HELSINGIN YLIOPISTO

Activism and Xenophobia

Corpus-Based Critical Discourse Analysis of Immigration-Related Tweets

Jonas Haverinen
ENG-3103 Corpus Linguistics
English Philology
Department of Modern Languages
University of Helsinki
December 21, 2018

Introduction

Immigration has been a constant topic of news, scrutiny and debate for some years, and with each new crisis, discussion around the topic seems to increase noticeably. Twitter, as social media usually does, allows its users to stay informed of ongoing events and react to them instantaneously and with quite few restrictions. Twitter discussions, especially ones regarding issues such as immigration, often contain users' personal beliefs and are commonly very direct or even hostile. Motivated by recent immigration-related events and media coverage thereof, I believe there is much to be gained from applying methods of corpus linguistics and critical discourse analysis to the "twittersphere" of related topics. As language often reflects the ideological stances of its users, social media can be a vital resource for mapping these ideologies.

I borrow some of my ideas and methodology from Gabrielatos & Baker (2008) and Coats (2017), as the studies aided me in coming to the research question. Gabrielatos & Baker (2008) produced, among other things, consistent collocates of immigration-related lexicon. The diachronic study employed texts from the British press, establishing certain collocates and word formations that occurred more consistently in broadsheet media compared to tabloids. The findings displayed very negative representations of immigrants, refugees and other such groups, especially in tabloid media. The corpus-based study is an excellent example of how corpus linguistics can be used to determine the views possessed and spread by certain entities. Coats (2017), on the other hand, collected English, gender-tagged tweets from Finland and the United States, hoping to establish linguistic differences between Finnish male and female users of English, compared to American ones. The collected tweets were divided into subcorpora based on user location and gender. The study found that while typical gender differences in CMC apply – for example, female speakers adopt new linguistic features more readily and use more affect markers – local norms still apply strongly. While the study does not represent critical discourse analysis, it functions as an example of mapping regional variation in a corpus-based approach and uses subcorpora to classify the data efficiently. In this paper, I hope to utilize some approaches and tools provided by these studies. My primary aim in this rather small-scale study is to map the prominent views of immigrants and related terms on Twitter through scrutinizing collocates and keywords – and to determine whether these views vary noticeably by the search term or location.

My hypothesis is that due to the recent events in immigration, and especially the media reports regarding them, much of the language around immigrants, refugees and asylum seekers is negative. This may stem from the news themselves portraying them in a negative light or the users having such perceptions of them. I also predict that tweets originating in the United States will display particularly negative ideologies regarding the issues. For this small-scale study, I created my own script to collect and parse tweets – the program in its entirety can be found here.

Methods

The data used for this study consists entirely of tweets collected through the Tweepy streaming library for the Python 3.6 programming language. Through Twitter's API (application program interface), the Tweepy library allows Twitter-registered programs easy access to incoming public tweets, which can then be stored for later analysis. Tweets are transferred as JSON objects, with specific fields for all data contained within a tweet, including the user, mentions of other users, hashtags, the (automatically detected) language and country of origin. My primary requirement for collecting a tweet was the English language tag. The tweet's text would also have to include the strings *refugee*, *migrant* or *asylum seeker* – this ensured the relevancy of the tweet. The formatting also takes into account fords which the three search terms are substrings of, which means plural forms and the words *immigrant* and *emigrant* were also accepted. I rejected all retweets – which consist solely of a previously posted tweet – as they could lead to a notable amount of duplicates in the data as well as occupy unnecessarily large amounts of memory. The data, a total of 2 gigabytes, was collected over the months of November and December 2018. The resulting corpus contains 491,517 tweets and a total of 13,511,098 words.

It must be mentioned that the program written for data collection accepts quoted tweets; such tweets consist of an existing tweet and a user's response to it. When a quoted tweet is collected by the script, only the response part is stored, which ensures that no duplicates are produced. While the tweet is relevant to the topic, the actual search term may not appear in the reply's text. This is accounted for in the data analysis process, however – for example, collocation analysis requires the presence of the word in the text, and the originally quoted tweets do not appear in the subcorpora.

From each collected tweet, its text, country of origin and hashtags were collected and stored as a separate object. All hashtags and mentions, which often function as proper nouns, were removed from the appended texts. Although hashtags were collected, I use them in this study only for contextual placement; they help in defining the phenomena that affect certain frequencies. In addition, the text was assigned a degree of positivity (positive, neutral, or negative) based on an automatic analysis by NLTK's (Natural Language Toolkit) VADER library, which uses the connotations and syntactic positions of words to conduct sentiment analysis on texts.

