Collecting Tweets to Investigate Regional Variation in Canadian English

Filip Miletic, Anne Przewozny-Desriaux, Ludovic Tanguy

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6255–6264
Marseille, 11–16 May 2020

Collecting Tweets to Investigate Regional Variation in Canadian English

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

Abstract

We present a 78.8-million-tweet, 1.3-billion-word corpus aimed at studying regional variation in Canadian English with a specific focus on the dialect regions of Toronto, Montreal, and Vancouver. Our data collection and filtering pipeline reflects complex design criteria, which aim to allow for both data-intensive modeling methods and user-level variationist sociolinguistic analysis. It specifically consists in identifying Twitter users from the three cities, crawling their entire timelines, filtering the collected data in terms of user location and tweet language, and automatically excluding near-duplicate content. The resulting corpus mirrors national and regional specificities of Canadian English, it provides sufficient aggregate and user-level data, and it maintains a reasonably balanced distribution of content across regions and users. The utility of this dataset is illustrated by two example applications: the detection of regional lexical and topical variation, and the identification of contact-induced semantic shifts using vector space models. In accordance with Twitter's developer policy, the corpus will be publicly released in the form of tweet IDs.

Keywords: Twitter, corpus construction, Canadian English, regional variation

1. Introduction

This paper presents a 1.3-billion-word corpus aimed at studying regional variation in Canadian English. It is part of a wider research effort which builds on previous sociolinguistic studies (Poplack et al., 2006; Boberg, 2012; Rouaud, 2019) and specifically seeks to investigate contact-induced semantic shifts in Quebec English. We aim to identify these linguistic traits by contrasting the English used in Quebec, where the majority of population is French-speaking, to varieties of English spoken in Canadian regions where the use of French is limited. Our approach relies on data-intensive methods such as distributional semantic models, but we maintain a variationist sociolinguistic perspective (Labov, 1972; Tagliamonte, 2006) grounded in fine-grained analysis of the linguistic behavior of individual speakers.

This methodological framework translates to the following corpus design criteria: (1) the corpus should reflect the specificities of the English spoken in Canada, as opposed to corpora of other national varieties of English or more generic datasets; (2) additional geographic metadata is necessary to compare different regional varieties of Canadian English: the province of origin of individual texts in the corpus is required as a minimum; (3) each regional subcorpus must meet a minimum size threshold of $\approx$ 100 million words in order for the proposed data processing methods to produce reliable results; (4) the reliance of these methods on features such as co-occurrence frequencies entails the need to limit sources of bias such as an irregular distribution of content across authors or a pervasive presence of spam or other types of noise; (5) sociolinguistic analysis of ongoing synchronic language variation requires data that is recent, largely contemporaneous, and produced in a reasonably spontaneous communicative context by individually traceable speakers; (6) the identification of individual speakers should allow us to examine inter-speaker variation within the local community: a description of the languages the individuals speak is necessary given Canada's multilingual environment and our focus on language contact.

As we were unable to find any existing corpus that could meet these criteria, we turned to Twitter, which provides large amounts of geotagged linguistic data together with basic user information. We collected tweets by identifying speakers geolocated in three cities corresponding to distinct dialect regions — Toronto, Montreal and Vancouver — and then crawling their Twitter timelines. We verified the pertinence of the collected data by filtering the location information indicated in the user profile, eliminating non-English tweets, and automatically excluding near-duplicate content. The resulting corpus contains 78.8 million tweets posted by 196,000 recently active users, with the content roughly equally distributed across the three regions. Its utility is illustrated by two case studies, respectively focusing on the detection of regionally-specific lexical variants and of contact-induced semantic shifts. In accordance with Twitter's developer policy, the corpus will be released as a list of tweet IDs together with instructions on how to collect the complete data using off-the-shelf software.[1]

The remainder of this paper is organized as follows. Section 2 presents an overview of related work; Section 3 discusses data collection and filtering methods; Section 4 outlines the corpus structure; Section 5 introduces the case studies conducted on the corpus; Section 6 provides a conclusion and possible directions of future work.

2. Related work

Diachronic semantic change, a research question closely related to synchronic semantic variation, has been addressed in recent years using distributional semantic models, which represent each word as a vector whose values reflect its co-occurrence statistics. A frequent approach involves training distributional models on large corpora from different time periods, and then comparing vector representations of a single word across the periods in order to detect semantic change (Gulordava and Baroni, 2011; Kim et al., 2014; Hamilton et al., 2016; Dubossarsky et al., 2017;

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

Del Tredici et al., 2019). Limited comparable work has also been done on synchronic semantic variation across domains (Del Tredici and Fernández, 2017; Fišer and Ljubešić, 2018).

The approach outlined above can be applied to the detection of synchronic semantic variation across Canadian regions by training distributional models on corpora from different regions. However, while large generic diachronic corpora of English are readily available, that is not the case for regional varieties of English used in Canada.

Corpus	Tokens	Geographic information	
Strathy	50m	country	text metadata
GloWbE	134m	country	text metadata
iWeb	308m	country	website domain
NOW	898m	country	text metadata
ENCOW16	222m	city	website IP address
JSI	1.3b	city	place of publication

Table 1: Existing corpora containing Canadian English data, with the size of the Canadian section (best estimates at the time of writing) and the granularity and origin of geographic information

Existing publicly available corpora of Canadian English are presented in Table 1 above. They include the Strathy Corpus of Canadian English (Strathy Language Unit, 2011), comprised of written and oral texts covering a variety of genres and historical periods, as well as the Canadian sections of multinational corpora such as Global Web-based English (GloWbE) (Davies, 2013a), News on the Web (NOW) (Davies, 2013b), and iWeb (Davies, 2018). However, these are all of limited utility in studies of regional variation, as the only provided geographic information is the country from which individual texts originate.

