The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data

Mikko Laitinen[1][0000-0003-3123-6932], Jonas Lundberg[2][0000-0001-9775-4594], Magnus Levin[2][0000-0002-5613-7618], and Rafael Martins[2][0000-0002-2901-935X]

[1] University of Eastern Finland, Agora 235, 80101 Joensuu, Finland
[2] Linnaeus University, Universitetsplatsen 1, 35195 Växjö, Sweden
mikko.laitinen@uef.fi

Abstract. This article presents the Nordic Tweet Stream (NTS), a cross-disciplinary corpus project of computer scientists and a group of sociolinguists interested in language variability and in the global spread of English. Our research integrates two types of empirical data: We not only rely on traditional structured corpus data but also use unstructured data sources that are often big and rich in metadata, such as Twitter streams. The NTS downloads tweets and associated metadata from Denmark, Finland, Iceland, Norway and Sweden. We first introduce some technical aspects in creating a dynamic real-time monitor corpus, and the following case study illustrates how the corpus could be used as empirical evidence in sociolinguistic studies focusing on the global spread of English to multilingual settings. The results show that English is the most frequently used language, accounting for almost a third. These results can be used to assess how widespread English use is in the Nordic region and offer a big data perspective that complement previous small-scale studies. The future objectives include annotating the material, making it available for the scholarly community, and expanding the geographic scope of the data stream outside Nordic region.

Keywords: Twitter, corpus linguistics, language choice, English as a lingua franca.

1 Introduction

This paper introduces a new real-time monitor text corpus of tweets from the Nordic countries. This corpus is a result of cross-disciplinary collaboration between computer scientists and a group of sociolinguists interested in language variability in general and English as a lingua franca (ELF) in particular. ELF is understood as second language use outside educational settings (cf. Mauranen et al. 2015). Our collaboration aims at better methodological accuracy in collecting new types of ELF data in multilingual settings, and we show how new sources of real-time data can lead to new insights in linguistics that are not only related to language learning (Bradley, 2016) and language assessment (García Laborda et al., 2016), but increasingly also to variationist sociolinguistics and its applications (Laitinen et al. 2017).

As is generally known, big and rich data from social media such as blogs, Facebook and Twitter have turned the web into a user-generated repository of information in ever-increasing numbers of areas. For instance, Twitter data have been used in social sciences to study the Arab spring (Campbell, 2011), to predict political campaigns (Gayo Avello et al., 2011; Tumasjan et al., 2010) and to predict stock markets (Bollen et al., 2011), and to model the geographic diffusion of new lexis (Eisenstein et al., 2014). Recent attempts also include incorporating data from various sources for applied purposes, such as the modelling of the impact of social networks in purchase intentions (Wang et al., 2016). Recently in linguistics, there have been various successful attempts to build both mono-lingual (Scheffler, 2014), and multilingual (Barbaresi, 2016) text corpora of tweets. In our field, corpus-based English sociolinguistics, more attention has been put on social media discussion fora (Mair, 2013), but tweets have also been explored (Knight et al., 2014; Huang et al., 2016; Coats 2016, 2017). At the same time, it has been argued that using real-time data streams is still in its infancy (Davies, 2015).

The Nordic Tweet Stream (NTS) initiative was started in April 2016, and it downloads tweets in real time from the Nordic region. The overarching goal is to tackle the role of social media and big language data in the global expansion and diversification of English. We explore the prospects of using Twitter data as a diagnostic tool in evaluating the changing role of English in lingua franca contexts and, as opposed to much of previous research in the field, our objective is to integrate two types of empirical data. We not only rely on traditional structured corpus data as is commonly done but also use unstructured data sources that are often big and rich in metadata, such as Twitter streams.

While we focus on a single geographic setting, the objective is to create a scalable tool that can be implemented in any location. The restriction to the five Nordic countries is justified in view of their many similarities. The countries constitute a geographically restricted region, and the main languages spoken there are largely related (Finno-Ugric Finnish being the exception to North Germanic Danish, Icelandic, Norwegian and Swedish), and English has a strong, though largely unofficial role (see, e.g., Leppänen et al., 2011; Bolton and Meierkord, 2013) in spite of there being no previous colonial ties between Britain and the Nordic region. English is the first foreign language taught in schools from an early age, and it is being used increasingly in research and higher education, business and the media.

