Using Elasticsearch for Linguistic Analysis of Tweets in Time and Space

Adrien Barbaresi[◦], **Antonio Ruiz Tinoco**[•]
[◦] Academy Corpora, Austrian Academy of Sciences
Sonnenfelsgasse 19, 1010 Vienna
adrien.barbaresi@oeaw.ac.at
[•] Faculty of Foreign Studies, Sophia University
7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554
a-ruiz@sophia.ac.jp

Abstract

The collection and analysis of microtexts is both straightforward from a computational viewpoint and complex in a scientific perspective, they often feature non-standard data and are accompanied by a profusion of metadata. We address corpus construction and visualization issues in order to study spontaneous speech and variation through short messages. To this end, we introduce an experimental setting based on a generic NoSQL database (Elasticsearch) and its front-end (Kibana). We focus on Spanish and German and present concrete examples of faceted searches on short messages coming from the Twitter platform. The results are discussed with a particular emphasis on the impact of querying and visualization techniques first for longitudinal studies in the course of time and second for results aggregated in a spatial perspective.

Keywords: Social media, Twitter, NoSQL databases, Time series, Spatial data mining

1. Introduction

Web documents in general and computer-mediated communication in particular put linguists in front of new challenges. As empirical evidence is now widely used in linguistics, the Web turns to a major source, which changes the way research is conducted. That is why Leech (2006) speaks of linguists as "inhabiting an expanding universe". The shift towards web texts took place around the year 2000, while focus on computer-mediated communication (CMC) is closely related to the emergence of social networks in the course of the 2000s. The sharp decrease in publication of work documenting web corpus construction and use after 2008 hints at a further development towards a particular focus on CMC data, which can be explained by the growing popularity of short message services (also called microblogs) in the form of social networks. The latter often provide larger number of identifiable users as well as clear data on network activity and range, for example by tracking followers. Because of the profusion of data and metadata, schemes and methods are needed to live up to the potential of these sources. It is indeed widely acknowledged that social and humanities scientists need to acquire and develop the skills to do data analysis and experiment with the visualization tools necessary to manage and interpret big data (Manovich, 2011).

At the time of Web 2.0 and APIs, a URL is merely an access point, the resulting website is often tailored not only to what the user has just typed in or said, but to a whole navigation and engagement history. Furthermore, the actual text may become accessory compared to the profusion of metadata which encase it. The clearly structured machine-readable formats combine up to hundreds of different fields ranging from previously studied information such as the user name or the date of posting to new, typical information such as the number of followers or the time zone of the user. That being said, not all information is as relevant or objective as one would expect, so that it has to be refined before reaching any conclusion.

Following from a first approach to download and indexing (Barbaresi, 2016), the present article tries to document both research methodology and research objects in an articulated fashion. Beyond the selection of a medium and the collection and preprocessing of messages, our work follows from two distinct procedures for handling social media information (Tsou and Leitner, 2013): "transform the original information formats into analytic forms" and "explore multiple analytical methods", to try to answer the following research question: "What should a comprehensive space-time analysis framework look like for different scenarios or questions?". Consequently, the visualizations introduced here are an attempt to find "the best combination of analytical methods to analyze social media information in depth from both temporal and spatial aspects" (Tsou and Leitner, 2013). We address corpus construction and visualization issues in order to study spontaneous speech and variation through short messages. We present and discuss an experimental setting to observe language through tweets, with a particular emphasis on the impact of visualization techniques on time series (messages seen in the course of time) and space (aggregated projections on maps of geolocated messages).

2. Experimental setting

2.1. Twitter as a source

The interest in Twitter is generally considered to reside in the immediacy of the information presented, the volume and variability of the data contained, and the presence of geolocated messages (Krishnamurthy et al., 2008). Other social networks do not deliver the same amount of text, especially for languages other than English, for example on Reddit (Barbaresi, 2015). Most importantly, they cannot be deemed as stable in time in terms of popularity and API access (Barbaresi, 2013). Nevertheless, because of access restrictions – mostly mechanical constraints on

the free streaming access point – it is not possible to retrieve all tweets one would need. It is indeed necessary to enter search terms or a geographic window, which may greatly affect results especially for highly frequent keywords (Ljubešic et al., 2014). The API is supposed to deliver a random sample representing a volume of about 1% of all tweets when used with a worldwide geographic window.