Each text was then added to subcorpora by the search terms that appear in it, as well as the country of origin. For example, a tweet from Australia that mentions immigrants is added to both the Australian subcorpus as well as the 'immigrant' subcorpus. As the data collection program did not set restrictions regarding it, tweets from several countries – including non-English-speaking ones – were categorized. Tweets with no location (from users who had not enabled geotagging) composed a surprising 95.84% of the data, and so these tweets were only added to subcorpora matching the appearing search terms. See Figure 1 for the distribution of successfully geotagged tweets:

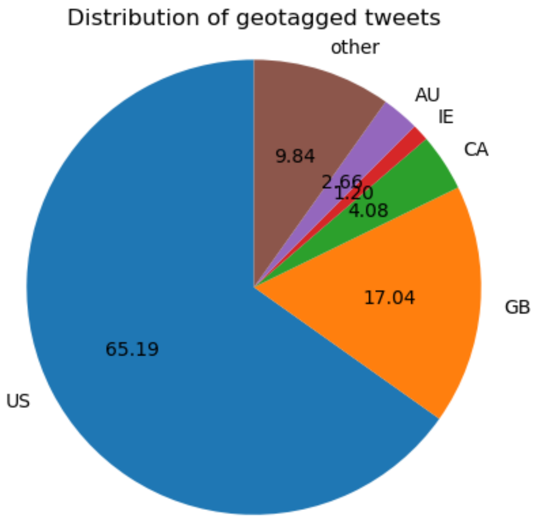

Figure 1. The distribution of successfully geotagged tweets.

Evidently, this gives rise to questions of representativeness. The subcorpora of search terms are substantially larger than the ones for countries of origin. Furthermore – although predictably – the noticeable majority of the data consists of American tweets. These issues could in part be corrected via a larger collective corpus; a larger corpus would achieve sufficient country-specific representativeness, although the balance between different subcorpora would most likely remain uneven due to the sheer amount of American Twitter traffic alone. For the purposes of this study, only countries with upwards of 200 tweets received their own subcorpora. The resulting subcorpora by country are as follows: US (United States of America, 13,315 tweets); GB (Great Britain, 3,480 tweets); CA (Canada, 833 tweets); AU (Australia, 543 tweets) and IE (Ireland, 245 tweets). This is both due to consistency and to ensure that the countries in question are English-speaking. Figure 2 shows the division of subcorpora by search term:

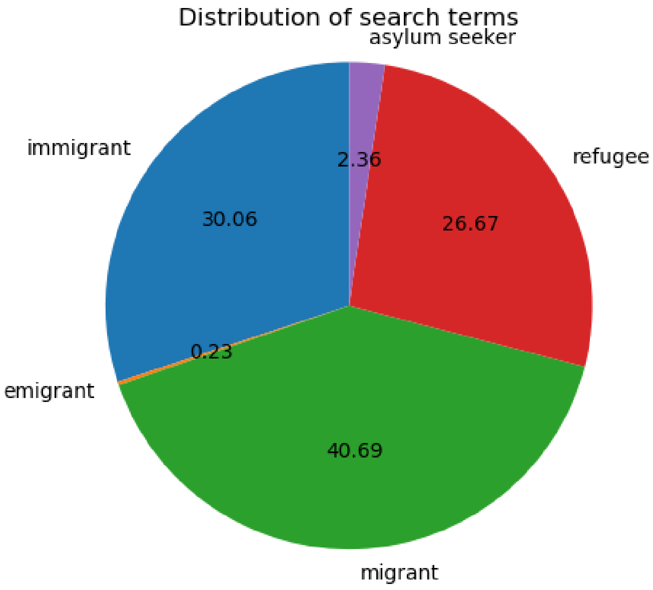

Figure 2. Distribution of search terms.

The distribution displayed in Figure 2 also shows imbalance between categories. The search terms *emigrant* and *asylum seeker* pale in comparison to the much more frequent *immigrant*, *migrant* and *refugee*. All three dominant terms are common in discourse regarding immigrant crises and the migrant caravan approaching the United States border from South – almost ubiquitous topics at the time of data collection and the writing of this paper.

After dividing the data into subcorpora, I created and ran a script to extract the figures for analysis. For each subcorpus, I charted its sentiment distribution, listed its most common hashtags and made a list of its 20 most frequently occurring words (stemmed using NLTK for consistency) to determine keywords. In addition, for the country subcorpora, I determined the most frequent bigrams (excluding stop words); for the search terms, I collected the most frequent collocates. Besides sentiment distribution and hashtags, these figures were attained by creating bigrams of the tokenized, lower-cased texts. All instances of the string ‘u.s.’ were changed to ‘US’ before tokenization, however, to prevent issues with punctuation, ensuring that the correct word was accounted for. Once the data for a subcorpus is collected, a graph (or csv file in the cases of bigrams and hashtags) is produced using the matplotlib library for Python.