City-level geolocation is available in two large web-based corpora with Canadian content, but it is of questionable reliability. ENCOW16 (Schäfer and Bildhauer, 2012; Schäfer, 2015) derives geographic information from website IP addresses, meaning that it locates the servers hosting the websites rather than their users. In contrast, the JSI Newsfeed Corpus (Bušta et al., 2017) geotags online journalistic content based on its place of publication, but the solidity of this information is counterbalanced by considerable divergences in the amount of data originating from different Canadian regions. Moreover, other corpus design criteria, such as the ability to identify all linguistic content produced by the same speaker, are not met by any of the 6 cited corpora.

As for Twitter, it has been used to study variation and change in different regional varieties of languages including English (Doyle, 2014; Eisenstein et al., 2014; Huang et al., 2016; Shoemark et al., 2017) and Spanish (Gonçalves and Sánchez, 2014; Donoso and Sánchez, 2017). External demographic information has been used to approximate variables such as gender (Bamman et al., 2014) and ethnicity (Jones, 2015; Jørgensen et al., 2015; Blodgett et al., 2016).

3. Data collection and filtering

Similarly to previous work on collecting geotagged Twitter data (Ljubešić et al., 2014; Barbaresi, 2016), our data collection pipeline, illustrated in Figure 1, comprises two main

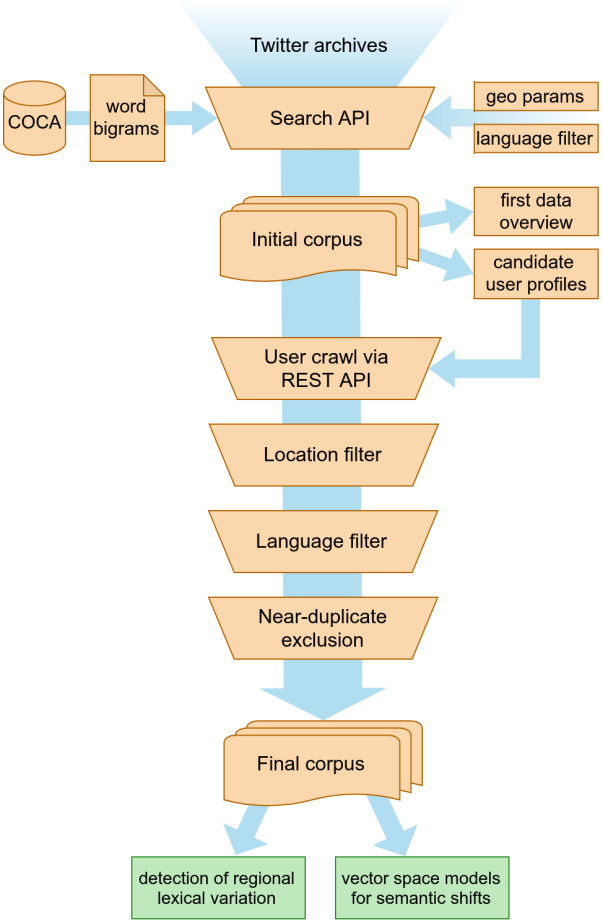

Figure 1: Data collection and filtering pipeline with possible applications of the corpus

steps: (1) an initial data collection which principally aims to identify Twitter users in geographic areas of interest; and (2) a subsequent crawl of the indexed users' timelines.

The first step was implemented by repeatedly querying Twitter's Search API in conjunction with geographic and linguistic filters. We used as search terms the 20,000 most frequent word bigrams in the 1-billion-word Corpus of Contemporary American English (COCA) (Davies, 2011). COCA is composed of texts that are roughly equally distributed over 30 years (1990-2019) and 8 genres, ranging from academic to spoken language. While the most frequent bigrams in the list are sequences of function words (e.g. *of the*), the majority include content words in commonly occurring patterns (e.g. *they work*, *my car*, *interest in*). Our approach is similar to the use of mid-frequency words to crawl web corpora (Baroni and Bernardini, 2004; Schäfer and Bildhauer, 2012), but like Scheffler (2014) we found that high-frequency search terms were more efficient on Twitter. Overall, this stage allowed us to identify English-speaking users living in Toronto, Montreal and Vancouver, and more generally to gain an initial insight into the gathered data.

The second step consisted in collecting all available tweets published by the initially indexed users. The aim was to increase the amount of available data while balancing the size of the regional subcorpora, as well as to obtain enough tweets published by individual users to analyze speaker-specific linguistic patterns. Tweets written in all languages

were initially retained to allow for a description of the overall linguistic profile of the corpus.

The collected data were subsequently filtered by (1) verifying user profile locations to confirm that they reference one of the targeted cities; (2) excluding tweets written in languages other than English; (3) excluding near-duplicate tweets to limit the impact of repetitive or automatically generated messages.

3.1. Choice of geographic areas

Tweet collection was geographically constrained to Canada's three most populous cities: Toronto (Ontario), Montreal (Quebec), and Vancouver (British Columbia). From a practical point of view, the choice of these cities was motivated by the need for a sufficiently large local user base allowing us to collect enough data over a reasonably short period of time. Moreover, and more crucially, the cities belong to distinct dialect regions (Boberg, 2005), as is also evidenced by their demographic profile.

Montreal is home to 80% — or around 890,000 — of Quebec's English speakers, but they represent only 22% of the city's population. Conversely, in Toronto and Vancouver the dominant language is English, with only 1.7% and 1.3% of population, respectively, speaking French (Statistics Canada, 2017). We aim to detect contact-related phenomena, and limit the impact of those deriving from unrelated regional variation, by examining the linguistic properties that are specific to Montreal and distinguish it from both Toronto and Vancouver.

