Collection of Twitter Corpora for Human and Social Sciences: Sampling Methodology and Evaluation

Louise-Amélie Cougnon[1], Louis de Viron[2,3], Patrick Watrin[3]
[1]MiiL, ILC, UCLouvain, [2]Datatext SRL, [3]Cental, ILC, UCLouvain
louise-amelie.cougnon@uclouvain.be, louis@datatext.eu, patrick.watrin@uclouvain.be

Abstract

The increasing popularity of electronic messages challenges social science researchers, particularly regarding data representativeness and quality. We propose and evaluate a methodology to create a corpus of Twitter data for a given population by sampling the targeted user population. In our case, the population is the groups of citizens, media and politicians using Twitter in Belgium (French and Dutch languages), Norway and France. We present in particular a machine-learning based methodology that enables the population sampling. We also present a methodology to evaluate the representativeness of our corpus compared to the Full Twitter stream of the targeted population.

Keywords: Dynamic Corpus Collection, Open Corpus, Twitter, Social Media, Representativeness, Machine Learning

1. Introduction

Social media has become a powerful driving force for information acquisition and dissemination in just a decade. One of the reasons for the popularity of these media is the ability to receive, create, and share public and private messages in an uninterrupted and low-cost manner (González-Bailón et al., 2014). This growth in the exchange of electronic messages has led to an accumulation of data, which has been called Social Media Big Data, a very rich resource, in particular for scientific research, which also brings about a series of problems that researchers must consider, but which they aren't always aware of (Tromble, Storz and Stockmann, 2017).

The fundamental problem with digital data collection is as much about the quality and volume of data, as it is about the representativeness and veracity of the accessible data (Stieglitz et al., 2018). Researchers are very often required to pay a social network or a private company for (very high priced) exhaustive corpora. They can also use APIs provided for free by social networks, but these are often limited and biased, without knowing exactly what the bias consists of. This causes problems for researchers: which APIs to use? What reliability, inference, or replicability of conclusions are possible, on the basis of corpora whose biases are unknown (Neuhaus and Webmoor, 2012; Tromble, Storz and Stockmann, 2017)? This unfamiliarity with the material creates many problems for research based on corpora from social networks, including a lack of clarity in the results, and a lack of replicability and comparability of the results across researchers (Stieglitz et al., 2018).

The social network Twitter, founded in 2006, participates actively to the growing interest it induces among researchers, as much as to the problems related to its collection. This social network is proven to be very useful for studying a wide range of phenomena, from the dissemination and veracity of information to trends in public opinion, as well as peaks in collective attention (Honey and Herring, 2009; Boyd, Golder and Lotan, 2010; Harrigan, Achananuparp and Lim, 2012; González-Bailón et al., 2014; Qin, Cai, and Wangchen, 2015; Mitra, Wright, and Gilbert, 2017; Cougnon and de Viron, 2021). One reason for this is the platform's prominence as a media of public communication and its importance in political and media discussions. Twitter's popularity in research is also due to the fact that its accounts are public by default and that the platform makes numerous APIs available, which allow researchers to build up corpora without much computer knowledge (González-Bailón et al., 2014).

However, corpus collection from Twitter is not exempt from the aforementioned problem of biases (Hino and Fahey, 2019), starting with the population that uses the platform actively or passively: only "22% of American adults who use Twitter are representative of the broader population" (Wojcik and Hughes, 2019: 2). The overrepresented population is the youngest, the Democrats, the most educated, and the most socioeconomically advantaged. It is also at the level of the data collection that biases are reinforced: Twitter itself imposes limits to the collection and is often unclear about the internal processes used to select and filter the data available through its APIs. This vagueness makes it impossible to infer scientific findings to the general Twitter population (Rafail, 2018).