Results

The most drastic notion is the most frequent keyword for each search term. The 10 most frequent words in the ‘asylum seeker’ corpus are shown in Figure 3:

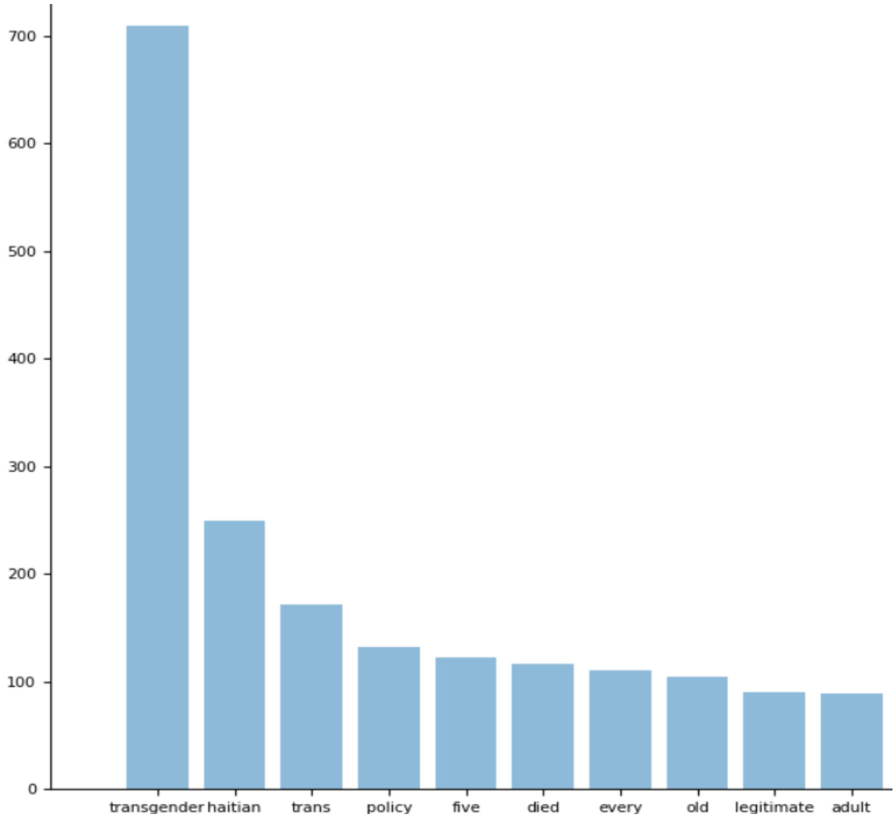

Figure 3. The most common words in the ‘asylum seeker corpus.

As can be seen, the difference in frequency between the most common word and the next is tremendous. As usually with frequency spikes such as this, there is an explanation: a recent widely circulated news report regarding a transgender asylum seeker dying in ICE custody. The case has caused much debate and criticism regarding the ways asylum seekers, especially those in the minorities, are treated. Spikes are present in the other term subcorpora as well, although perhaps not as drastic:

