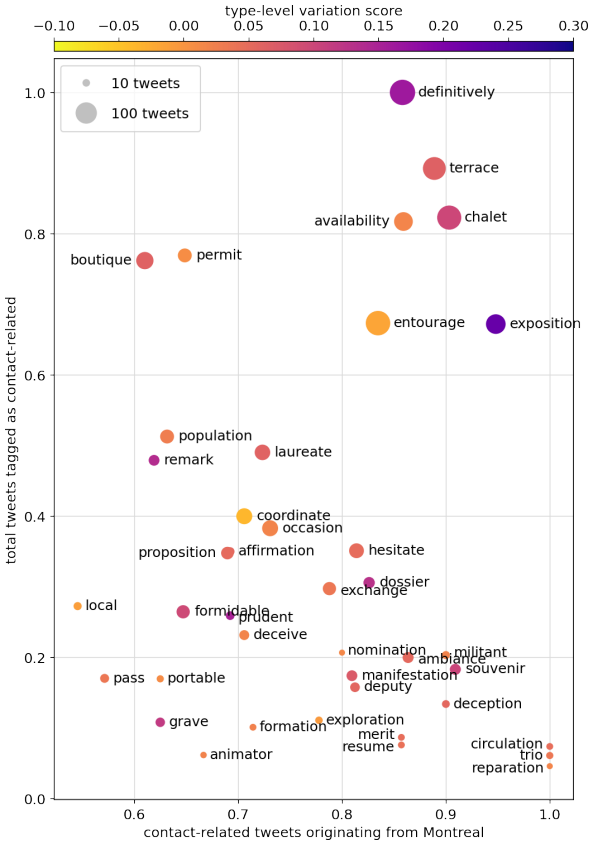

Figure 2: Scatter plot of annotated words. Y-axis: proportion of tweets that were tagged as contact-related (out of all annotated tweets). X-axis: proportion of tweets from the Montreal subcorpus (out of all tweets tagged as contact-related). Marker size reflects the total number of tweets that were tagged as contact-related. Color coding indicates the variation score computed on type-level models.

5.3 Patterns of semantic change

We analyze general trends in the annotated data in order to determine if they are related to the variation scores established using type-level models. We specifically focus on two issues that may limit the performance of type-level models: (i) if a contact-related use concerns a minority of occurrences, it is unlikely to be captured by `word2vec`; (ii) if it is frequent but not regionally specific, if will not be reflected by our variation score. We compute two corresponding measures: (i) the proportion of tweets, out of all annotated tweets, which appear in clusters that are tagged as contact-related; and (ii) the proportion of tweets, out of all tweets in the clusters that are tagged as contact-related, which originate from the Montreal subcorpus.

The results plotted in Figure 2 point to two overarching tendencies. On the one hand, several semantic shifts (*definitively* 'definitely', *exposition* 'exhibition' etc.) are characterized by a high proportion of contact-related tweets and, among those, a high proportion of tweets from Montreal. This is indicative of overwhelming contact-related influence which is moreover regionally specific. On the other hand, a larger number of examples (*circulation* 'traffic', *animator* 'group leader' etc.) present limited contact-related usage which additionally varies in terms of regional specificity.

The annotation-based measures are weakly correlated with the type-level variation score ($\rho = 0.23$ for both measures), as well as with one another ($\rho = -0.13$). This reflects contrasting patterns, like in the case of *entourage*, which has a large number of contact-related tweets and a low type-level variation score (-0.02, ranked 39th out of the 40 items). This is related to the relatively small difference between the conventional sense

'people attending an important person' and the contact-induced sense 'group of friends, family'. The distinction is immediately apparent to the annotator, but it is often underpinned by referential knowledge rather than differences in distributional contexts. Compare this with the adjective *grave*, which has a high type-level variation score (0.12, ranked 6th), but appears in few contact-related clusters. This is due to most of its clusters being excluded because they involve its French homograph, as in the expression *ce n'est pas grave* 'it doesn't matter'. However, the use of French implies drastic distributional differences which are easily captured by type-level models.