Since August 2009, Twitter has allowed tweets to include geographic metadata, which are considered to be a valuable source for performing linguistic studies with a high level of granularity, although the geolocation is not always accurate. Currently, the public Twitter Application Programming Interfaces (APIs) can provide five types of geocoding sources: geo-tagged coordinates, place check-in location (by way of a bounding box), user profile location, time zones, and texts containing explicit or implicit locational information (Tsou et al., 2017). The geo-tagged coordinates are the most frequently used type of information. The geolocalized messages can conveniently be projected on maps, which is highly relevant for various research fields, for instance variation studies in linguistics. That being said, it is important to note that geolocated tweets are a small minority, with estimates as low as 2% of all tweets (Leetaru et al., 2013). However, even the bounding boxes used to retrieve tweets do not always function as expected due to systematic errors (Tsou et al., 2017). Additionally, the geolocation results of profile locations are not a useful proxy for device locations, and the success at being able to place users within a geographic region varies with the peculiarities of the region (Graham et al., 2014).

From the point of view of corpus and computational linguistics, tweets are both highly relevant and difficult to process. Short messages published on social networks constitute a "frontier" area due to their dissimilarity with existing corpora (Lui and Baldwin, 2014), most notably with reference corpora. Some metadata are more useful than others, and some languages fit more easily into the allocated space than others (previously 140 characters, now 280 for most languages). Regarding the content itself, the quantity of information in general or the relevance for linguistic studies in particular may vary greatly. In spite of the restrictions on the API, the delivered volume may already be sufficient for diverse types of studies, and focusing on a given geographical region can be a way to provide enough relevant linguistic evidence. After appropriate filtering and selection, it is possible to draw maps to compare the geolocations of tweets with population density as a preliminary to study language variation (Arshi Saloot et al., 2016) or to use as input for classification tasks to determine language use in terms of variants on the social network (Alshutayri and Atwell, 2017).

2.2. Corpus building

While some studies ground on a collection process which is limited in time, the corpora described in this article are monitor corpora, as data collection goes on they grow with time. So-called "heavy tweeters" (Krishnamurthy et al., 2008) as well as peculiarities of the API (Morstatter et al., 2013) raise the question of sampling processes. The random sampling methodology used by Twitter to generate the streams of tweets is rarely put into question. This means that steps have to be taken in order to minimize or at least to assess the impact of differences in user activity as well as potentially unknown sampling biases. Studies have shown that it is desirable to gather a set of users which is both large and diverse (Zafar et al., 2015), so that the collection process is opportunistic. Such steps are described in previous work. It is possible to take decisions based on relevant metadata such as the number of followers or retweets as well as on information contained in the tweets themselves, such as the mention "RT" for retweet (Ruiz Tinoco, 2013). Additionally, it is possible only to take tweets coming from selected sources into account, for example by review the source fields in the collected tweets and focusing on common access points and clients (Tsou et al., 2017), most notably the website itself and the official Twitter mobile apps.

2.3. A suitable database infrastructure

The volume of storage space required to record and process the tweets is on the order of magnitude of several terabytes. Web data are a typical challenge for linguists working at public research institutions who do not dispose of large amounts of computing power (Tanguy, 2013). Additionally, Twitter data come in a form which has to be refined to suit the needs of linguists, as not all information and all metadata fields are linguistically relevant. In order to keep up with the challenges related to data structure and growing amount of tweets, we present our search engine of choice. The interest of NoSQL databases is known as regards the feature-rich content returned by the Twitter API (Kumar et al., 2014), they make it possible to access the corpus and see through it in various ways by using faceted searches. Their logic also supports indexing a variable number of metadata and efficiently divide the corpus into several subcorpora. In that sense, our purpose is to be opportunistic enough during corpus creation in order to allow for subcorpora which match particular interests.