Among the Twitter APIs made freely available, the *Streaming API* provided, for a few years, a stream of tweets based on search parameters (e.g. specific hashtags), but only offered real-time data, without the ability to search for historical tweets. The resulting corpus was limited to roughly 1% of the real Twitter traffic results and the selection process of these 1% tweets was again unclear. Twitter recently shut down public access to the Streaming API, making obsolete many of the methods used by other researchers to collect data. Out of the APIs still available, the *Search API* provides access to a sample of tweets dating back 7 days. This method also provides a biased sample (Tromble, Storz and Stockmann, 2017): Twitter itself acknowledges that "the standard search API is focused on relevance, not completeness" (González-Bailón et al., 2014). Finally, the *User Timeline API*[1], which we used for our collection, provides access to posts from specific accounts on Twitter and can provide with the history of these accounts. It is the only free API that provides complete data and not a filtered subset. Its limitation is that

[1] https://api.twitter.com/1.1/statuses/user_timeline.json

it only retrieves the most recent 3200 tweets from each account. It therefore limits the representativeness of the accounts, especially when they are pro-active (when they publish tens or hundreds of tweets per month). It is precisely in order to counter this problem that the methodology described in this paper has been designed.

2. A Project Designed for Social Sciences

2.1 The Representativeness of the Corpus

The question of corpus representativeness in linguistics and literature is not new and is not limited to digital data (Arbach and Ali, 2013). It is indeed always essential to be able to determine a population that can be represented by its language, its discourse, or its content production: "A corpus must be 'representative' in order to be appropriately used as the basis for generalizations" (Biber, 1993: 243). In this context, as we have seen, digital corpora are rich resources and, most of the time, more easily stored and analyzed by automatic tools. However, the criteria of representativeness are questionable and depend on the research purposes as well as the media that is studied. "Typically, researchers focus on sample size as the most important consideration in achieving representativeness" (Biber, 1993: 243). However, we now know that this is not the only criterion to be considered in the construction of a corpus. The notion of representativeness also includes an objective of variety and veracity in relation to the population that is being studied (Stieglitz et al., 2018). Thus, researchers attempt to cover language practices in their homogeneity as well as in their heterogeneity. For this reason, we are more in line with a desire to represent the variety of profiles rather than attempting to achieve exhaustiveness in Twitter discourse. To do so, our project is focused on the representation of geographical regions within Twitter, units of analysis which, as we will see, have their own particularities, but which are homogeneous enough for our methodology.

2.2 Towards an Open Corpus of Tweets per Region

The goal of our project is to dynamically collect a sample of tweets and allow researchers to obtain easy access to this sample without prior knowledge of the research topics under consideration, and without being limited by a filter whose biases the researchers are not aware of, or by the obligation to focus exclusively on real-time data. The constructed corpora will be made available to researchers through a consultation interface developed specifically for human science research.

The methodology created is designed for creating an open corpus for a given geographical region. Currently, the regions covered are French-speaking Belgium (BE-FR), Dutch-speaking Belgium (BE-NL), France (FR-FR) and Norway (NO-NO). However, the methodology is straightforwardly transposable to new regions. For each region, three sub-corpora are constituted: a media corpus, a political corpus and a citizen corpus. While the first two are derived from a manually constructed list of accounts that are more or less easy to build up, the citizen corpus is created from an automatically generated and validated list of accounts.

The idea of building "open corpora" or "monitor corpora" that is to say "corpora that never stop growing" Habert (2000: 12-13), was born, in opposition to the so-called closed corpora, developed once and for all, from the desire to continuously capture data in order to make possible the study of the evolution of certain phenomena, originally linguistic (Renouf, 1993). However, open corpora require that we define what exactly they focus on and from when onwards. Our project of collecting open corpora from Twitter requires the selection of a panel of "citizen", "media" and "politician" Twitter accounts from a given region. The selection of the last two corpora does not require a very complex methodology of representativeness, given that the media and political accounts in the concerned areas are not extremely numerous. However, the "citizen" accounts requires a real thought on the sampling process and this methodology is described in detail in this paper.

3. Corpus Collection Methodology

The methodology for collecting and evaluating the Twitter data corpus is designed to meet the various challenges described in section 1 and the various expectations described in section 2.

3.1 Accounts Selection

The approach adopted for the selection of citizen accounts is semi-automatic, and is based on a supervised machine learning algorithm, aiming at selecting the accounts that belong to the targeted population among the followers of popular accounts (henceforth "hub") within the population. We start by manually building a list of about 500 hub accounts[2] within the target region. This manually built "hub list" of profiles covers various domains in order to target varieties of discourses and opinions: it includes artists (singer, painter, writer...), sportsmen, influencers (make-up, politics, history, fashion...), large commercial brands, politicians, media, public and private institutions and NGOs. The lists are built in collaboration with members from the targeted region to ensure that they offer the broadest possible representation of the sensitivities present in the region. While this methodology does not offer representativeness in the statistical sense, it does allow researchers to clearly define what the corpus represents and, more importantly, what it does not.