Data collection was limited to tweets sent from the metropolitan areas of the three cities, all of which are highly multicultural. This means that our corpus may contain messages posted by non-native speakers of English. We therefore attempted to create corpora of smaller, more homogeneous communities (West Island of Montreal; Oshawa–Whitby, ON; Victoria, BC), but this led to a multifold decrease in collected data and was deemed too inefficient.

3.2. Initial tweet collection

An initial corpus was created using Twitter's Search API, which looks up queries in a sample of recently published tweets. The queries were filtered geographically by indicating the targeted areas as a radius around a point of interest, defined using geographic coordinates. Since this stage only aimed to identify English speakers, data collection was restricted to tweets tagged by Twitter as written in English. Moreover, search parameters were used to exclude retweets from the results: the diffusion of content posted by others may be indicative of the popularity of different subjects across regions, but our focus is on individual users' linguistic production rather than their topical interests.

As mentioned above, we queried the Search API using the 20,000 most frequent word bigrams from COCA. For each bigram in the list, all available tweets in the targeted geographic areas were collected. As a single iteration over the entire list takes an average of 5 days, repeating iterations allows us to move chronologically through Twitter's archives. By the time an iteration is completed, the temporal window of available tweets (6–9 days preceding the query) also shifts, meaning that the next iteration mostly returns previously unavailable data.

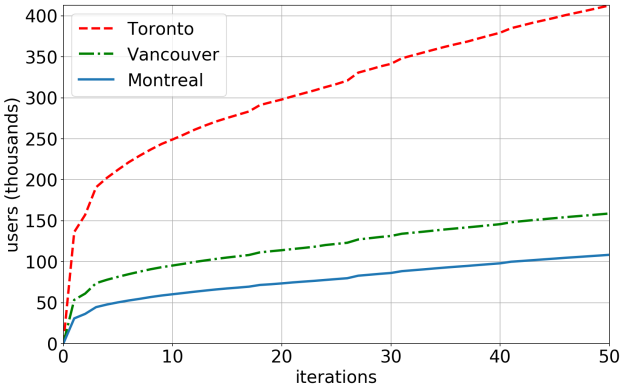

Figure 2: Cumulative number of identified users per subcorpus

A total of 50 iterations were completed between mid-January and mid-November 2019. The resulting corpus contains 58,451,998 tweets published by 679,785 distinct users. As shown in Figure 2, 50.6% of users were identified in the first 5 iterations, but subsequent queries still provided a constant and non-negligible flow of new data. However, the number of collected tweets per user varies considerably (top 1% of users account for 36.6% of tweets), as does the number of identified users across regions (108,383 in Montreal, 158,762 in Vancouver and 412,640 in Toronto). That said, this initial dataset is a valuable starting point for more controlled user-level tweet collection.

The search method was chosen over the better-known Streaming API, which returns a real-time sample of tweets, as it yielded considerably more data. For comparison, we ran the Streaming API for 30 days in October 2019 with comparable geographic parameters, obtaining 925,668 tweets published by 57,218 individual users. Over the same period of time, 6 iterations of the Search method were completed, yielding 8,332,629 tweets published by 303,538 users. In other words, the use of the Streaming API led to a roughly ninefold decrease in collected data and a fivefold decrease in identified users compared to our approach.

This is largely due to the fact that the Streaming API only takes into account tweet-level location data: precise geolocation, when the tweet is tagged with the precise geographic coordinates of the user's location at the time of tweeting; or manual geolocation, when the user chooses the place associated with the tweet from a list of proposed options or by looking up a specific place. These features are only available on mobile devices and are actively used by a fraction of all users, which limits the availability of geotagged tweets. In our case, this is further affected by tight geographic constraints and a comparatively small number of targeted users (especially English-speaking Montrealers). An alternative solution was proposed for the German Twitter Snapshot (Scheffler, 2014), which collected tweets published in Germany by tracking words specific to German rather than applying geographic filtering. We could not implement this method, as only a fraction of all English-language tweets are posted in Canada.

As for the Search API, it maximizes the amount of data returned by geographic queries by interpreting non-geotagged

tweets (93% of our initial dataset) as sent from the location indicated in the user profile. In related previous work, corpora were created for three closely related languages with limited coverage in the Streaming API — Croatian, Serbian and Slovene — by querying the Search API using words specific to the targeted languages, without geographic parameters (Ljubešić et al., 2014). While this approach avoids issues related to the reliability of geolocation, it is not directly applicable to our case. Lexical variants distinguishing dialect regions are both less numerous and less frequent than words differentiating distinct languages, which would impact the efficiency of this method and would potentially introduce a bias towards speakers who use regionalisms more frequently.

Although both tweet-based and profile-based geolocation may introduce a demographic bias (Pavalanathan and Eisenstein, 2015), our reliance on manually indicated user profile location is justified by a considerable increase in collected data as well as by its sociolinguistic significance: this location corresponds to the place users consciously associate with their online presence. Moreover, precise tweet-level geolocation will become limited to tweets containing photos or videos,[2] which will affect data collection pipelines relying on this type of geographic information.

3.3. User profile crawling

After identifying a sufficient number of users, their entire timelines were crawled subject to the Twitter-imposed limit of 3,200 most recent tweets per user (including retweets). In order for the final regional subcorpora to be comparable in size, we included the 108,383 users indexed in the initial Montreal subcorpus, as well as the same number of randomly sampled users in each of the larger Toronto and Vancouver subcorpora. The crawl was performed for two batches of users, in April and November 2019, respectively. In addition to excluding retweets based on Twitter metadata, we followed common practice in eliminating the messages that contain the `RT @` string in their text. This case, affecting 0.7% of collected tweets, corresponds to comments of other users' messages embedded in tweet text so it has the potential to distort user-level word frequencies. Moreover, we only retained tweets with at least 2 words in addition to any hashtags, user handles and URLs. While this led to the exclusion of 8.7% of collected tweets, it ensured that each retained tweet contained at least some linguistic content as opposed to being a list of Twitter-related entities. Unlike in the initial data collection, no language restrictions were sent to the Twitter API in order to allow for a subsequent analysis of the languages that are actively used by individual users.