2 Tapping into Twitter to create the NTS

Twitter is a microblogging platform that enables users to exchange short messages, tweets (www.twitter.com). Since its launch in 2006, it has expanded rapidly, and in November 2017 it was ranked as one of the most popular websites in the world by the Alexa ranking (http://www.alexa.com/topsites) with an estimated 310 million users publishing over half a billion tweets each day (www.internetlivestats.com/twitter-statistics/). Each Twitter user has a unique username (prefixed with @, e.g., @ThisIsHarryPotter) that can be used both as a signature and a reference. Each user has a number of friends (users they follow) and followers (users following them). Users can group

posts together by topic or type by use of hashtags (words or phrases prefixed with a # sign, e.g., #EurovisionSongContest). To repost a message from another Twitter user and share it with their own followers, a user can retweet the message.

In addition to the actual message, each tweet comes with a rich set of metadata (a selection is illustrated in Table 1), enabling researchers to make use of various metadata attributes, both user-generated and service-provided ones.

Table 1. A selection of the metadata parameters in the NTS

User-related info	**Description**
Name	user name
screen_name	user's Twitter name
Location	user's location
Description	descriptions of themselves
verified*	information whether an account is verified by Twitter (True/False)
followers_count*	number of Twitter followers
friends_count*	number of Twitter friends
account_identifier*	a unique account identifier number
tweets_issued*	number of tweets from one user
created_at*	date the account was created
time_zone	reported time-zone of the Twitter user
lang	reported language of the Twitter user
Place-related info	
place_type*	place of residence (country/city/ etc.)
place_name*	name of place of residence
country_code*	name of country of residence
geo_location*	[GPS Coordinates]
Tweet-specific info	
Date*	2016-07-03
Time*	00:00:31
Weekday*	Sunday
Lang*	En
Tweet	Why does Davos seem to be the only one around Stannis with his head on right? <HT>#emeliewatchesgot</HT> <HT>#got</HT> <HT>#GameofThrones</HT>

NB: * Indicates that these pieces of information are automatically generated as opposed to being user-provided information that can be misleading and inaccurate.

One of the advantages for sociolinguists is the fact that Twitter tries to assign a geolocation to each tweet (cf. the groundbreaking work done by Huang et al. (2016) for instance). We are here not referring to the user provided home location which often is misleading or missing (e.g. “Mars”, see Graham et al. 2013), we refer to the geolocation information provided by Twitter. Depending on user’s privacy settings and the geolocation method used, tweets either have an exact location specified as a pair of latitude

and longitude coordinates or an approximate location specified as a rectangular bounding box. Alternatively, no location at all is specified. This type of geographic information ('device location') represents the location of the machine or device on which a user sent a Twitter message. The data are derived either from the user's device itself (using the GPS) or by detecting the location of the user's Internet Protocol (IP) address (GeoIP). The primary source for locating an IP address is the regional Internet registries allocating and distributing IP addresses among organizations located in their respective service regions (e.g. RIPE NCC at www.ripe.net handles the European IP addresses). Exact coordinates are almost certainly from devices with built-in GPS receivers (e.g. phones and tablets).

Another reason why Twitter has been tapped into in various scientific projects is that it comes with an open policy allowing third-party tools or users to retrieve at most a 1% sample of all tweets. This service, the Twitter Streaming API, enables programmers to connect to the Twitter server and to download tweets in real time. The Streaming API provides three parameters – keywords, hashtags and geographical boundaries – which can be used to delimit the scope of tweets to be downloaded. Once the number of tweets matching the request starts to reach 1% of all available tweets, Twitter will begin to sample the data returned to the user.