We then extract the followers of the 500 hub accounts using the *Twitter Followers API*[3], resulting in a large list of followers. These accounts must be filtered in order to meet the definition of the corpus: they should be speakers of the target language living or coming from the target region (BE-FR, BE-NL, FR-FR and NO-NO).

3.2 Accounts Validation

The validation of accounts among the list of followers is based on a supervised machine learning model that relies on a set of features described in Table 1.

[2] The number limit of 500 hub accounts was arbitrarily defined because it already represents a very large number of followers. As an example, for the French speaking Belgian population, the 500 hub accounts represented 18M followers which is larger than the target population itself.

[3] https://api.twitter.com/1.1/followers/ids.json

geo_in_location	Number of local geographic entities in the account geolocation (string matching based on Wikidata (Vrandečić and Krötzsch, 2014))
out_geo_in_location	Number of non-local geographic entities in the account geolocation (words outside the Wikidata entities list)
geo_in_tweets	Number of local geographic entities in the last 200 tweets (string matching)
lang_in_tweets	Percentage of tweets written in the target language among the last 200 tweets
tld_in_tweets	Number of URLs from the country-code TLD[4] in the last 200 tweets
seed_followed_accounts	Number of accounts followed among hub accounts

Table 1: Features used in the account validation model

The features generation algorithm is generic and is used for each region. However, it requires the adaptation of resources to cover the geographical specificities of the targeted region (list of geographical entities, list of official languages...).

We test our approach by training a first model for the French-speaking part of Belgium. A training set of 500 accounts randomly retrieved from the list of followers is constituted and manually tagged. The human annotator indicates as "TRUE" the accounts (1) that belong to citizens of the target region, and (2) whose tweets are written in the target language. When both conditions are not met or not predictable, the annotator tags it as "FALSE". A classifier is trained on these annotations in order to generalize the annotation task on all the extracted followers, and thus constitutes the final list of followed accounts.

For the training, we use the XGBoost algorithm, an optimized open source implementation of the gradient boosting trees algorithm (Chen and Guestrin, 2016). The model is trained in a 5-sample cross-validation, with an average accuracy of 90%.

The XGBoost algorithm helps us to understand the results through an analysis of the features importance within the model, in which each feature is associated with a weight (Table 2).

Feature	Weight
geo_in_location	0.28
tld_in_tweets	0.28
out_geo_in_location	0.17
lang_in_tweets	0.14
seed_followed_accounts	0.07
geo_in_tweets	0.06

Table 2: Feature weights of in the BE-FR model

We observe that the features are globally well balanced, and are all contributing to the model, even if the geography of the account and the shared "national" URLs are the most important features in the model.

Since the feature generation algorithm is common to all regions, it is necessary to annotate a separate training set and to train a proper model per region. Indeed, the importance of the features varies depending on the particularities of each region. For example, the fact of speaking Norwegian is very discriminating in defining whether an account is Norwegian, which is not the case for French in France or Belgium, whereas French is spoken in 88 countries all over the world[5].

The analysis of the features in the Norwegian model confirms this hypothesis (Table 3): the language of the tweets is the most discriminating feature, and has a much higher weight (0.44) than in the French-speaking Belgian model (0.14).

Feature	Weight
lang_in_tweets	0.44
geo_in_location	0.20
out_geo_in_location	0.11
tld_in_tweets	0.10
geo_in_tweets	0.09
seed_followed_accounts	0.06

Table 3: Feature weights of in the NO-NO model

The models are then tested on an independent set of 200 accounts per region, annotated in the same way as the training set. The evaluation of these models (Table 4) shows good results, with an F-Measure close to 90% for all models, but above all, very good precision.

Region	Precision	Recall	F1-score
BE-FR	0.90	0.90	0.90
BE-NL	0.96	0.90	0.93
NO-NO	0.99	0.76	0.86
FR-FR	0.93	0.88	0.90

Table 4: Evaluation metrics of the account validation model

As we have seen, our approach is semi-automatic, and replicable to any region of the world. To do this, it is necessary to select a list of hub accounts specific to the region concerned, collect their followers and train a new validation model on a sample of these followers. The model can be applied on the full list of followers in order to select the final list of accounts.