3.4. Location filtering

Since we are interested in the linguistic communities of Toronto, Montreal and Vancouver, we aim to restrict data collection to the users who explicitly declare that they live in these cities. While the geographic parameters used with the Search API correspond to these areas, some users in the corpus may have been identified independently of their profile locations, based solely on individually geotagged tweets. Others still may have been retained even though multiple cities are indicated in their profile.

We therefore used a heuristic to additionally filter the places indicated in the `location` field in the user profile. In order for a user to be retained, the field must include the name of one of the examined cities (e.g. *Montreal*). It may additionally include the name of the corresponding province (e.g. *Quebec*), the name of the country (e.g. *Canada*), as well as generic geographic descriptors (e.g. *north, greater, metro* etc.). No other elements were accepted.

In the Montreal subcorpus, profile locations were indicated in 7,719 distinct ways (after being lowercased and stripped of punctuation and diacritics). Of these, 46 meet the above criteria and were used by 69% of the identified users. The individual realizations differ in terms of the order and precision of included information (*Montreal* vs. *Montreal West, Quebec*), orthographic choices (*Montreal* vs. *Montréal*), use of abbreviations (*Quebec* vs. *QC*) and punctuation. Out of the 7,673 rejected locations, 6,872 (used by 22% of users) indicate multiple targeted cities (*Montreal & Toronto*), places outside of the search area (*Ottawa*) or insufficient geographic information (*Canada*). The remaining 801 locations (used by 9% of users) refer to neighborhoods (*Plateau Mont-Royal*) or points of interest (*McGill University*) in the search area, but were excluded due to the presence of lexical items which are too specific to include in the filtering heuristic. Based on the number of classified users, the Montreal subcorpus heuristic obtained an F-score of 0.94. Comparable patterns were also observed in the Toronto and Vancouver subcorpora.

3.5. Language identification

As previously mentioned, the populations of Toronto, Montreal and Vancouver are all highly multilingual. While the initial data collection parameters ensure that the identified users have sent at least one tweet tagged as English, crawling their entire timelines provides a clearer picture of the languages they actually use. The distribution of language tags outlined in Table 2 shows that English is by far the most frequent language in the corpus, but, in addition to the expected use of French in Montreal, immigrant languages are also present. Since we only aim to investigate regional differences affecting English, tweets tagged as written in other languages (15.5% overall) were excluded.

Montreal		Toronto		Vancouver	
en	69.7%	en	93.4%	en	92.4%
fr	22.6%	es	1.2%	es	1.6%
es	2.3%	tl	.8%	pt	1.1%
pt	.7%	pt	.7%	tl	.9%
ar	.6%	fr	.6%	fr	.6%
other	4.1%	other	3.3%	other	3.5%
total	100.0%	total	100.0%	total	100.0%

Table 2: Distribution of tweets across the top language tags (components may not sum to totals due to rounding)

The decision to use Twitter-provided language tags was preceded by an evaluation of third-party systems on a manually annotated sample of 494 monolingual English tweets

[2] `https://twitter.com/TwitterSupport/status/1141039841993355264`

and 420 monolingual French tweets, grouped into balanced categories with 2, 5, 10, 15 or 20 words per tweet. We focused on English and French because, in addition to being Canada's two official languages and the center of our research objectives, they correspond to the most frequent language tags in the corpus. We tested three widely used off-the-shelf language identification systems — `langid.py` (Lui and Baldwin, 2012), `cld2` (McCandless, 2014) and `langdetect` (Nakatani, 2010) — and a majority-vote system combining the three methods, proposed in an earlier evaluation (Lui and Baldwin, 2014). The results in Table 3 show that all systems are consistently reliable except on very short tweets. As expected, the vote-based system performs on par with or improves on the best individual F-scores.

We further compared the performance of the evaluated systems to the language tags indicated in tweet metadata. While we are unable to report the quantitative results because Twitter's developer policy[3] prohibits the benchmarking of their services, we do not find it necessary to implement a third-party language identification system in our pipeline. The systems we evaluated on English and French occasionally provide marginal improvements compared to Twitter's tags, but their performance is overall less consistent.

System	Words per tweet					
	2	**5**	**10**	**15**	**20**	**all**
langid	.822	.964	.989	**.994**	**1.000**	.963
langdetect	.896	.917	.989	.989	**1.000**	.963
cld2	.793	.898	.971	.967	**1.000**	.935
vote	**.902**	**.976**	**.994**	**.994**	**1.000**	**.979**

Table 3: Macro-averaged F-score on manually annotated English and French tweets of different lengths

However, the use of Twitter's language tags raises another potential issue. Practices such as borrowing and codeswitching are frequent among bilingual speakers, meaning that multiple languages may be used in a tweet, whereas only one language tag is indicated in the metadata. This problem was evaluated on a balanced sample of 1,000 tweets tagged by Twitter as English or French. We manually identified other-language content in 65 tweets: 60 written in these two languages, and 5 written in English or French and another language. Note that most identified tweets (56 out of 65) were tagged as French.

We attempted to automatically identify the languages in multilingual tweets using the top 2 predictions produced by each of the tested language identification methods. A majority vote system was also implemented based on the 2 most frequent language tags from the individual predictions. The best accuracy was obtained by `langdetect`, which correctly analyzed 25% of tweets.