6 Conclusion

In this study, we applied embedding-based semantic change detection methods to a new set of experimental conditions. Focusing on the descriptive relevance of the results, we developed an 80-item test set for English–French language contact, and manually annotated corpus data for 40 lexical items. Our results demonstrate that current type-level methods are of limited practical use in detecting new semantic shifts: despite robust performance on standard quantitative evaluation, they obtain very low precision on the discovery task. We further implemented a token-level analysis of the individual senses with which target items are used. This accelerated manual data exploration, leading to a detailed account of the issues impacting type-level methods. These results more generally represent the first quantitative corpus-based analysis of contact-induced semantic shifts in Quebec English.

Our work also has more general implications, which reaffirm the central role of evaluation practices and corpora in advancing computational analyses of semantic change (Hengchen et al., 2021). The comparison of evaluations on the test set and on the discovery task underscored the stark difference between the two approaches. This should be taken into account when choosing evaluation methods, especially where the aim is to establish practical usability. And while some reported issues are specific to our corpus, similar problems may affect other semantic change studies, as noisy datasets and complex sense distributions are not unique to our work. Finally, the comparison of type-level and token-level analyses highlighted diverging trends in the data which indicate that semantic shifts involve multiple dimensions of variation. Future work should therefore aim to identify different types of semantic change in addition to quantifying its presence.

Acknowledgements

We are grateful to the anonymous reviewers for their valuable comments and suggestions. Experiments presented in this paper were carried out using the OSIRIM computing platform, administered by the IRIT research lab and supported by the CNRS, the Région Occitanie, the French Government and the ERDF (see https://osirim.irit.fr).

References

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

Katherine Barber, editor. 2004. *Canadian Oxford Dictionary*. Oxford University Press, Don Mills.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 diachronic lexical semantics (DIACR-Ita) task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*.

Pierpaolo Basile and Barbara McGillivray. 2018. Exploiting the web for semantic change detection. In Larisa Soldatova, Joaquin Vanschoren, George Papadopoulos, and Michelangelo Ceci, editors, *Discovery Science*, volume 11198, pages 194–208. Springer International Publishing, Cham.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

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

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

Barbara Bullock, Wally Guzmán, and Almeida Jacqueline Toribio. 2019. The limits of Spanglish? In *Proceedings of the 3rd Joint SIGHUM Workshop*

on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 115–121, Minneapolis, USA. Association for Computational Linguistics.

Hélène Cajolet-Laganière, Pierre Martel, Chantal-Édith Masson, and Louis Mercier. 2014. Usito. https://usito.usherbrooke.ca/.

Paul Cook and Laurel J Brinton. 2017. Building and evaluating web corpora representing national varieties of English. *Language Resources and Evaluation*, 51(3):643–662.

Stefano De Pascale. 2019. *Token-Based Vector Space Models as Semantic Control in Lexical Lectometry*. Doctoral dissertation, KU Leuven, Leuven.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long Papers*.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. 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)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacques Dendien and Jean-Marie Pierrel. 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.

Zhubo Deng, Weijia Shi, Pei Zhou, Muhao Chen, and Kai-Wei Chang. 2019. Computational analysis of French reborrowing process for English loanwords. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 1–4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Dollinger and Margery Fee. 2017. DCHP-2: The dictionary of Canadianisms on historical principles, second edition. http://www.dchp.ca/dchp2.

Gonzalo Donoso and David Sánchez. 2017. Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 16–25, Valencia, Spain. Association for Computational Linguistics.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-Out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Jacob Eisenstein. 2018. Identifying regional dialects in on-line social media. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, Blackwell Handbooks in Linguistics, pages 368–383. Wiley Blackwell, Hoboken, NJ.

Irene Eleta and Jennifer Golbeck. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432.

Margery Fee. 1991. Frenglish in Quebec English newspapers. In *Papers of the Fifteenth Annual Meeting of the Atlantic Provinces Linguistic Association*, pages 12–23. Atlantic Provinces Linguistic Association.

Margery Fee. 2008. French borrowing in Quebec English. *Anglistik: International Journal of English Studies*, 19(2):173–188.