Indeed, the drawback of using the Streaming API is the 1% limitation and the fact that Twitter is secretive about the sampling mechanism used. Morstatter et al. (2013) compare the sample provided by the Streaming API with the expensive Firehose API allowing access to 100% of all public tweets. Their investigation shows that the sample provided by the Streaming API was a sufficient random sample when the stream was filtered using geographical boundaries, and that the sample contained 43.5% of all tweets when the geographical boundary was a rectangle large enough to enclose the entire country of Syria. The sample received when filtering on keywords and especially hashtags was not as good. Their attempts to replicate the actual top-100 lists of most used hashtags using the Streaming API gave mixed results "indicating that the Streaming data may not be good for finding the top hashtags."

The NTS data collection makes uses the free Twitter Streaming API. As our downloading mechanism we use hbc (https://github.com/twitter/hbc), which is the default Twitter client when programming is done in Java. To collect tweets, we first specify a geographic region covering the five Nordic countries. A second filtering is added to select only the tweets tagged with a Nordic country code (DK, FI, IS, NO or SE). This second filtering is necessary to exclude tweets from neighboring countries (e.g., Germany and Russia) located within the chosen geographic boundary. Hence, NTS uses the geolocation information in each tweet to identify Nordic tweets and consequently, Twitter users who do not want to share their location are not included. Previous studies suggest that the proportion of geolocated tweets is low, between 0.7% and 2.9% depending on geographic contexts (e.g. Barbaresi, 2016). In an in-house experiment to test the accuracy of our data (during a 10-day period in August 2017), we compared the number of NTS tweets in which the language tag was Swedish (n=53,614) with tweets taken from another download project that tries to capture all tweets language-tagged as Swedish independent of geolocation or not (n=1,880,844). This results indicates that

only 2.8% of all tweets, tagged as Swedish, are geolocated to one of the Nordic countries. It should also be noted that GeoIP based device location can easily be tricked by using proxy gateways, allowing a user anywhere in world to "appear" to be located at a certain GeoIP address. The use of proxy gateways to hide the location is probably most common among user with a malicious intent. Figure 1 visualizes the corpus creation pipeline.

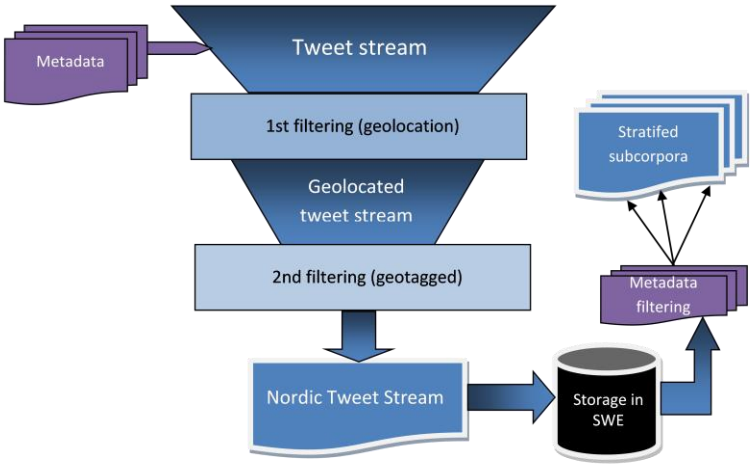

Fig. 1. The pipeline for creating the NTS

In order to test the coverage of the Streaming API we set up an automatic tweet generator publishing one tweet per hour with Sweden as the country code. In 67 days this generator published 1,608 tweets, 1,606 of which were captured by the NTS. We have also identified three other users from Finland, Norway and Sweden that publish tweets at regular intervals. We tracked them for 80 days and found that on average 98.9% of their tweets were captured by NTS. It thus seems likely that this way of downloading tweets includes a large majority of all geolocated tweets in the region.

As is common in studies that use geolocated tweets, the raw data also include tweets that are generated by automated bots (i.e. non-personal and organization-initiated machines), which often skews sampling (cf. Huang et al. 2016). In a spin-off project, we are using machine learning algorithms to recognize suspected bot accounts and use the method developed by Lundberg et al. (2018). This algorithm recognizes bot-generated tweets written in English and in Swedish, and we currently expanding the algorithm to the other main languages in the region. Since writing this article, we have implemented the bot-filtering algorithm, but the initial results reported here are based on data that contain English and Swedish bots.