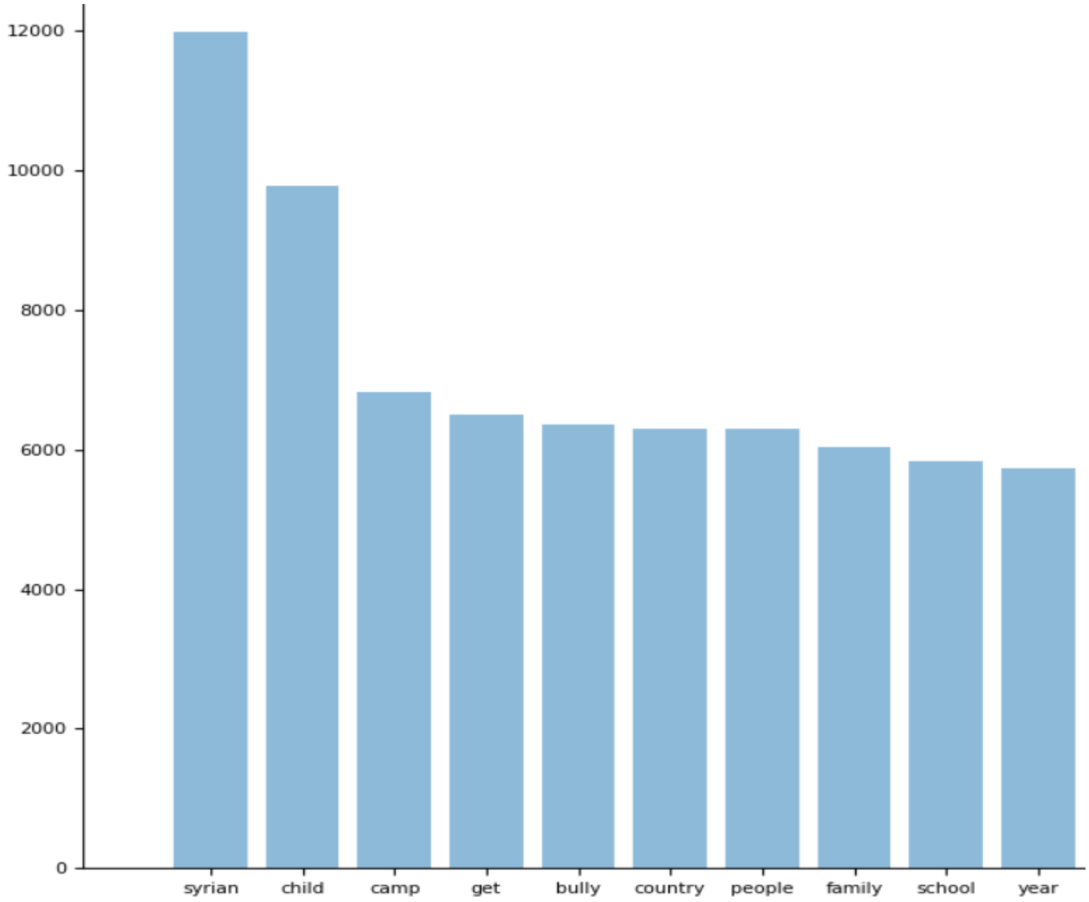

Figure 4. The most frequent words in the ‘refugee’ subcorpus.

Again, the most frequent words are related to news reports and current events: the Syrian refugee crisis (*crisis* is indeed the 20th on the list); family separation, especially highlighting children; and a 15-year-old Syrian refugee, brutally bullied in Northern England. More such tragic and widely discussed events are represented in the top 16 ‘migrant’ subcorpus results (see Figure 5):

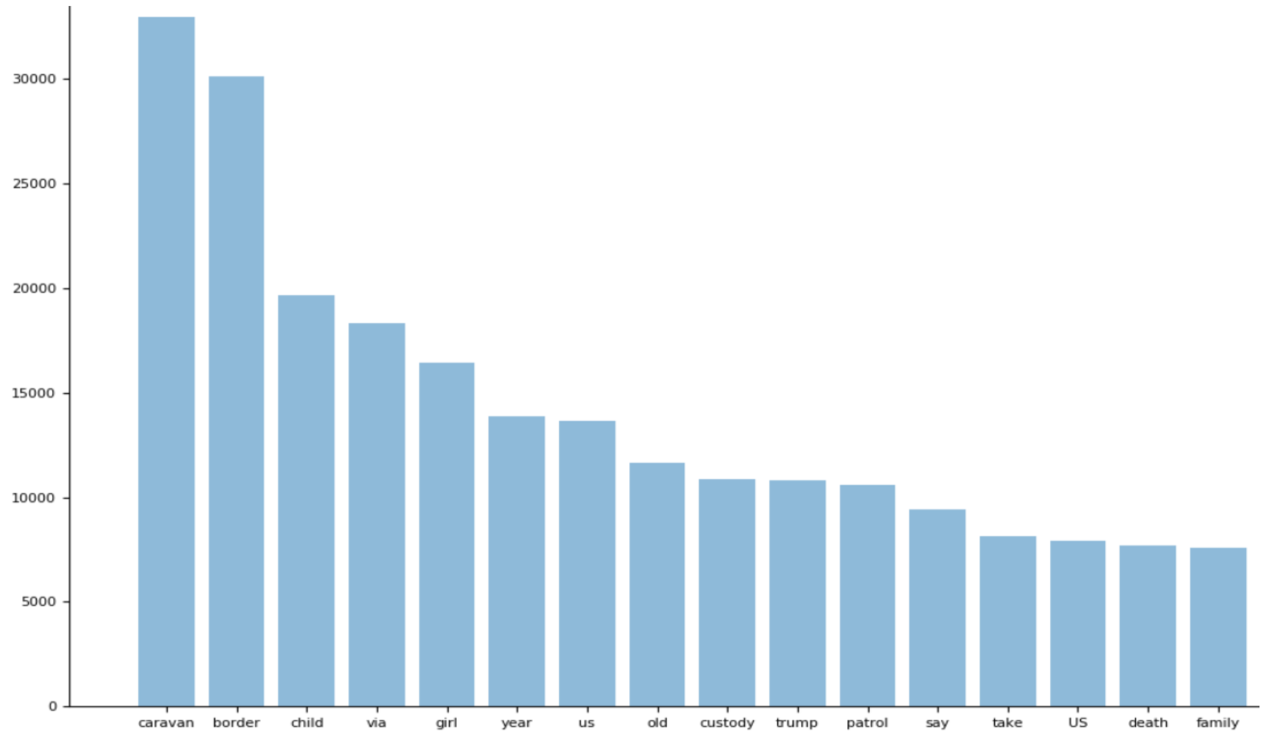

Figure 5. The 16 most frequent non-stopword results in the ‘migrant’ subcorpus.