3.3 Tweet enrichment and indexing

The set of accounts is the basis for the creation of the corpus. The tweets produced by these accounts are downloaded respecting the limitations imposed by Twitter: (1) Historical corpus: retrieving the last 3,200 tweets (REST API, cf. part 1) produced by these accounts, which in a large majority (94%) represents all the tweets produced by these accounts; (2) Monitor corpus: collecting daily, starting from the date mentioned in Table 5, the new tweets

[4] Country Top-level domain (i.e. suffix) of an URL (e.g.: *be* for Belgium, *no* for Norway…)

[5] https://www.francophonie.org/88-etats-et-gouvernements-125

produced by these accounts continuously in order to enrich the database.

The tweets are then indexed in Elasticsearch[6], a search engine that allows users to query the textual content of the corpus of tweets and its metadata and to create custom corpora for specific researches. On top of the indexation, an NLP enrichment is applied by using spaCy (Hannibal and Montani, 2017), a Python library that provides pre-trained language models for the languages that are in the scope of the project[7]. The following information is indexed: (a) lemmas (nouns, verbs and adjectives); (b) multi-words units, based on predefined syntactic patterns and applied based on spaCy's pos-tagging; (c) named entities (places, people and organisations).

The current state of the open corpora constituted so far is described in Table 5, while Figure 1 illustrates our complete process in a synthetic way.

	BE-FR	**BE-NL**	**NO-NO**	**FR-FR**
Hub accounts	542	600	825	775
Followers	18 M	3.6 M	2.9 M	67 M
Selected accounts	187 K	277 K	240 K	947 K
Start collection date	Sep. 20	May 21	Jul. 21	Nov. 21
Tweets	115 M	143 M	100 M	**320 M**

Table 5: Statistics related to the *Twitter Open Corpus* per region (January 2022)

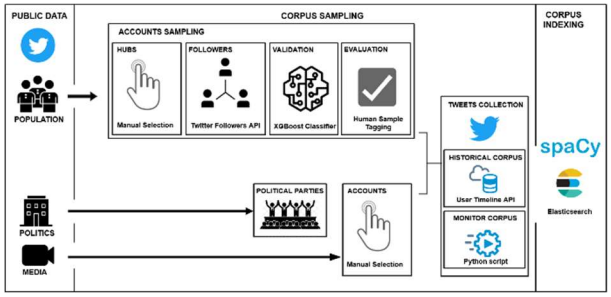

Figure 1: Summary of the Corpus construction process

4. Evaluation of the Representativeness

The evaluation of our method is crucial as it addresses a prevalent issue in research on Twitter corpora (Tromble, Storz and Stockmann, 2017). However, as we mentioned earlier, the population that uses Twitter does not coincide with the general population of a given region. Therefore, the definition of its representativeness cannot be based on socio-demographic measures of the population: the question is rather to know if our results are representative of the general results for all Twitter users of a given region, by ensuring in particular that our corpus is large and varied enough. To achieve this, we base our assessment on Hino and Fahey (2019), who compare the representativeness of a sample of tweets to the general Twitter stream.

To perform this analysis, we focus on tweets from Belgium by comparing our Belgian corpus (BE-FR & BE-NL) with the Twitter stream filtered on tweets in French sent from Belgium (geolocation-filtered Twitter stream, with a radius of 150 km around the geographical center of Belgium[8]).

On top of this analysis, we have to validate that the filtered Twitter stream is itself a representative sample of the BE-FR Twitter stream. To that purpose, we select one topic highly specific to Belgium: *Codeco* (the Belgian concertation committee in the fight against Covid-19). We assume that the topic is so specific to Belgium that it will rarely occur in a tweet outside of the country. Figure 2 presents the distribution of this topic over a 15-day period (15 December 2020 - 31 December 2021) in both corpora.