3 Basic statistics

As of April 30, 2017, the NTS has downloaded 12,443,696 tweets from 273,648 user accounts and over 0.7 billion points of metadata. Figure 2 presents the number of tweets per day in the material.

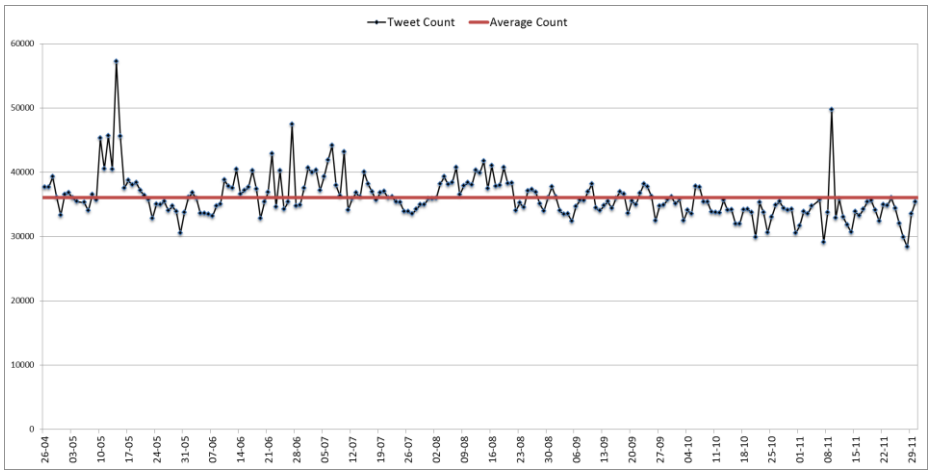

Fig. 2. Distribution of tweets in the five Nordic countries during the first months of data streaming

The average count per day is just below 36,805 tweets per day, and during the time period visualized here, there were two days during when the downloading system crashed. Some of the peaks in the frequencies of tweets are connected to events covered by the media. For instance, four of the five highest spikes in the data overall occurred on the 10, 12, 14 and 15 of May, and the Eurovision Song Contest, one of the most watched TV programs every year in the Nordic countries, took place on May 14, spilling over into May 15. The peak on June 27 is largely due to Iceland unexpectedly defeating England in the Euro 2016 football tournament. In 47,000+ tweets there were more than 5,000 occurrences or hashtags with the names of the two countries (e.g., Island till kvartsfinal!!!! ('Iceland to the quarterfinals' (Swedish)) and I love you Iceland. (tweeted by a Norwegian)). Immediately after the game, more than ten Nordic tweets per second were registered that discussed the game. Other days when the activity was greater include the day after the Brexit vote in Britain in the end of June, and the day after the presidential elections in the U.S. in November 2016.

In the near future, we plan on making parts of the data available for academic purposes via an intuitive search interface, which is possible according to Twitter's Terms of Service (Twitter TOS). This work is currently in progress, but it builds on the idea that the interface should be designed to be suitable for scholars with limited competences in data mining.

4 Case study: Some observations of language choice in the NTS

The empirical part demonstrates the potential uses of this corpus. Our approach is variationist sociolinguistic, meaning that we see language use as variable, in which a speaker/writer makes choices between alternative forms that are drawn from the pool of resources available (Tagliamonte, 2012: 3). We provide broad cross-country data on language choice, making use of the automatically generated metadata parameter of a tweet language in the data (Table 1 above). While we recognize that automated language identification methods are not entire accurate, the agreement between human coders and Twitter's language recognition system is fairly high for languages written in the Latin alphabet (Graham et al. 2013).