Given the relative rarity of other-language items and the poor performance of the tested language identification systems, multilingual content filtering has not been implemented. Word-level language identification may provide more precise results and is a possible direction of future work.

[3] `https://developer.twitter.com/en/developer-terms/agreement-and-policy`

3.6. Near-duplicate exclusion

A frequent issue in Twitter-based corpora is the presence of near-duplicate messages generated by both automated spam accounts and prolific human users. Attempts are usually made to filter out this content as it can bias word frequencies. A common approach consists in excluding accounts that exceed defined cut-off points in terms of the number of tweets, followers, followees etc., or in excluding all tweets containing URLs or other specific strings. These methodological decisions are based on the potential link between these user account features and spam production (Yardi et al., 2010).

Such solutions, however, do not take into account the fact that user behavior on Twitter is often heterogeneous. We manually analyzed the 20 users in our corpus with the highest number of posted tweets in their profiles. We observed that 7 accounts indeed publish exclusively near-duplicate content such as song titles played by radio stations, while another 2 post a mix of similarly generated tweets and spontaneous messages. However, the remaining 11 accounts are all consistent with genuine human communication.

As 2 of these are corporate Twitter profiles where different social media managers interact with the public, we focused on the 9 accounts which are used by individual speakers. To varying extents, they all produce genuine tweets as well as ones that are automatically generated by, for example, posting content on other social media sites. In some cases, the high number of published tweets is actually driven by retweets, while the content of original posts is similar to that of average accounts. Moreover, while some tweets containing URLs merely reference external content (e.g. titles of linked videos), others include fully acceptable messages.

Taking into account this variety of behaviors, we implemented a system whose aim is not to exclude all tweets posted by the users most likely to produce spam, but rather to distinguish, within the production of each individual user, the tweets that are of genuine interest from near-duplicate content. For each user, a distance matrix was calculated for all their tweets. We used Levenshtein's distance, which quantifies the difference between two strings of characters as the number of edit operations (character insertions, deletions or substitutions) necessary to modify one string of characters into the other.

As our aim is to exclude messages with similar linguistic content independently of Twitter-specific entities, we removed hashtags, user handles and URLs from tweet text. In calculating the absolute Levenshtein's distance, replacement operations were assigned a weight of 2 in order for the distance between entirely different strings of characters to be equal to the sum of their lengths. This distance was then normalized by dividing it with the total number of characters in a pair of tweets. A normalized score of 0 corresponds to identical strings, and a score of 1 to strings with no overlapping characters.

After calculating the distance matrix, near-duplicate tweets were identified using hierarchical clustering. We excluded all clusters where the distance between individual tweets did not exceed 0.45. This cut-off point was determined empirically, as it was found to correspond to a reasonable balance between precision and recall. While the identifi-

cation of near-duplicates published by different users may further improve the quality of our data, it is computationally prohibitively expensive with the current method.

We more closely examined the performance of our system by focusing on precision; recall was not evaluated at this stage due to the significantly higher complexity of the task for human annotators. We analyzed a random sample of 10 user profiles, for which we had collected a total of 6,317 English-language tweets. The system excluded 1,956 tweets, grouped into 691 distinct clusters, as near-duplicates. For each cluster, we manually annotated the degree of similarity of the identified tweets: strong, for tweets that are near-identical in form and meaning; partial, when repetitions concern parts of the text, while the meaning remains similar across the tweets; weak, when isolated words are shared between tweets whose overall meaning is not strongly related; none, when similarity is only reflected by character patterns rather than words.

Similarity	Clusters		Tweets	
Strong	209	30.2%	896	45.8%
Partial	278	40.2%	645	33.0%
Weak	165	23.9%	337	17.2%
None	39	5.6%	78	4.0%
Total	691	100.0%	1,956	100.0%

Table 4: Analysis of clusters of tweets identified as near-duplicates (components may not sum to totals due to rounding)

As indicated in Table 4 above, 21.2% of analyzed tweets were affected by some type of misclassification, i.e. exclusions based on weak or inexistent similarity. However, the latter category, corresponding to clearly erroneous exclusions related to the simplicity of Levenshtein's distance as a similarity measure, was limited to 4% of tweets. As for the tweets presenting weak similarity (17.2%), they are admittedly related to minor overlaps in content, but it is difficult to estimate how beneficial their inclusion would have been for the corpus as they are often limited size or informational value. Moreover, these exclusions are related to structural similarity rather than, for example, specific topics, so they are not expected to negatively affect co-occurrence statistics. We overall consider this loss of data to be outweighed by the benefits of cleaner, less repetitive content.

4. Corpus description

The corpus obtained after crawling individual user profiles, performing language and location filtering and excluding near-duplicate content contains 78.8 million tweets posted by 196,431 individual users. After tokenizing the corpus with `twokenize` (Gimpel et al., 2011; Owoputi et al., 2013), this corresponds to 1.3 billion tokens. On average 401 tweets were collected per user; the top 1% of users account for 6.2% of tweets, which represents a considerable improvement compared the initial stage of data collection. The data is roughly equally distributed across the three regional subcorpora.

The structure of the final corpus is presented in Table 5 below. Token counts were limited to the metadata-indicated display text range, i.e. tweet text stripped of tweet-initial user handles referring to conversation chains and of tweet-final URLs mostly used to embed media. This represents 97.7% of analyzed text content. No further removal of Twitter-related entities was performed, as they are often syntactically integrated in the tweet text and can also provide insights into bilingual communication (e.g. hashtags used in a language different from the rest of the tweet).