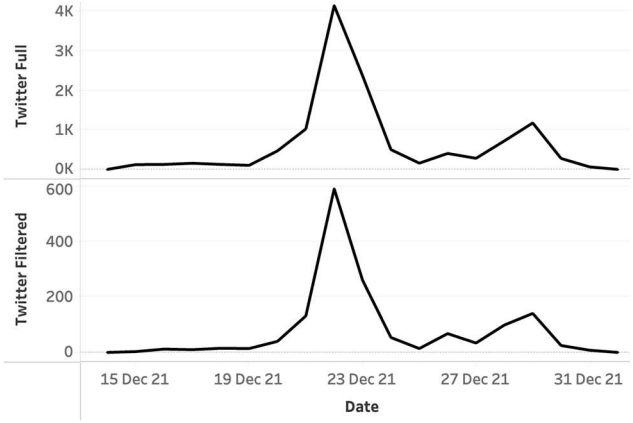

Figure 2: Distribution of "Codeco" in subcorpora over time

The visual similarity in the frequency distributions is confirmed by an R-squared score of 0.992. This first analysis therefore allows us to show the frequency representativeness (distribution of tweets over time) of the Twitter Filtered corpus. This first result could lead us to question our methodology: if the geolocation information is sufficient, our selection process could appear useless. However, Figure 3 shows that our solution collects 2.5 times more tweets than the geolocation filter for selected query. The reason is that Twitter limits the filtered feed to tweets whose users are explicitly geolocated in Belgium (whose users have given explicitly the information). This finding confirms the interest and the originality of our approach that relies on more implicit features to select the accounts that will feed the corpus.

[6] https://www.elastic.co/

[7] One must note that the SpaCy models are not trained on Twitter data. This can lead to annotation errors. Using Twitter-specific models is under investigation.

[8] https://api.twitter.com/1.1/search/tweets.json?geocode=50.64012,4.66668,150km

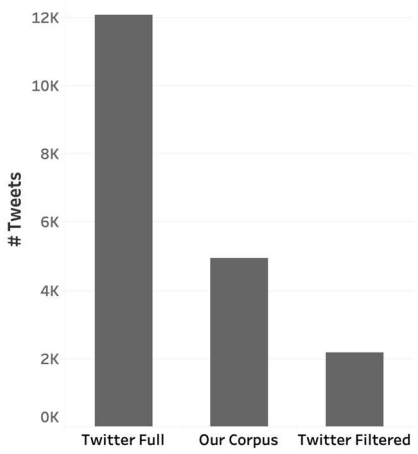

Figure 3: Number of tweets about "Codeco" / "Overlegcomité" in the corpora (Dec 15-31, 2021)

In addition to the distribution of tweets, Hino and Fahey (2019) also study the distribution of topics to assess the semantic representativeness of their corpus. To estimate this diversity, they suggest using *latent Dirichlet allocation* (LDA: Blei et al., 2003), which identifies the topics present within each tweet and then enables daily comparison of the distribution of these topics between two corpora. The underlying assumption is that a high similarity in the distribution of topics should reflect a similarity in content and allow us to claim a certain semantic representativeness.

Following their methodology, we train a LDA model on an independent corpus of tweets covering the period from 16 to 31 December[9]. Then we fold each tweet of both samples into a vector of 100 topics, which we subsequently combine to obtain one vector per day and per corpus (the vector of a day is obtained by averaging the vectors of topics of this day). The distribution similarity is then estimated by calculating the cosine similarity.

Figure 4 shows a high cosine similarity between the vectors of the two corpora (always above 80 %), showing a similarity of content..

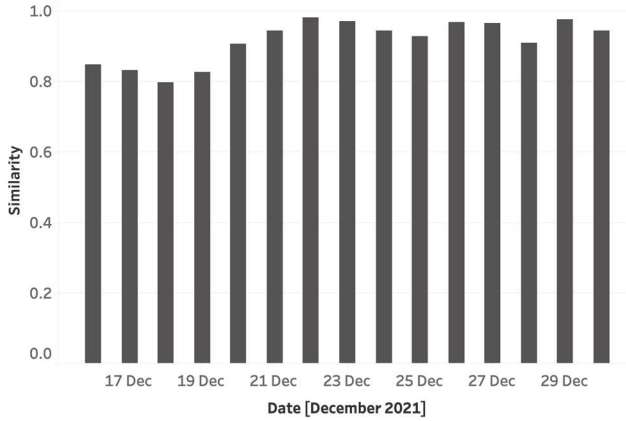

Figure 4: Cosine Similarity between average daily LDA vectors (Full Twitter vs. Twitter Filtered)

These two analyses (frequency and topics distributions) allow us to conclude that we can use the Twitter filtered stream as the reference corpus (Ground Truth) for analyzing the representativeness of our own corpus.