The migrant caravan approaching the United States border has repeatedly made headlines as politicians have given their views on the issue. With their arrival resulting in violent conflict, and with the topic being central in the media cycle of the largest country in the data, the numbers are much higher than any other word’s. The words *year*, *old*, and *girl* are explained by the terrifying report of a 7-year-old migrant girl dying in the custody of U.S. Border Patrol (7 is also by far the most frequent numeric in the dataset). Notably, as U.S. President Donald Trump has been vocal about the caravan especially, his name appears more than 10,000 times. Similar keywords appear in the ‘immigrant’ subcorpus (Figure 6):

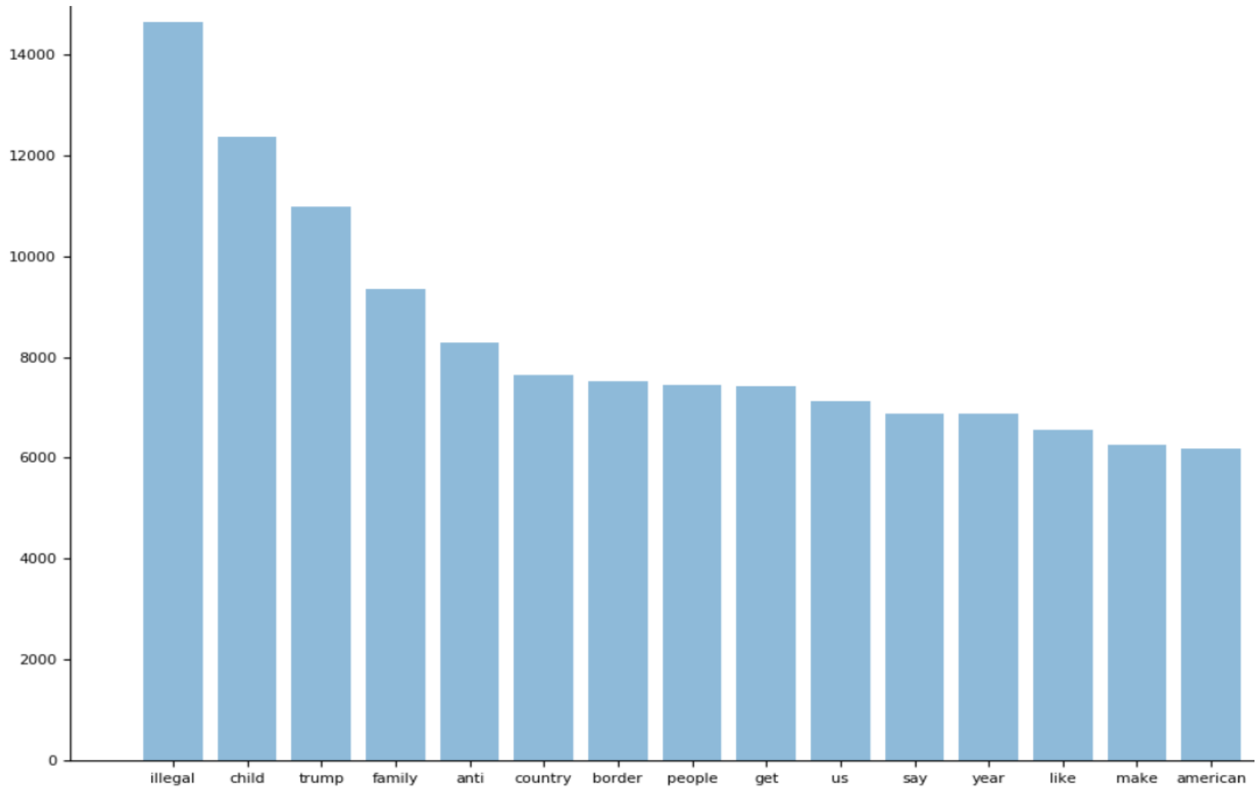

Figure 6. Frequencies in the ‘immigrant’ subcorpus.

As Figure 6 shows, Trump, families, and children appear in this data as well. Many of these terms are, as such, defining keywords of the corpus. The fact that *illegal* is the most frequent word in the subcorpus also echoes Gabrielatos & Baker’s (2008) analysis of this typical, negative collocation. See also Figure 9 for this consistent collocate.