Two main components of the open-source ELK stack (Elasticsearch, Logstash, Kibana) are used, namely Elasticsearch[1] to index the tweets and its front-end Kibana[2] to provide a user-friendly interface to queries, results, and visualizations. Installation and configuration are straightforward on most platforms, it is also possible to directly append servers to an existing cluster. Additionally, a sharding structure is implemented: shards (whether on the same machine or not) can be activated or deactivated to suit particular needs. Even with a single-server installation, it is convenient to process large amounts of tweets, that is on the basis of 10 Gb of data collected per day on the streaming API. The creation of subcorpora is possible through facets corresponding to a number of constraints acting on the text or the metadata (countries, precise time or date intervals, geographical window, etc.). Finally, the software is open-source and currently updated frequently, which gives access

[1] https://www.elastic.co/ Elasticsearch seems to be among the top-10 most popular database software at the moment https://db-engines.com/en/ranking

[2] https://www.elastic.co/de/products/kibana

to the latest optimizations or customizations (for example through Kibana's plugins).
Although it is not primarily a search engine for linguists, Elasticsearch takes advantage of the native JSON format of the tweets as well as of a number of relevant field types after a subsequent mapping, which allows for refined queries on text and metadata, which in turn ca be relevant for linguists, as we discuss in the remainder of this article. In order to give a user-friendly access to the results, dashboards can be configured out of a series of indicators. Despite its advantages for the structuration of non-standard data, the main drawback of nested JSON format resides in the lack of familiarity or compatibility with the formats commonly used in corpus linguistics, however the integration is possible through a number of conversion tools (e.g. plugins). The main drawbacks result at the time being from the built-in linguistic processing as well as a lack of integrated linguistic annotation. Considering non-standard speech, the standard lemmatization/normalization of queries and results by the search engine may be imprecise. Language-specific analysis modules can be selected, the default lemmatization is employed here due to the multilingual nature of our data, so that tokens can mainly be accessed on token or surface level.

3. Faceted querying and visualizations

In this section, we present three case studies focused on structured data synthesis in order to demonstrate the characteristics of language use on Twitter as well as the kind of information made available by our experimental setting.

3.1. Results presented as a dashboard

The first case deals with the amount of information available, which has to be filtered and presented it a synthetic fashion in order to allow for linguistic interpretation. We start from a classical display of results in corpus linguistics, the "word in context" (or KWIC) feature. Figure 1 shows a version of it which has been adapted to the tweets: results from several fields are aggregated into a dashboard, in that case the date, the text of the tweet (with the actual results), the name as chosen by the user, the "screen name" or account ID on Twitter, and the follower count. That way, it is straightforward to make a number of assumptions regarding the status of both user and tweet (for example concerning their influence). This example concerns colloquial German, where the contraction of *denkst du* in *denkste* has been considered to be typical for computer-based communication (Bartz et al., 2013). The query[3] contains the search for the exact expression *denkste* (without normalization), it also restricts the context to tweets in German which are not explicit retweets.

3.2. Combined time series

In a second example, we use the leverage provided by the amount of data to shed light on particular phenomena and use and visualization to corroborate hypotheses on language. Since metadata include the time of posting, it is possible to split the corpus in units of time. It is also quite natural to look at the axis of time in the monitor corpus, both in a linear and in a aggregated way. We present a longitudinal study which would be costly and tedious to realize in a classical fashion but where comparable information can be derived from the tweets themselves without artefacts linked to metadata. Figures 2a and 2b display the results of a time series query comparing two characterizing variants in Spanish in the course of the day: *buenos días* and *buen día*, Figure 2a focuses on tweets sent from Argentina while Figure 2b focuses on tweets sent from Spain, both are rendered using the Kibana's timelion plugin.[4] While a global figure would show patterns relative to the time zone of the users, these two uncouple the information to show the difference: a predominancy of *buen día* in Argentina and of *buenos días* in Spain, with a roughly similar pattern in the course of time highlighting that this expression is almost exclusively used in the morning. All this information is gathered in a suitable fashion for variation studies and is interpretable provided it is presented with the adequate circumspection.

3.3. Spatial analysis

Finally, our collection processes and infrastructure allow for the spatial studies on languages. A higher granularity of both queries and display is possible, as well as a direct access to the geolocalized tweets, which are then naturally interpretable in the form of a map. The tweets are automatically grouped by built-in clustering processes. The circles on the map display the number of tweets available for a particular place or region.
In order to illustrate the immediacy and practicality of the information available this way, we take Spanish diminutives as example, as there is a fair proportion of geolocated tweets to be found from different countries. There are several known diminutives for the word *café*, Figure 3a depicts the spatial distribution of tweets for the token *cafecito*, mostly in Central and South America, whereas Figure 3b focuses on *cafelito*, nearly exclusively to be found on the Spanish Peninsula. This comparison on maps using millions of tweets collected in 2017 confirms empirically this fact known to variation studies. On this order of magnitude, map processing from already indexed data is a matter of seconds and thus suitable for exploratory research from a linguistic and from a practical point of view.