To perform this evaluation, we collect two corpora based on the 50 most popular hashtags in Belgium during one week (January 2nd - January 9th, 2022)[10]. The first is obtained by using the Twitter filtered stream (Ground Truth) and the second is obtained by querying our own corpus (BE-FR & BE-NL). We then apply the same two methods as above to evaluate the tweets and topics distribution between both corpora.

To evaluate the frequency representativeness, we first normalize the tweets distribution (number of tweets for each day divided by the total number of tweets in the subcorpus). Then we plot the distribution (Figure 5) and analyze the correlation of both dimensions. We obtain a R-squared score of 0.903, which expresses a strong correlation.

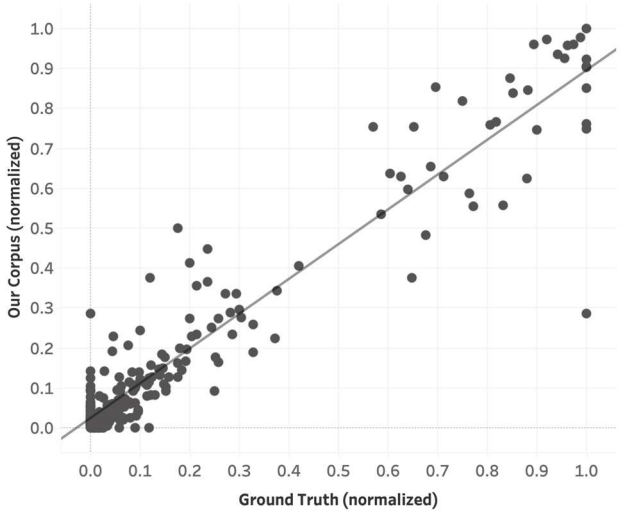

Figure 5: Normalized distribution of the top 50 hashtags over time in both corpora

To evaluate the semantic representativeness, we train a LDA model in the same way as above. Then we fold the tweets of both corpora in the model in order to build the daily vectors. Figure 6 shows a high cosine similarity between the vectors of the two corpora (always above 75 %), meaning that their content is comparable.

[9] The choice of an independent corpus is motivated by the need to have access to a large number of tweets in order to obtain a more expressive model.

[10] We obtain these hashtags thanks to the Getdaytrends tool. https://getdaytrends.com/belgium/

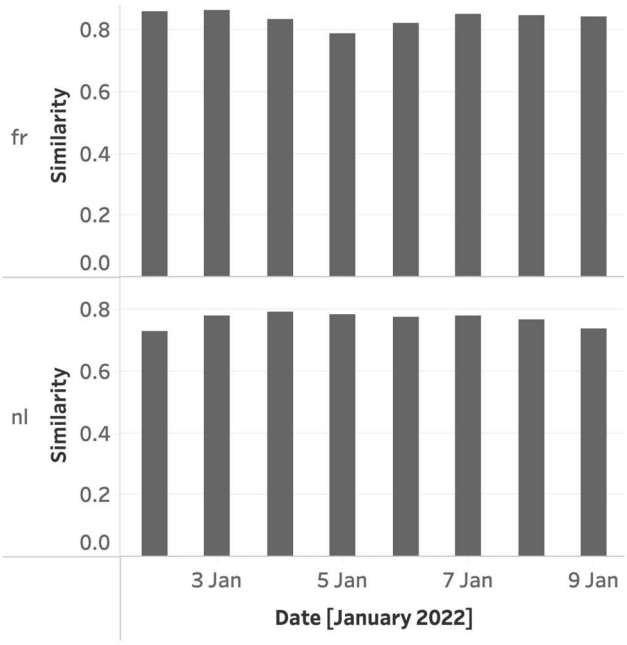

Figure 6: Cosine Similarity between average daily LDA vectors (split by corpus)

Based on the results of these two analyses, we can conclude that our corpus is representative of the production of tweets in Belgium. This result validates our method, which, in addition to being more productive than a geolocation-based filtering, is also easily adaptable to other regions.