Darja Fišer and Nikola Ljubešić. 2018. Distributional modelling for semantic shift detection. *International Journal of Lexicography*, 32(2):163–183.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 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*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Pamela Grant. 2010. English usage in contemporary Quebec: Reflections of the local. In Elaine Gold and Janice McAlpine, editors, *Canadian English: A Linguistic Reader*, number 6 in Strathy Occasional Papers on Canadian English, pages 177–197. Queen's University, Kingston, ON.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics*, 46(4):293–319.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*,

pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 341–372. Language Science Press, Berlin.

Amélie Josselin. 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, Lyon.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Peter Koch. 2016. Meaning change and semantic shifts. In Päivi Juvonen and Maria Koptjevskaja-Tamm, editors, *The Lexical Typology of Semantic Shifts*, number 58 in Cognitive Linguistics Research, pages 21–66. De Gruyter Mouton, Berlin, New York.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? Quantifying the geographic variation of language in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):615–618.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 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*, pages 192–202, Online. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020a. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 343–349, New York, NY, USA. Association for Computing Machinery.

Tom McArthur. 1989. *The English Language as Used in Quebec: A Survey*. Number 3 in Strathy Occasional Papers on Canadian English. Queen's University, Kingston, ON.

Barbara McGillivray, Simon Hengchen, Viivi Lähteenoja, Marco Palma, and Alessandro Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4):893–907.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Filip Miletic, Anne Przewozny-Desriaux, and Ludovic Tanguy. 2020. Collecting tweets to investigate regional variation in Canadian English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6255–6264, Marseille, France. European Language Resources Association.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?": A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 439–448.

Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 666–669.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 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*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.

Benedicte Pierrejean and Ludovic Tanguy. 2018. Predicting word embeddings variability. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 154–159, New Orleans, Louisiana. Association for Computational Linguistics.

Ondřej Pražák, Pavel Přibáň, Stephen Taylor, and Jakub Sido. 2020. UWB at SemEval-2020 task 1: Lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online). International Committee for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Martina A Rodda, Alessandro Lenci, and Marco S G Senaldi. 2017. *Panta Rei*: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3(1):11–24.

Suzanne Romaine. 1995. *Bilingualism*. Wiley Blackwell, Oxford.

Julie Rouaud. 2019. *Lexical and Phonological Integration of French Loanwords into Varieties of Canadian English since the Seventeenth Century*. Doctoral dissertation, Université Toulouse - Jean Jaurès, Toulouse.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 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*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23. International Committee for Computational Linguistics.

Jacqueline Serigos. 2017. Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of Anglicisms in Spanish. *International Journal of Bilingualism*, 21(5):521–540.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

S. A. Tagliamonte and D. Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83(1):3–34.

Sali A. Tagliamonte. 2002. Comparative sociolinguistics. In J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, editors, *The Handbook of Language Variation and Change*, pages 729–763. Blackwell, Malden and Oxford.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press, Berlin.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 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*, pages 1195–1204, Valencia, Spain. Association for Computational Linguistics.

Elizabeth Closs Traugott. 2017. Semantic change. *Oxford Research Encyclopedia of Linguistics*.

Ana Uban, Alina Maria Ciobanu, and Liviu P. Dinu. 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*, pages 161–166, Florence, Italy. Association for Computational Linguistics.

Uriel Weinreich, William Labov, and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann and Yakov Malkiel, editors, *Directions for Historical Linguistics*, pages 95–188. University of Texas Press, Austin.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2703–2708, Austin, Texas. Cognitive Science Society.

A Evaluation of type-level models

Accuracy for all evaluated type-level models is presented in Table 5.

Type	Win	Dim	Accuracy
OP	5	100	**0.800**
SR	10	100	0.775
SR	2	300	0.750
SR	2	100	0.725
SR	5	100	0.725
OP	2	100	0.700
OP	10	100	0.700
SR	5	300	0.675
OP	10	300	0.675
OP	2	300	0.675
OP	5	300	0.650
SR	10	300	0.625

Table 5: Evaluation results for all model configurations. OP: Orthogonal Procrustes; SR: Spatial Referencing; Win: window size; Dim: vector dimensions.

B Type-level semantic shift candidates

The full list of the top 50 semantic shift candidates output by the best performing type-level model is presented in Table 6.