Subcorpus	Users	Tweets	Tokens
Montreal	72,305	23,469,526	384,740,451
Toronto	64,164	28,442,928	481,126,844
Vancouver	59,962	26,924,158	473,322,674
Total	196,431	78,836,612	1,339,189,969

Table 5: Corpus structure

We initially crawled 325,000 Twitter profiles across the three cities. Of these, nearly 11,000 were inaccessible at the time of the crawl because they had been deleted or had become private following their initial identification. While this is a tolerable loss of data (3.2% of accounts), we plan to improve the efficiency of our pipeline by crawling individual user profiles as soon as they are identified by the Search API.

More significantly, 118,000 accounts (36.2%) were excluded based on their profile location. Out of the 132 million tweets retained after user-level geographic filtering, 15.5% were rejected because they were not written in English and a further 24.7% were excluded as near-duplicates. The filters we implemented led to a considerable reduction in corpus size, but they ensure the reliability of collected data.

Before the exclusion of non-English-language content from the corpus, the users were analyzed according to the languages they use on Twitter. For each user, we calculated the proportion of English language tweets (out of all English and French tweets) and the proportion of tweets in English and French (out of all tweets). Figure 3 above suggests that we identified predominantly English-speaking individuals, as well as some demonstrably bilingual speakers. As expected, the use of French is more frequent in the data collected in Montreal compared to the other two cities, whereas

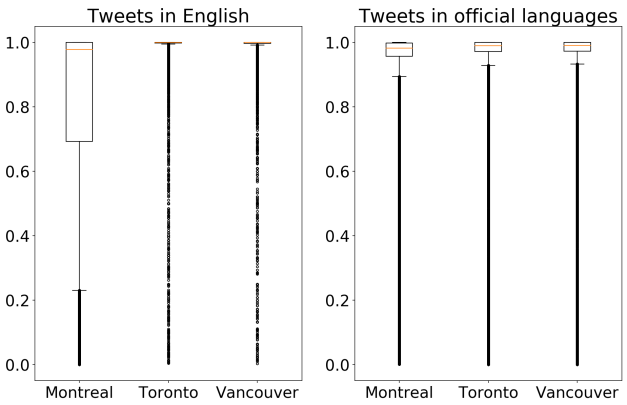

Figure 3: Left: proportion of tweets in English per user (out of tweets in English and French). Right: proportion of tweets in English and French per user (out of tweets in all languages). Results based on language tags produced by Twitter prior to the exclusion of non-English content.

the use of non-official languages (i.e. languages other than English and French) is comparable across the subcorpora.

5. Example applications

5.1. Regional lexical and topical variation

Following previous studies on social media (Eisenstein, 2015; Chandrasekharan et al., 2017; Shoemark et al., 2018), we investigated the most distinctive regional lexical variants in the corpus using the publicly available Python implementation[4] of the Sparse Additive Generative model (SAGE) (Eisenstein et al., 2011). SAGE estimates the deviation in log-frequencies of terms in a corpus of interest relative to their log-frequencies in a background corpus using the maximum-likelihood criterion, with a regularization parameter ensuring that rare terms are not overemphasized. A high value of the deviation estimate indicates that a term is overrepresented in a given corpus, and a low value that it is underrerpresented.

We more closely examined the 100 most distinctive lexical items from the Montreal subcorpus and manually categorized them into distinct types of lexical phenomena. Representative examples for each category are presented in Table 6 and further discussed below.

Category	**Examples**
Regionalisms	*metro* (57.7), *supper* (19.5)
Local referents	*montreal* (791.3), *habs* (228.1), *drouin* (20.3), *poutine* (35.5)
Spelling	*café* (14.6), *center* (65.7), *youre* (43.1)
Chatspeak	*loll* (24.5), *lolll* (14.9), *fkn* (54.4)
French items	*une* (16.7), *dans* (16.7), *merci* (32.4)

Table 6: Categories of lexical items specific to Montreal, with frequency per million words indicated in brackets

As expected, this method identified known regional variants that are related to contact with French. The word *metro* is associated with the French term *métro*, used as the official name of Montreal's underground railway system. The preference for the term *supper* may likewise be related to the similarity of the corresponding Quebec French term *souper* (Boberg and Hotton, 2015). Importantly, the alternative variants used in other regions (*subway* and *dinner*, respectively) feature among the most underrepresented items in the Montreal subcorpus.

Some lexical items are more frequent in Montreal than elsewhere because of the local importance of their referents. This is the case of the French borrowing *poutine*, which denotes the typical Quebec dish consisting of French fries topped with gravy and cheese. This category also includes local toponyms, as well as the names of sports teams (*habs* 'Habs', the nickname of the Montreal Canadiens hockey team) and of their players (*drouin* 'Jonathan Drouin').

Other identified lexical items point to understudied regional spelling preferences. For instance, the American spelling variant *center* is overrepresented in the Montreal subcorpus, whereas the British variant *centre*, typically associated with Canadian English, is underrepresented. A potentially related issue is a higher prevalence of apostrophe dropping in Montreal, as in *youre* (vs. *you're*). While more work is needed to explain these two patterns, other cases, such as the frequent use of accented letters in words such as *café*, appear more directly related to the influence of French.

Some chatspeak featues also seem to reproduce patterns typical of French. The variants of *lol* 'laugh out loud' typical of Montreal emphasize the final consonant (*loll*, *lolll*), similalrly to the corresponding French initialism *mdr* (e.g. *mdrr*, *mdrrr*), based on the expression *mort de rire*, literally meaning 'dead of laughter'. Conversely, the forms salient for Toronto feature an orthographic lengthening of the vowel (e.g. *lool*, *loool*). As for the abbreviation *fkn* 'fucking', its prevalence in Montreal may be related to the widespread use of the expletive *fuck* and derived forms in Quebec French, where they are stripped of the vulgar connotation present in English (Meney, 2017). Given the influence of French, bilinguals may perceive the term as inoffensive in both languages and therefore use it more frequently in English, but this hypothesis should be tested more extensively.