5. Conclusions

We defined a methodology for the collection of open Twitter corpora that are representative of regional communities on the social network. This methodology is efficient and replicable to new regions with little effort. However, this methodology has its own biases. Its strength is that it can define them: (1) The use of the REST API to collect the last 3,200 tweets from each account imposes a very diverse history for the whole corpus (some tweets going as far as 10 years backwards, others only a few months). Thus, our corpus can only be considered complete from a given time, when the collection is done on a daily basis (Table 5). (2) The choice of hub accounts has an impact on the nature of the tweets sent by their followers: it is therefore important to always know the nature of these accounts, especially in the variability that they may have in each specific region (religious communities, presence of political parties, engagement of influencers, number of private accounts...).

With these biases in mind, we therefore propose a method that enables scholars to draw inferences from the data with caution and escape from the vicious circle of "we simply do not know what we do not know" which is very common in Twitter-based research (Tromble, Storz and Stockmann, 2017).

6. Acknowledgements

Since 2018, this research has been made possible thanks to the intervention of 3 different grants: (1) the Opinio project[11], which was partly funded by the Fédération Wallonie-Bruxelles (a French-speaking Belgian institution); (2) the FuturICT 2.0 project[12], which was funded by the FLAG-ERA Joint Transnational Call; and (3) the 2O2CM project[13], which was funded by the Solstice JPI Climate Call.

7. Bibliographical References

Arbach, N. and Ali, S. (2013). Aspects théoriques et méthodologiques de la représentativité des corpus. *Corela*, HS-13.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4): 243–257.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Boyd, D., Golder, S.A., Lotan, G. (2010). Tweet, Tweet, Retweet: conversational aspects of retweeting on Twitter. *HICSS-43 IEEE*, Kauai, HI.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 785–794, San Francisco California, USA.

Cougnon, L.-A. and de Viron, L. (2021). Covid-19 et communication de crise. Focus linguistique sur les tweets francophones de Belgique. In S. Cotelli Kureth, M. Dubois, & A. Kamber (Eds), *La linguistique appliquée à l'ère digitale*, Bulletin suisse de linguistique appliquée, Bulletin suisse de linguistique appliquée, No spécial,1, pp. 103--128.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J. and Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, 38: 16–27.

Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ?. In Bilger, M. (Ed.), *Linguistique sur corpus. Études et réflexions*, 31, pp. 11--58.

Harrigan, N., Achananuparp, P., Lim, E.-P. (2012). Influentials, novelty, and social con-tagion: the viral power of average friends, close communities, and old news. *Social Networks,* 34: 470–480.

Hino, A., and Fahey, R.A. (2019). Representing the Twittersphere: Archiving a representative sample of Twitter data under resource constraints. *International Journal of Information Management*, 48: 175–184.

Honnibal, M., and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Honey, C., Herring, S.C. (2009). Beyond microblogging: conversation and collaboration via Twitter, system sciences. In *Proceedings of the 42nd Hawaii International Conference on HICSS '09*, pages 1–10, Big Island, HI.

[11] https://uclouvain.be/fr/instituts-recherche/ilc/miil/opinio.html

[12] https://futurict2.eu/

[13] https://change4climate.eu/

Mitra, T., Wright, G. and Gilbert, E. (2017). Credibility and the Dynamics of Collective Attention. In *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW).

Rafail, P. (2018) Nonprobability sampling and twitter: Strategies for semibounded and bounded populations. *Social Science Computer Review*, 36(2): 195–211.

Stieglitz, S., Mirbabaie, M., Ross, B., and Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39: 156–168.

Tromble, R., Storz, A., and Stockmann, D. (2017). We don't know what we don't know: When and how the use of twitter's public APIs biases scientific inference. *Social Science Research Network*, n°3079927.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57 (10): 78–85.

Qin, Z., Cai, J., and Wangchen, H. Z. (2015). How rumors spread and stop over social media: A multi-layered communication model and empirical analysis. *Communications of the Association for Information Systems*, 36: 369 – 391.

Wojcik, S., and Hughes, A. (2019). Sizing Up Twitter Users. *Pew Research Center*. Available at: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/