3.4. Discussion

Twitter is a particular medium from an informational as well as from a sociological point of view. Age or gender biases are difficult to assess (Peersman et al., 2011), although they are certainly impacting both the structuration of the networks and the linguistic analysis. Additionally, a corpus may be affected by certain trends or keywords, beyond simple retweets or more complex repetitions of certain patterns. "Early adopters" and "power users" can distort a corpus quantitatively and qualitatively on the side of the observer and at the same time influence other users in their use of the language on the productive side.
Additionally, our experimental setting is not neutral and greatly contributes to shape the potential experiments. Nu-

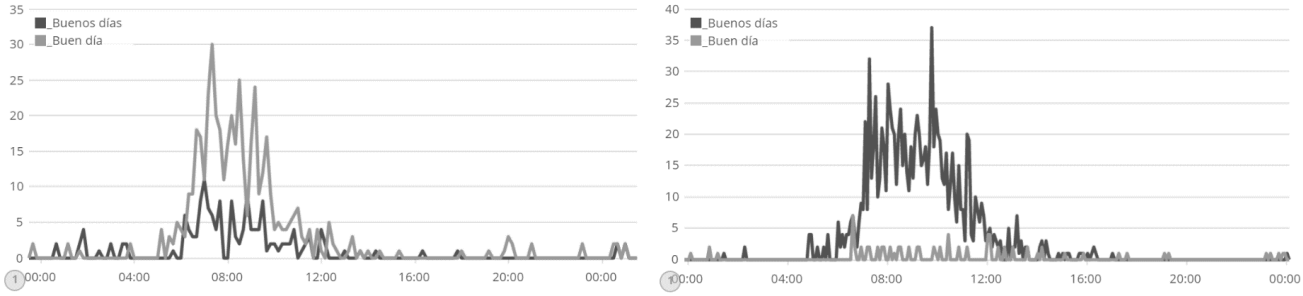
[3] text:"denkste" AND lang:de AND retweeted:False

[4] https://www.elastic.co/guide/en/kibana/current/timelion.html

text:denkste AND lang:de AND retweeted:false

@timestamp	text	user.name	user.screen_name	user.followers_count
April 16th 2017, 04:32:44.000	watt denkste @SirQLate ? Bombe, oder? https://t.co/V86HNpnphN	Frithjof Klepp	yakkul	89
March 8th 2017, 20:40:31.000	Da denkste oh lecker frisches Basilikum in der Tomatensoße und dann ist es Koriander...	Nadja	Suki9911	52
April 23rd 2017, 20:08:46.000	Da denkste hast nen schönen Sonntag, sind die Fellnasen andere Meinung https://t.co/rbd0CtuHbF	Papa Beck	_Papa_baer_	815
April 30th 2017, 03:00:00.000	achte isn typischer fall von denkste! bäm bäm bäm bäm bäm bäm bäm bäm	Rathausuhr Neukölln	rh_neukoelln	1,421
April 17th 2017, 16:40:20.000	Präsid.-kandidat #Macron: Deutschlands wirtsch. Stärke nicht mehr tragbar. Sollen wir schlechter werden? Denkste! https://t.co/YPpkG8C3wT	Heinz Scholz	HSderSchreiber	239
April 9th 2017, 01:22:25.000	Und dann fühlen die Finger zärtlich über die Veltins Relief-Flasche, und zack, denkste an Rudi. #S04	Torsten Wieland	TorstenWieland	4,507
April 29th 2017, 17:26:33.000	"130kmh sonne Rotze da denkste de stehst auf der kackautbahn" - mein Vater, 54, hat sich wiedermal blitzen lassen	Johann Scholz	ScholziWins	64
April 24th 2017, 05:20:00.000	Scheisse Da denkste die ganze zeit "Geile TL Heute" zack biste auf der @AnnaNymchen ihre Seite. Gönnt euch	Schattendasein	140zeichensucks	215
May 2nd 2017, 07:49:48.000	Da denkste, irgendwas in der Ecke bewegt sich doch und dann ist es Tim der auf dem Balkondeckenstapel hinter dem Vo… https://t.co/PgzP43ciqj	Çapulcu B	_blickwinkel_	3,728
April 5th 2017, 18:26:01.000	Sitz im Garten und hau eben mal ne Salami weg. Dann guckste auf die Inhaltsstoffe und denkste tut das not der ganze… https://t.co/yIypxnSNTW	inFarbeundBunt	sabinehartl1977	382