The results provide empirical evidence to two theoretically relevant questions related to ELF. Firstly, in many accounts ELF is still seen to be restricted to domain-specific uses such as academia and international business/law (cf. Mair 2013). Note that such a restricted view is not always shared among ELF scholars, and Pietikäinen (2017) has shown that ELF is used in the family settings for instance. Our big social media data approach enables testing this empirically and estimate the extent to which ELF is used in the Nordic setting. Secondly, we add a big data perspective that complements traditional survey studies; we use a sample of over 200,000 informants. This figure can be contrasted with the samples used in a few previous studies. For instance, the results of a traditional mail-in survey in Finland in 2007 were based on a stratified sample of 1,495 respondents (Leppänen et al., 2011). Similarly, an exploratory interview study of the role of English Sweden drew data from 28 respondents (Bolton and Meierkord, 2013). Naturally, the amount of information extracted through a carefully-designed survey or an interview study can be extensive, but we wish to highlight the need to combine methods from both traditional methodologies and studies making use of big and rich data.

Figure 3 shows the language distribution in our data. English is the main language, and its share is 32.9%. This figure is slightly smaller than the share of English in the Austrian monitor tweet corpus (Barbaresi, 2016), in which the share was 42.2%. The main languages in the region are the next most frequent: Swedish (25.9%), Finnish (11.5%), Norwegian (5.5%), Danish (4.9%), and Icelandic (1.9%).

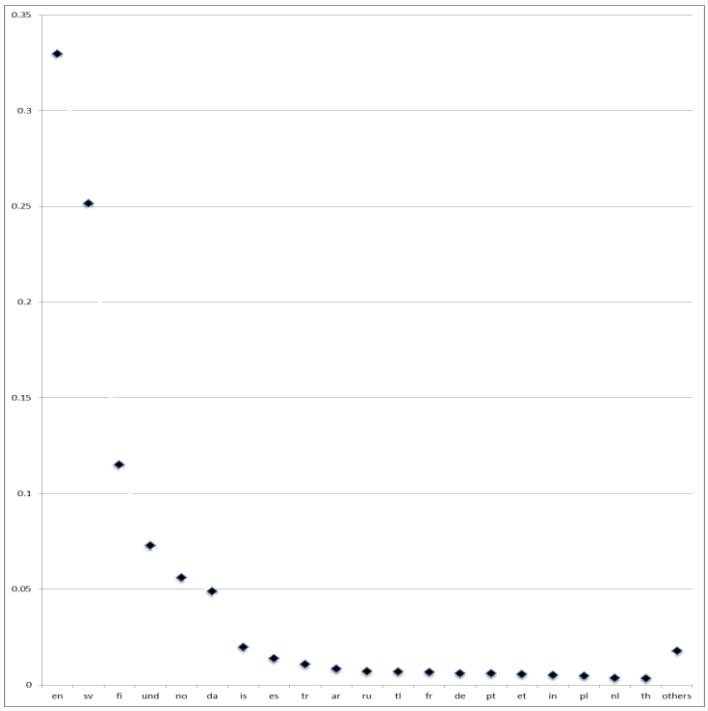

Fig. 3. Language choice in the NTS data.

Some brief comments are needed about the some of the other classifications. The category "und(efined)", which represents about 7%, to a great extent consists of messages from two different categories: on the one hand promotions of hashtags or URLs, either to individual friends or to all followers, and on the other instances where there is too little linguistic material to identify the language (e.g., simply ? or ? wtf ? <URL> … </URL>"). The unexpectedly large share of Tagalog (code=tl) stems from laughter such as hahhah erroneously being coded as tl.

Apart from English and the main/national languages, the shares of other languages are small. Most of them are European languages, but also immigrant languages in Europe are among most frequently used ones, i.e. Arabic, Turkish, Russian, Indonesian, and Thai.

The distributions of the languages show both regional similarities and differences. With regards to similarities, when we divide the data according to the five countries, English is among the top two languages in every country (Table 2). Its share varies between the lowest share (26%) in Finland and the highest (46%) in the Danish data. Table 2 shows the five most frequently used languages. It excludes the tweets with unidentified language codes.