There was noticeable overlap between many search terms and countries of origin in terms of most occurring words; the country-specific subcorpora display pronounced emphasis on certain topics. For example, Figure 7 shows the most frequent words in the GB subcorpus:

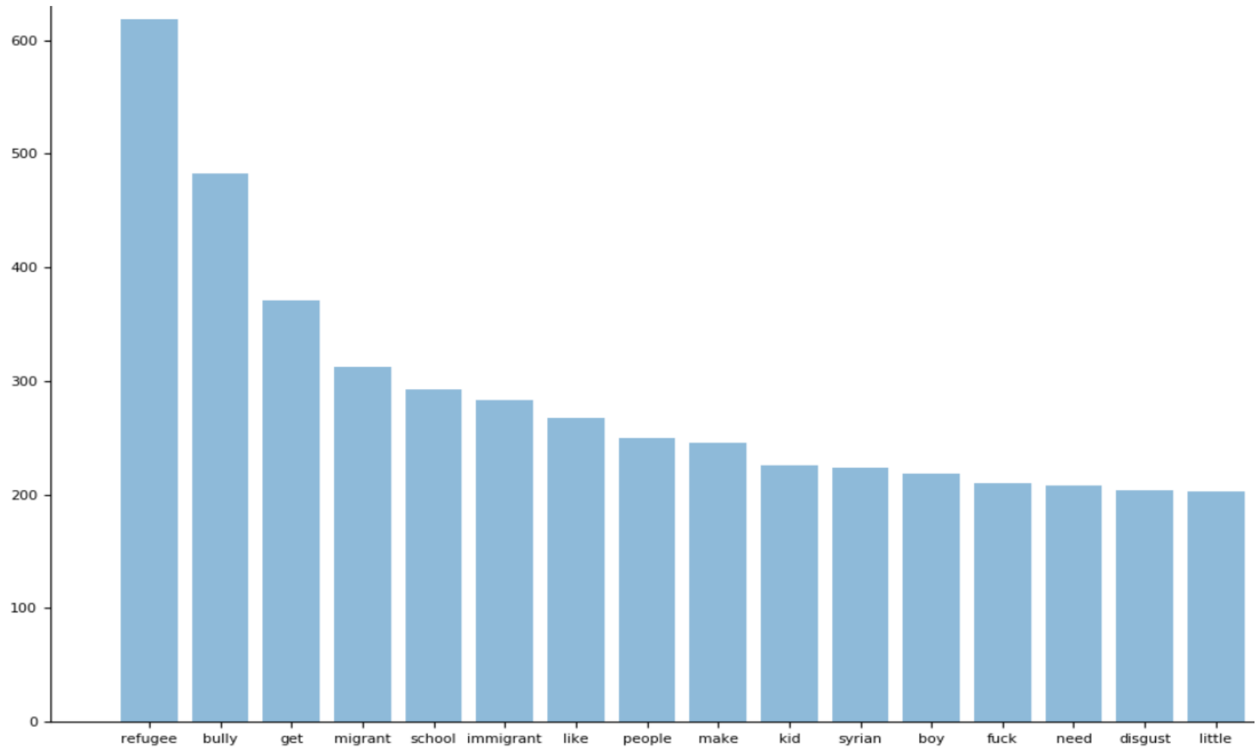

Figure 7. The 16 most frequent words in the GB subcorpus.

Refugee is the most common word, with *bully* as the second most frequent. The top two words and the difference between *refugee* and *migrant* are explained once again by the incident of a refugee boy being bullied at school. Interestingly, this subcorpus is the only one where a swearword appears; the other words present form combinations which function as criticism and expressions of displeasure. The two most common bigrams in the subcorpus were ‘Syrian refugee’ and ‘absolutely disgusting’, which exemplifies this. Figure 8 displays the US subcorpus.