1	pour	26	exposition
2	plateau	27	s
3	nt	28	corona
4	den	29	encore
5	rapport	30	trustee
6	sous	31	coupe
7	ont	32	2
8	en	33	dispatch
9	mb	34	dire
10	aux	35	appraisal
11	saison	36	vie
12	tout	37	premier
13	svp	38	overdose
14	vers	39	petite
15	pour	40	fort
16	bec	41	mtg
17	de	42	plus
18	pa	43	vu
19	trough	44	nest
20	gorge	45	staging
21	detached	46	basin
22	le	47	br
23	parfait	48	ce
24	still	49	lever
25	#venom	50	bologna

Table 6: Top 50 semantic shift candidates

C Sample clusters based on token-level representations

Additional examples of cluster-based analyses are presented in Tables 7 (*manifestation*), 8 (*deception*), and 9 (*definitively*). Detailed explanations of the observed patterns are provided alongside each set of examples.

	There was a	**manifestation**	in Montreal against the proposed religious
(1)	Montreal's	**manifestation**	protesting against loi 21 banning of « religious
	in Quebec's history . This walk is the biggest	**manifestation**	for this week . And 52 more towns in the province
	This is the most visual	**manifestation**	of patriarchal privilege . That's why it's especially
(2)	Probably the best	**manifestation**	of the benefits of physical/digital retail integration
	about the the fact that Disneyland is the physical	**manifestation**	of 1950s American exceptionalism and right-wing
	Giving a Voice to the Voiceless — attending	**Manifestation**	Contre Projet De Loi 128 , Protest Against Bill
(3)	to the streets this afternoon in Montréal Grande	**Manifestation**	contre la haine et le racisme . Demonstrators
	currently having a brownout the night before the	**manifestation**	pour le climat here in Montreal . How odd it is .

Table 7: Sample clusters for *manifestation*, which is typically used in English to signify 'instance, display'. Cluster 1 illustrates the contact-related sense 'protest, demonstration' (cf. Fr. *manifestation*). Cluster 2 corresponds to the conventional English sense. Cluster 3 contains occurrences of the French homograph *manifestation* attested in codeswitched tweets.

	#ducks ?? With big expectations come biggest	**deceptions**	... #nhlhockey #Game7Curse #game7
(1)	Great expectations , few	**deceptions**	and stunning debuts make a unique
	fantastic example of endurance and overcoming	**deceptions**	! Thank you so much !
	The Coffee	**Deception**	: 13 Little Known Facts About Coffee
(2)	The grand	**deception**	: Looking for love , validation & peace outside of
	Kavanaugh's testimony : The immaculate	**deception**	.
	The new song	**Deception**	Bay , from Milk & Bone's second album , is out !
(3)	From their second album , out now :	**Deception**	Bay . I wonder if they are familiar with the doggy
		Deception	Bay is a masterpiece

Table 8: Sample clusters for *deception*, which in English refers to the action of deceiving (misleading) someone. Cluster 1 reflects the contact-related sense 'disappointment' (cf. Fr. *déception*). Cluster 2 is a case where no determination was made by the annotators as the contexts were deemed insufficiently specific to disambiguate the possible senses. Cluster 3 exemplifies the use of the target word as a proper noun, here referring to the song "Deception Bay" by the Montreal band Milk & Bone.

	Pouring coffee beans in the water tank ... I	**definitively**	need coffee !!!
(1)	again some developers after all these years ! I	**definitively**	want to come back to Montréal next year for the
	thank you very much ♡ I'm touched , I would	**definitively**	love to work with you one day !
	This is	**definitively**	a job that should've been replaced by a small script
(2)	65% of everything in school is	**definitively**	a waste of time . Useless subjects and more > : l
	party that would bring us decades back . A party	**definitively**	far from the interests of Quebeckers and
	In 2018 ? Most	**definitively**	!
(3)		**Definitively**	!
		Definitively	Yay !!!

Table 9: Sample clusters for *definitively*, whose conventional meaning in English is 'conclusively, indisputably'. Clusters 1 and 2 indicate different contexts in which it is used with the more general contact-related sense 'definitely, certainly' (cf. Quebec French *définitivement* 'definitely'). Cluster 3 shows a further generalization of that use, including as an emphatic interjection ('yes!').

.