Table 2. The top-5 language per five countries in the NTS

Rank	Denmark	Finland	Iceland	Norway	Sweden
1	English (46%)	Finnish (55%)	Icelandic (46%)	English (37%)	Swedish (52%)
2	Danish (30%)	English (26%)	English (36%)	Norwegian (31%)	English (29%)
3	Spanish (2%)	Estonian (2%)	Spanish (2%)	Danish (5%)	Spanish (1%)
4	Norwegian (2%)	Russian (2%)	French (1%)	Spanish (2%)	Arabic (1%)
5	Swedish (2%)	Swedish (1%)	German (1%)	Swedish (2%)	Turkish (1%)

If we add up the proportions of English and the main/national language of each country in Table 3, the two most frequent languages account for over 80% of the languages used in Iceland (82%), in Sweden (81%) and Finland (81%). The total shares in Denmark (76%) and Norway (68%) are substantially lower. A notable fact is that immigrant languages primarily appear among the most frequent language in Sweden (Arabic and Turkish) and in Finland (Estonian and Russian). We do not yet know if these differences are reflections of real variability or whether they have been brought about by technical factors.

Figure 4 below shows a pattern over the average day in the Nordic region. It visualizes the hourly averages of all the material captured by the stream from April 2016 to April 2017 and the share of English (lang=en) material. As for a measure of variability, the vertical bars for the standard error of the mean help to estimate how significant the hourly shifts are.

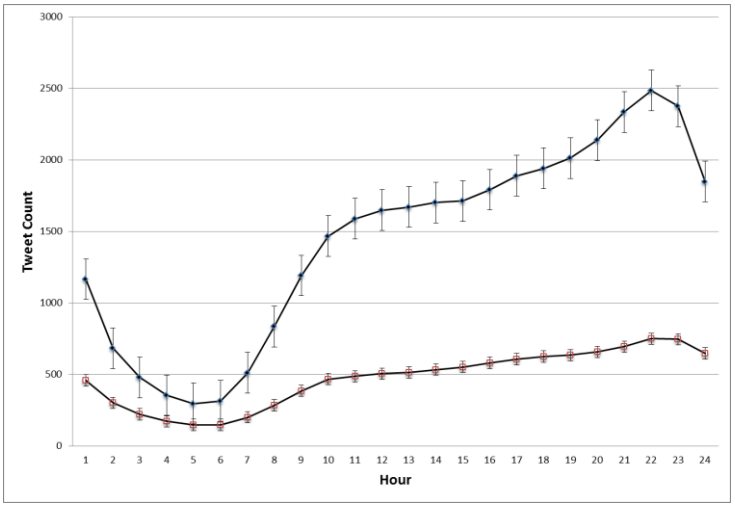

Fig. 4. Tweet stream and the share of English per hour compensated for time zones in NTS

The figure presents a clear pattern of the distribution of tweets. On the one hand, the temporal distribution follows the daily patterns of most people. The pattern is highly similar to the one in the German Twitter snapshot (Scheffler, 2014), but there are also noticeable differences. Firstly, as could be expected, people tweet the least in the early morning, reflected in the dramatic drop in the hourly activity after midnight, and from then on the frequency increases steadily. The normal office hours see a constant increase in the activity, and the activity peaks at 9–11 PM from where it decreases. The high frequencies late in the evening support the finding that tweeting to a great extent is connected to late evening leisure activities. The main difference between our data and the data presented in Scheffler (2014) is that the peak in her German material occurs around 8 or 9 PM.

The proportion of English is the highest when Twitter activity is at its lowest at 5 in the morning, reaching almost 50% of all tweets, and at its lowest when the Twitter activity is at its highest, dropping just below the 30% mark. The high proportion of English can partly be explained by the low overall proportion of tweets, since some tweets that are automatically generated around the clock, such as weather reports, are produced in English.

Lastly, we are currently testing various options of visualizing the material. Figure 5 below illustrates how StanceXplore (Martins et al. 2017) could be used to visualize language choice. This tool was originally developed as an interactive tool for modelling stance taking in social media platforms. Stance and sentiment refer to the ways speakers position themselves in relation to their own or other people's beliefs, opinions, and statements in communicative situations with others.