Finally, a number of French items were identified due to their use in messages with both French and English content. As we have previously argued, codeswitching is overall rare in this dataset; the prominence of these items is related to the higher relative frequency of codeswitching in the Montreal subcorpus. We observe that the attested codeswitching patterns vary in structure and thereby reflect real-life usage. Consider the following examples:

1. **On devrait juste interdire les commentaires.** That's it. Then again, no more FB or Twitter...
   They should just forbid comments. *That's it. Then again, no more FB or Twitter...*

2. Hi there, guys! We always appreciate the support. You're the best! **Merci!**
   Hi there, guys! We always appreciate the support. You're the best! ***Thanks!***

In example (1), the user produced a complete sentence in French and then switched to English for the remainder of the tweet; the switch was possibly triggered by the use of the fixed expression *that's it*. In example (2), only the term *merci* was used in an otherwise English-language tweet. Since the message was addressed to a user from the United States, this can be seen as an expression of local identity.

Overall, this analysis shows that our corpus captures differences in the use of known lexical variants as well as locally-specific topics, confirming the regional representativeness and comparability of the data. Moreover, the variety of observed codeswitching patterns shows that language contact can manifest itself on Twitter in many of the ways it does in spoken language. Other, as yet understudied types of variation, such as regional spelling and abbreviation preferences, are specific to online communication and represent an added value of Twitter in sociolinguistic research.

5.2. Contact-induced semantic shifts

Having established the presence of regional linguistic features in the corpus, we explored the detection of contact-induced semantic shifts using distributional semantic models. We used `word2vec` (Mikolov et al., 2013) to train

[4] `https://github.com/jacobeisenstein/SAGE`

an embeddings model for each regional subcorpus. Default hyperparamaters were applied (skip-gram architecture with negative sampling rate of 5, window size of 5, embeddings of size 100, subsampling rate of 10^{-3}, number of iterations set to 5) with the minimum word frequency set to 100.

This method produces low-dimensional vector spaces, in which vector dimensions are not naturally aligned to the same coordinate axis, meaning that the models are not directly comparable. We followed work on diachronic embeddings (Hamilton et al., 2016) and aligned the models using Orthogonal Procrustes, available in a Python implementation.[5] This allowed us to directly compute the cosine-distance between each word's vectors in all pairs of models so as to detect the most prominent divergences in Montreal. This approach successfully identified a number of words exhibiting contact-induced meanings. Examples including *exposition* 'exhibition' and *terrace* 'restaurant patio' reflect the findings of previous sociolinguistic studies (Boberg, 2012), while newly identified cases such as *definitively* 'definitely' present comparable contact-related influence: the unconventional meanings are all likely related to French cognates (*exposition*, *terrasse* and *définitivement*, respectively). Other regional semantic variants are of more limited interest for sociolinguistics, as they are related to proper nouns (*plateau* denoting the borough of Plateau-Mont-Royal in Montreal) or cultural factors (*chum* referring to a species of salmon in Vancouver, which lies on the Pacific Ocean).

The obtained models tend to strongly emphasize a single meaning for each word. In the Montreal model, this creates the impression that language contact may have completely altered the way in which a word is used by an entire speech community, whereas the identified words are in fact attested with a variety of meanings. A clear case is illustrated by the following tweets, both posted by users from Montreal:

3. My fav thing about my new job is the rooftop **terrace**
4. Nothing like drinks on the **terrace** of your fave pub to end the semester

In example (3), the word *terrace* is used with the conventional meaning referring to a flat roof; in example (4) it is attested with the meaning typical of the French word *terrasse*, denoting an outdoor seating area.

A manual analysis suggests that the contact-related meanings of this and other words are overall more frequent in Montreal, but that they are mostly used by users who tweet in both English and French, including those who live in Toronto and Vancouver. This has important implications for the sociolinguistic status of contact-induced semantic shifts, as they appear to represent a variation in usage limited to bilinguals rather than being established regional variants typical of Quebec English.

We intend to refine these results using context-informed word representations such as BERT (Devlin et al., 2018), which should allow us to explicitly account for polysemy and more easily quantify aggregate and speaker-specific preferences. Already at this stage, though, we have shown that our corpus can be used for the automatic detection of regional semantic variants, while the availability of user-level data contributes to a more complete understanding of the precise patterns of variation that are at play.

[5] https://github.com/williamleif/histwords

6. Conclusion and future work

We have presented a new 78.8-million-tweet corpus aimed at studying regional variation in Canadian English. Tweets posted by users based in Toronto, Montreal and Vancouver were collected using a time-efficient pipeline complemented by location, language and near-duplicate filtering. The resulting corpus meets the initially defined design criteria: it mirrors both national and regional specificities of Canadian English, it is sufficiently large for data-intensive modeling methods as well as fine-grained user-level analysis, and it maintains a reasonably balanced distribution of reliable content across regions and users. Although Twitter's terms of use preclude us from publishing the entire corpus, the release of pre-filtered tweet IDs provides a useful starting point for other studies of Canadian English.

The presented example applications show that the corpus facilitates the study of both previously described and novel regional linguistic variants. The large amount of available data allowed us to use methods such as word embeddings, which otherwise could not have been applied to regional variation in Canadian English, whereas the focus on individual users brought more clarity to the precise status of regional linguistic variants. These observations are complemented by ongoing work aiming to identify distinct profiles of users based on the use of contact-related linguistic variants and on associated extra-linguistic factors reflected by Twitter metadata. This line of inquiry will allow us to move beyond Twitter-focused analysis and formulate more precise research hypotheses on the status and representations of regional linguistic forms in spoken Canadian English.