Figure 1: Detail/qualitative analysis in the dashboard mode, exact search for the contraction *denkste* in tweets identified as being in German and excluding retweets



(a) Focus on Argentina

(b) Focus on Spain

Figure 2: Two variants of salutation (singular and plural), visualized through Kibana according to the hours of use (in each respective time zone)

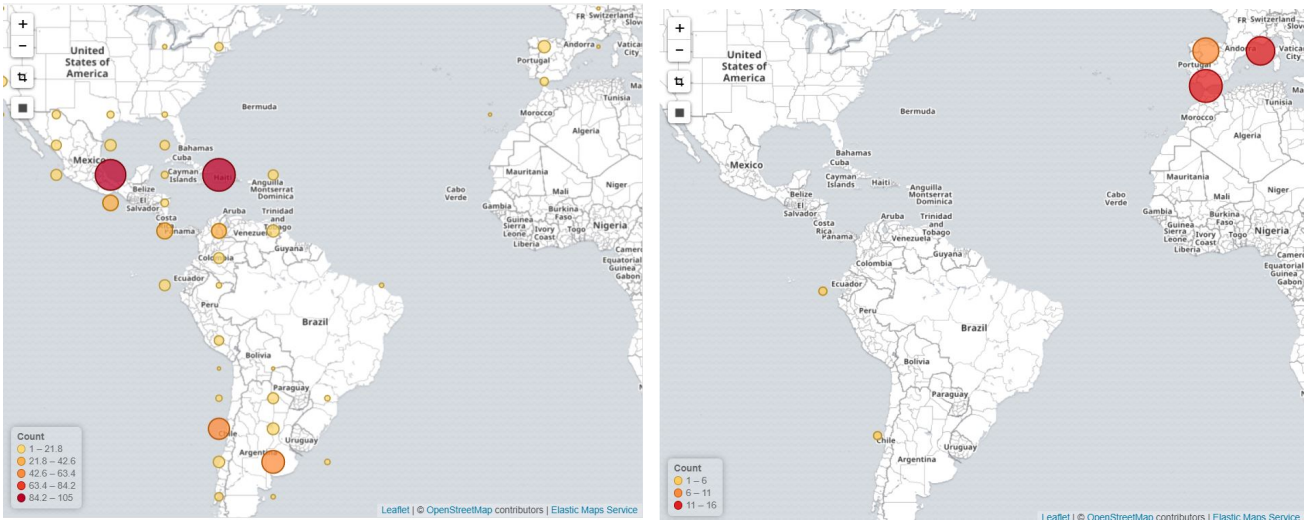
meric indicators tend to be favored as they allow for immediate realization of graphs such as bar charts. The analysis of text can get much more complicated, starting from the linguistic annotation tasks: language identification of tweets is error-prone. Catalan for example very often resorts to the "undetermined" category according to Twitter data. Furthermore, tweets are used for various purposes and may entail a diverging amount of "text" in a traditional linguistic sense (statements, replies, phatic formulas as opposed to hashtags, gifs, emoticons, and emojis).

Installations and tweet corpora are maintained at both institutions[5], they are used by colleagues and students alike, mostly for exploratory studies on spontaneous non-standard language. In this context, despite the lack of deep linguistic analysis, the interface is more user-friendly, easier to explain to students for example, and thus more directly usable for linguistic studies. The aggregation of data into a dashboard provides a way to find the right balance between profusion of data and relevance. We can still highlight two main artefacts of the apparatus in this case. First, there is a strong tendency to adjust the queries to the output, that is to test for hypotheses which lead to clear-cut results. Second, the projection on maps is the most prominent feature. Geolocated tweets distinguish this platform from other social networks, and corpus users are fond of reactive, interpretable maps.