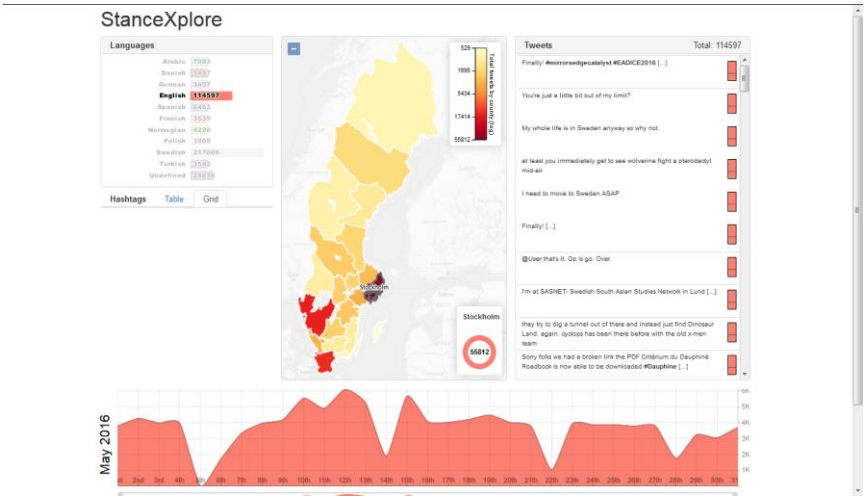

Fig. 5. StanceXplore used to visualize language choice

The tool makes it possible to capture a more fine-grained view of language choice in specific regional contexts, as is the case with the counties in Sweden here. This snapshot shows how 114,597 tweets in May 2016 are distributed regionally and chronologically. The tool visualizes the number of messages in various languages and the user can select individual languages and see their regional frequencies and the actual messages keyed in by the user.

5 Conclusions

This article has introduced a new multilingual Twitter corpus covering five countries in the Nordic region. The corpus is a real-time monitor corpus that is both big in size and rich in metadata. The article has presented some of the early observations in the first months of the streaming process, which started in spring 2016. The objective is to continue the streaming for at least a year, thus updating the corpus with nearly 37,000 tweets per day. The data collection is taking place on a two-layered model in which we limit ourselves geotagged tweets in a specified geographic region, and we hope to expand the method to new regions.

We are currently working on annotating parts of the material. This work consists of lemmatizing the tweets and running parts-of-speech tagging to the material. The work has started from the English and the Finnish material. Another future objective is to build an intuitive search interface for the material.

References

1. Barbaresi, A.: Collection and indexation of Tweets with a geographical focus. Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 2016. In: Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC), 24–27. <hal-01323274>. (2016).
2. Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8. (2011).
3. Bolton, K. and Meierkord, C.: English in contemporary Sweden: Perceptions, policies, and narrated practices. Journal of Sociolinguistics 17(1), 93–117. (2013).
4. Bradley, L.: The Mobile Language Learner – Use of Technology in Language Learning. Journal of Universal Computer Science 21(10), 1269–1282. (2016). doi: 10.3217/jucs-021-10-1269.
5. Campbell, D. G.: Egypt Unsh@ckled: Using Social Media to @ # :) the System: how 140 Characters Can Remove a Dictator in 18 Days. Llyfrau Cambria/Cambria Books. (2011).
6. Coats, S.: Grammatical feature frequencies of English on Twitter in Finland. In: Squires, L., English in Computer-mediated Communication: Variation, Representation, and Change, 179–210. Berlin: De Gruyter. (2016).
7. Coats, S.: 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/. (Accessed 15 Jan 2018). (2017).