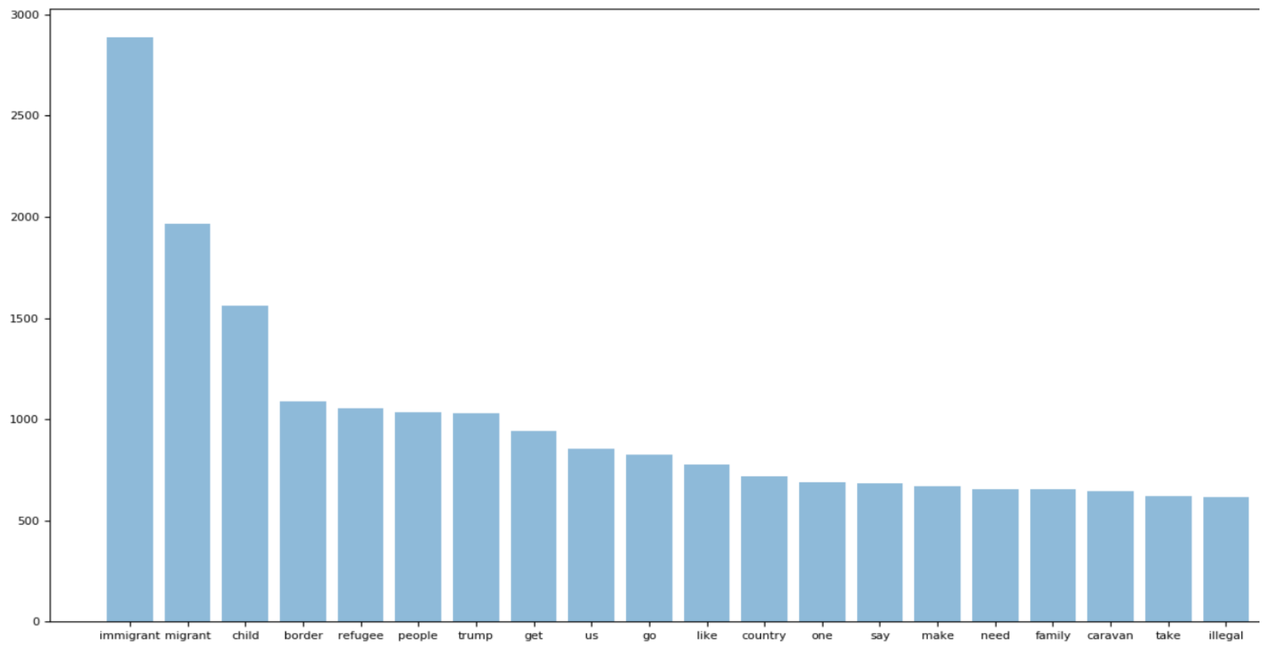

Figure 8. The 20 most frequent words in the US subcorpus.

In the US subcorpus, the most discussed topics (other than the search terms) include again children, the border, Donald Trump and the caravan. Somewhat interestingly, the top results in the CA and AU subcorpora have no references to these phenomena, making them more distant; Ireland, however, shares 9 out of its 20 most frequent words with the GB corpus, with *bully* and *Syrian* among them.

The most frequent collocates for each search term were in line with the keyword results: the most frequent collocate for *asylum seeker*, *transgender*, was more than twice as common as the number two, *Haitian*. Figure 9, where discussion of the 7-year-old girl's death is also visible, restates the inseparability of *illegal* and *immigrant*:

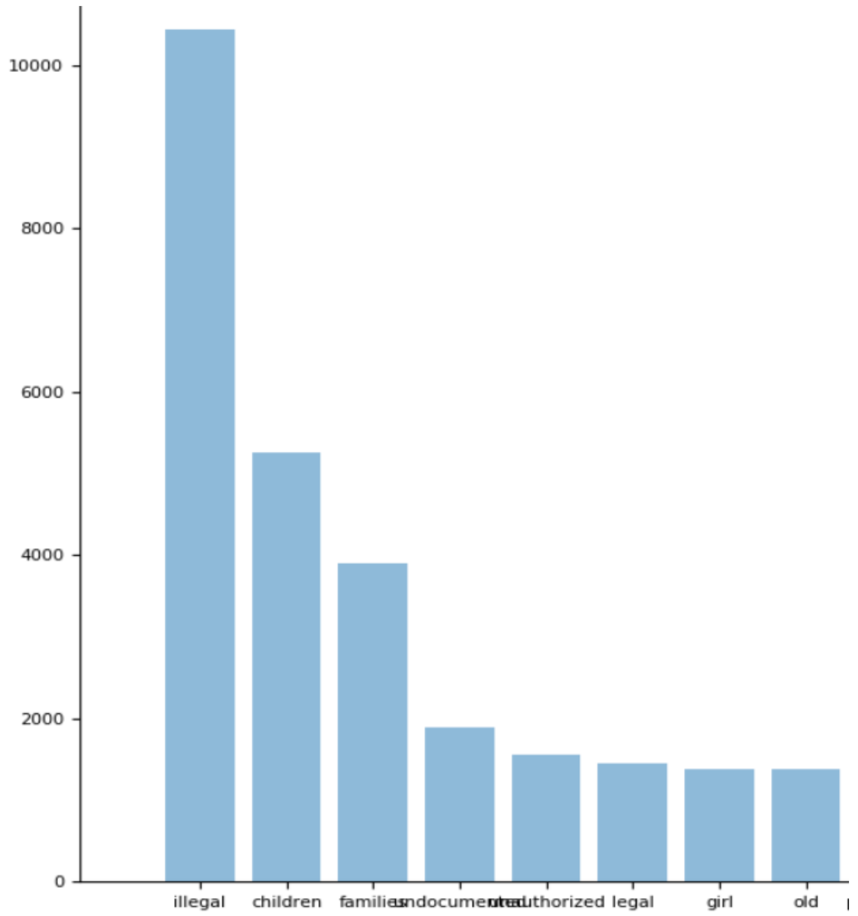

Figure 9. The most frequent collocates of *immigrant – illegal, children, families, undocumented, unauthorized, legal, girl, old*.