While previous studies have suggested that aggregate geographic patterns observed on Twitter correlate with traditional dialectological studies (Doyle, 2014), we aim to shed more light on the precise relationship between user-level linguistic choices observed on Twitter and real-life sociolinguistic behaviors. We seek to further investigate the computationally identified linguistic variants and their social correlates through sociolinguistic fieldwork. This specifically involves a face-to-face survey based on a well-established methodological framework (Durand and Przewozny, 2012), focusing on a sample of native Canadian English speakers who reflect the linguistic profiles identified in our corpus. Our objective is to explicitly evaluate the reliability of linguistic information and metadata provided by anonymous Twitter users in the context of variationist sociolinguistic studies. This will in turn help inform future work on the collection and interpretation of linguistic data on social media.

7. Acknowledgments

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

8. Bibliographical References

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, April.

Barbaresi, A. (2016). Collection and indexing of tweets with a geographical focus. In Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 24–27.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In LREC, page 1313.

Blodgett, S. L., Green, L., and O'Connor, B. (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.

Boberg, C. and Hotton, J. (2015). English in the Gaspé region of Quebec. *English World-Wide*, 36(3):277–314.

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

Bušta, J., Herman, O., Jakubíček, M., Krek, S., and Novak, B. (2017). JSI Newsfeed corpus. In The 9th International Corpus Linguistics Conference, Birmingham, UK.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31.

Davies, M. (2011). N-grams data from the Corpus of Contemporary American English (COCA).

Davies, Mark. (2013a). Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE).

Davies, Mark. (2013b). Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day.

Davies, Mark. (2018). The 14 Billion Word iWeb Corpus.

Del Tredici, M. and Fernández, R. (2017). Semantic variation in online communities of practice. In IWCS 2017 - 12th International Conference on Computational Semantics - Long papers.

Del Tredici, M., Fernández, R., and Boleda, G. (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), June.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donoso, G. and Sánchez, D. (2017). Dialectometric analysis of language variation in Twitter. In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial4).

Doyle, G. (2014). Mapping dialectal variation by querying social media. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 98–106.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1136–1145, Copenhagen, Denmark, September. Association for Computational Linguistics.

Durand, J. and Przewozny, A. (2012). La phonologie de l'anglais contemporain : usages, variétés et structure. *Revue française de linguistique appliquée*, 17(1):25–37.

Eisenstein, J., Ahmed, A., and Xing, E. P. (2011). Sparse additive generative models of text. In Proceedings of the International Conference on Machine Learning (ICML), pages 1041–1048.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114, November.

Eisenstein, J. (2015). Written dialect variation in online social media. *Charles Boberg, John Nerbonne, and Dom Watt, editors, Handbook of Dialectology. Wiley*.

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

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. 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, pages 42–47.

Gonçalves, B. and Sánchez, D. (2014). Crowdsourcing dialect characterization through Twitter. *PLoS ONE*, 9(11):e112074, November.

Gulordava, K. and Baroni, M. (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, GEMS '11, pages 67–71, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). 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, August. Association for Computational Linguistics.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255, September.

Jones, T. (2015). Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90:403–440.

Jørgensen, A., Hovy, D., and Søgaard, A. (2015). Challenges of studying and processing dialects in social media.

In Proceedings of the Workshop on Noisy User-generated Text, pages 9–18.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (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, June. Association for Computational Linguistics.

Labov, W. (1972). Sociolinguistic patterns. University of Pennsylvania Press.

Ljubešić, N., Fišer, D., and Erjavec, T. (2014). TweetCaT: a tool for building Twitter corpora of smaller languages. In Proceedings of LREC.

Lui, M. and Baldwin, T. (2012). langid. py: An off-the-shelf language identification tool. In Proceedings of the ACL 2012 system demonstrations, pages 25–30. Association for Computational Linguistics.

Lui, M. and Baldwin, T. (2014). Accurate language identification of Twitter messages. In Proceedings of the 5th workshop on language analysis for social media (LASM), pages 17–25.

McCandless, M. (2014). Chromium Compact Language Detector. `https://code.google.com/p/chromium-compact-language-detector`.

Meney, L. (2017). Le français québécois entre réalité et idéologie: un autre regard sur la langue: étude sociolinguistique. Presses de l'Université Laval.

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

Nakatani, S. (2010). Language detection library (slides). `https://www.slideshare.net/shuyo/language-detection-library-for-java`.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. 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, pages 380–390.

Pavalanathan, U. and Eisenstein, J. (2015). Confounds and Consequences in Geotagged Twitter Data. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2138–2148, Lisbon, Portugal. Association for Computational Linguistics.

Poplack, S., Walker, J. A., and Malcolmson, R. (2006). An English 'like no other'?: Language contact and change in Quebec. *Canadian Journal of Linguistics*, 51(2-3):185–213.

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

Scheffler, T. (2014). A German Twitter snapshot. In Proceedings of LREC, pages 2284–2289.

Schäfer, R. and Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3), Lancaster. UCREL, IDS.

Shoemark, P., Sur, D., Shrimpton, L., Murray, I., and Goldwater, S. (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, pages 1239–1248.

Shoemark, P., Kirby, J., and Goldwater, S. (2018). Inducing a lexicon of sociolinguistic variables from code-mixed text. In Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pages 1–6, Brussels, Belgium, November. Association for Computational Linguistics.

Statistics Canada. (2017). First official language spoken, language spoken most often at home, age and sex for the population excluding institutional residents of Canada, provinces and territories, census metropolitan areas and census agglomerations (table). 2016 Census.

Strathy Language Unit. (2011). Strathy Corpus of Canadian English.

Tagliamonte, S. A. (2006). Analysing sociolinguistic variation. Cambridge University Press.

Yardi, S., Romero, D. M., Schoenebeck, G., and boyd, d. (2010). Detecting spam in a Twitter network. *First Monday*, 15(1).