8. Davies, M.: Corpora: an introduction. In: Biber, D. and Reppen, R. (eds.), The Cambridge Handbook of English Corpus Linguistics, 11–31. Cambridge: Cambridge University Press. (2015).
9. Eisenstein, J, O'Connor, B., Smith, N.A and Xing, E.P.: 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(11). doi:10.1371/journal.pone.0113114
10. García Laborda, J., Magal Royo, T. and Bakieva, M.: 2015. Looking towards the Future of Language Assessment: Usability of Tablet PCs in Language Testing. Journal of Universal Computer Science 21(10), 114–123. (2015).
11. Gayo Avello, D., Metaxas, P. T. and Mustafaraj, E:. Limits of electoral predictions using twitter. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence. (2011).
12. Graham, M., Hale, S. A., and Gaffney, D.: Where in the world are you? Geolocation and language identification in twitter. The Professional Geographer 66, 568–578. (2013). doi 10.1080/00330124.2014.907699
13. Huang, Y., Guo, D. Kasakoff, A. and Grieve, J.: Understanding US regional linguistic variation with Twitter data analysis. In: Computers, Environment and Urban Systems 59 (2016) 244–255.(2016). doi:10.1016/j.compenvurbsys.2015.12.003.
14. Knight, D., Adolphs, S. and Carter, R.: CANELC: Constructing an e-language corpus. Corpora 9(1), 29–56. (2014).
15. Laitinen, M., Lundberg, J., Levin M., and Lakaw, A.: Revisiting weak ties: using present-day social media data in variationist studies. In: Säily, T., Palander-Collin, M., Nurmi, A. and Auer A. (eds.), Exploring Future Paths for Historical Sociolinguistics, 303–325. Amsterdam: John Benjamins. (2017.). doi 10.1075/ahs.7.12lai.
16. Leppänen, S., Pitkänen-Huhta, A., Nikula, T., Kytölä, S., Törmäkangas, T., Nissinen, K., Kääntä, L., Räisänen, T., Laitinen, M., Koskela, H., Lähdesmäki, S. and Jousmäki, H.. National Survey on the English Language in Finland: Uses, Meanings and Attitudes. Available at: <http://www.helsinki.fi/varieng/series/volumes/05/> (2011).
17. Lundberg, J., Nordqvist, J. Matosevic , A. On-the-fly Detection of Autogenerated Tweets, arXix preprint (2018).
18. Mair, C.: The World System of Englishes: Accounting for the Transnational Importance of Mobile and Mediated Vernaculars. English World-Wide 34, 253–278. (2013).
19. Martins, R. M., Simaki, V., Kucher, K., Paradis, C., and Kerren, A. StanceXplore: Visualization for the Interactive Exploration of Stance in Social Media. In: 2nd Workshop on Visualization for the Digital Humanities (VIS4DH'17), October 2017, Phoenix, Arizona, USA. (2017).
20. Mauranen, A., Carey R., and Ranta. E.: New answers to familiar questions: English as a lingua franca. In: Biber D. and Reppen R. (eds.), Cambridge Handbook of English Corpus Linguistics, 401–417. Cambridge: Cambridge University Press. (2015).
21. Morstatter, F., Pfeffer, J., Liu, H., and Carley. K.M.: Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In: Association for the Advancement of Artificial Intelligence International Conference on Weblogs and Social Media 7: 400–408. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071. (2013).
22. Pietikäinen, K. ELF in social contexts. In: Jenkins, J., Baker, W., and Dewey M. (eds.) The Routledge handbook of English as a lingua franca, 321 – 332. Abingdon: Routledge.
23. Scheffler, T.: A German Twitter Snapshot. In: Proceedings of LREC, 2284–2289. (2014).
24. Tagliamonte, S.: Variationist Sociolinguistics: Change, Observation, Interpretation. London: Blackwell. (2012)

25. Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM 10, pp. 178–185. (2010).
26. Twitter TOS, Twitter's Terms of Service, https://twitter.com/tos, last accessed 31 Sept. 2017.
27. Wang, H-W. et al.: Exploring the Impacts of Social Networking on Brand Image and Purchase Intention in Cyberspace. Journal of Universal Computer Science 21(11) 1425–1438, (2016). doi: 10.3217/jucs-021-11-1425.