4. Conclusion

The actual contents of a web corpus can often only be listed with certainty once the corpus is complete. In fact, corresponding to the potential lack of information concerning

[5]Academy Corpora, Austrian Academy of Sciences and Faculty of Foreign Studies, Sophia University.



(a) cafecito (b) cafelito

Figure 3: Spatial distribution of two diminutives of the word *café* in tweets

the metadata of the texts is a lack of information regarding the content, which has to be recorded and evaluated a posteriori, in a *post hoc* evaluation (Baroni et al., 2009). The notion of *a posteriori* analysis is a key concept regarding the study of tweets, whereas corpus design as considered by the tradition of corpus linguistics is systematically *a priori*. This bears both a risk and a chance: the linguistic relevance of the documents included is harder to assess, but it is also possible to determine new structural elements and discover relations in the data, for instance linguistic phenomena.

In this sense, we addressed corpus construction and visualization issues in order to study spontaneous speech and variation through short messages, notably metadata such as a location embedded in tweets. The technological stack based on Elasticsearch and Kibana has convinced us by its stability, scalability, and ease of use, although it ought to be complemented by further refinements and annotations in order to suit the needs of linguists.

We believe that our experimental setting can bring linguistic studies closer to actual language use. To this end, the adequation of the corpus with a given research goal has to be assessed. It is perfectly possible to adapt the geometry of the corpus to target a particular user type, region, or language. Yet beyond the scope of geographic variation as traditionally seen by examining utterances in a lexical or syntactic or other linguistic aspects, the study of online social networks also opens up new possibilities regarding user involvement and activity or general characteristics of the populations, features which would have needed a particular data collection effort in the past and which were not put into focus in variation studies.

5. Bibliographical References

Alshutayri, A. and Atwell, E. (2017). Exploring Twitter as a Source of an Arabic Dialect Corpus. *International Journal of Computational Linguistics (IJCL)*, 8(2):37–44.

Arshi Saloot, M., Idris, N., Aw, A., and Thorleuchter, D. (2016). Twitter corpus creation: The case of a Malay Chat-style-text Corpus (MCC). *Digital Scholarship in the Humanities*, 31(2):227–243.

Barbaresi, A. (2013). Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop*, pages 9–15.

Barbaresi, A. (2015). Collection, Description, and Visualization of the German Reddit Corpus. In *2nd Workshop on Natural Language Processing for Computer-Mediated Communication, GSCL conference*, pages 7–11.

Barbaresi, A. (2016). Collection and Indexing of Tweets with a Geographical Focus. In *Proceedings of CMLC-4, LREC 2016*, pages 24–27.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bartz, T., Beißwenger, M., and Storrer, A. (2013). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL*, 28(1):157–198.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A Few Chirps about Twitter. In *Proceedings of the First Workshop on Online Social Networks*, pages 19–24. ACM.

Kumar, S., Morstatter, F., and Liu, H. (2014). *Twitter Data Analytics*. Springer.

Leech, G. (2006). New resources, or just better old ones? The Holy Grail of representativeness. *Language and Computers*, 59(1):133–149.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5).

Ljubešic, N., Fišer, D., and Erjavec, T. (2014). TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. *Proceedings of LREC*, pages 2279–2283.

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)@ EACL*, pages 17–25.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2:460–475.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proceedings of ICWSM*.

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Ruiz Tinoco, A. (2013). Twitter como Corpus para Estudios de Geolingüística del Español. *Sophia Linguistica: working papers in linguistics*, 60:147–163.

Tanguy, L. (2013). La ruée linguistique vers le Web. *Texto! Textes et Cultures*, 18(4).

Tsou, M.-H. and Leitner, M. (2013). Visualization of social media: seeing a mirage or a message? *Cartography and Geographic Information Science*, 40(2):55–60.

Tsou, M.-H., Zhang, H., and Jung, C.-T. (2017). Identifying Data Noises, User Biases, and System Errors in Geo-tagged Twitter Messages (Tweets).

Zafar, M. B., Bhattacharya, P., Ganguly, N., Gummadi, K. P., and Ghosh, S. (2015). Sampling Content from Online Social Networks: Comparing Random vs. Expert Sampling of the Twitter Stream. *ACM Transactions on the Web (TWEB)*, 9(3):12.