Although the news of the child’s death is on the list of (worldwide, but most likely American-dominated) collocates, the case of the bullied Syrian refugee does not appear in the subcorpus for *refugee*; *Syrian* is by far the most common collocate, but as *crisis* is also among the most common collocates for the word, the Syrian refugee crisis seems like the most prominent cause here. The same phenomena can be seen in the most common bigrams for the country-specific subcorpora. The most frequent bigrams in the US subcorpus are ‘migrant caravan,’ ‘illegal immigrant,’ ‘border patrol,’ ‘immigrant children’ and ‘migrant children.’

The sentiment analysis conducted by VADER for each subcorpus gave especially negative results for the GB subcorpus – however, the reason is not the depiction of immigrants or refugees, but their treatment. Most of the negativity is criticism, angry responses to unjust events. This seems to permeate the data to some extent, with tragedies being highlighted more.

Discussion

Although some of the negative constructions of refugees, immigrants and asylum seekers mentioned by Gabrielatos & Baker (2008) persist in this data, most collocations and keywords shown by the results reflect very current phenomena and events, which is indeed what social media affords its users. While there are many negative words, they are typically headlines or complaints and responses aimed at media coverage and politicians. The results show people's reactions to the kinds of experiences immigrants are going through. The topics of discussion seem to be dictated by the media – a thought included in my hypothesis, but actualized in an entirely different way.

This disproves parts of my hypothesis while supporting others: for one, the US data has very few positive instances indeed, but the negativity is not aimed at refugees. While 'illegal immigrant' does not show signs of vanishing, other aspects mentioned by Gabrielatos & Baker (2008), such as impossible compounds ('illegal asylum seeker') or demeaning verb choices, are not present in the Twitter data. The country also seems to matter greatly, which supports the statement made by Coats (2017) as well; local entities and circumstances are vital even when the communication takes place on a global scale, as it does on social media. British and Irish results share many aspects in common, while Donald Trump's presidency seems to be inseparable from the migration-related events, unfortunately marked by numerous deaths, in the United States.

As such, Twitter is an excellent tool for mapping ideologies regarding current events. However, this study also presents various potential issues. Firstly, the immediate nature of Twitter discourse makes streaming – that is, collecting tweets as they are sent in – an extremely synchronic method of research. Diachronic studies may therefore be better off scraping the web for pre-existing texts. Second, the small scale of this study produces problems in representativeness. In order to collect a corpus of respectable size, one would need to pre-plan the variables, requirements and hypotheses much further in advance and gather data for a much longer time – not to mention, prepare enough memory. Third, as Coats (2017) mentions, it is difficult to filter out unwanted sources like automated tweets or news outlets. This may, in part, be remediable via a large enough corpus to ensure representativeness, which is the fourth issue: the small overall size and imbalanced distribution of the corpus means that the data is not very representative. The fifth and final issue is that while using a programming language provides its user with much more flexibility than pre-existing corpus software, errors are always a possibility, and the scripts for gathering and handling data need to be built very carefully and with the hypothesis and synchronicity in mind – in other words, a larger scale benefits this kind of study.

In this paper, I believe I have demonstrated both the advantages and pitfalls of Twitter streaming. Using similar methods, future studies could include analysis of precise attributives of immigrants, refugees and asylum seekers by more specific regions (perhaps by city, or in all English-speaking countries). The very same data I have collected could also be combined in different ways – it would be, for example, possible to study the correlation between sentiment and hashtags. Further qualitative analysis using more precise requirements, such as adjectives within a 3-token distance, would also doubtlessly produce some intriguing results.

Sources

Coats, S. (2017) Gender and lexical type frequencies in Finland Twitter English. In: Hiltunen, T., McVeigh, J. and Säily, T. (eds.), Big and Rich Data in English Corpus Linguistics: Methods and Explorations. (Studies in Variation, Contacts and Change in English 19). http://www.helsinki.fi/varieng/series/volumes/19/

Gabrielatos, C. and Baker, P. (2008) 'Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005', Journal of English Linguistics, 36(1), pp. 5–38. doi: 10.1177/0075424207311247.