Journal of Universal Computer Science, vol. 23, no. 11 (2017), 1038-1056
submitted: 1/9/16, accepted: 26/11/17, appeared: 28/11/17 © J.UCS

Utilizing Multilingual Language Data in (Nearly) Real Time: The Case of the Nordic Tweet Stream

Mikko Laitinen
(School of Humanities, University of Eastern Finland and Department of Languages
Linnaeus University, Växjö, Sweden
(mikko.laitinen@uef.fi / mikko.laitinen@lnu.se)

Jonas Lundberg
(Department of Computer Science, Linnaeus University, Växjö, Sweden
jonas.lundberg@lnu.se)

Magnus Levin
(Department of Languages, Linnaeus University, Växjö, Sweden
magnus.levin@lnu.se)

Alexander Lakaw
(Department of Languages, Linnaeus University, Växjö, Sweden
alexander.lakaw@lnu.se)

Abstract: This paper presents the Nordic Tweet Stream, a cross-disciplinary digital humanities project that downloads Twitter messages from Denmark, Finland, Iceland, Norway and Sweden. The paper first introduces some of the technical aspects in creating a real-time monitor corpus that grows every day, and then two case studies illustrate how the corpus could be used as empirical evidence in studies focusing on the global spread of English. Our approach in the case studies is sociolinguistic, and we are interested in how widespread multilingualism which involves English is in the region, and what happens to ongoing grammatical change in digital environments. The results are based on 6.6 million tweets collected during the first four months of data streaming. They show that English was the most frequently used language, accounting for almost a third. This indicates that Nordic Twitter users choose English as a means of reaching wider audiences. The preference for English is the strongest in Denmark and the weakest in Finland. Tweeting mostly occurs late in the evening, and high-profile media events such as the Eurovision Song Contest produce considerable peaks in Twitter activity. The prevalent use of informal features such as univerbated verb forms (e.g., *gotta* for (HAVE) *got to*) supports previous findings of the speech-like nature of written Twitter data, but the results indicate that tweeters are pushing the limits even further.

Keywords: Twitter, corpus linguistics, language choice, oral discourse style
Categories:

1 Introduction

This paper introduces a new real-time monitor text corpus of tweets from the Nordic countries. The corpus is a result of cross-disciplinary collaboration of a computer scientist and a group of sociolinguists. This collaboration aims at better methodological accuracy in collecting new types of data and builds on theoretical

relevance in analyzing them in those fields in humanities that could greatly benefit from cross-disciplinary collaboration. We show how new forms of computer-mediated communication data can lead to new insights in humanities and in linguistics that are not only related to language learning [Bradley, 2015] and language assessment [García Laborda et al., 2016], but increasingly also to sociolinguistics and its applications. In recent years, 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, and various big data approaches have started tapping into this rich material. For instance, data from the micro-blog platform Twitter 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], to predict stock markets [Bollen et al., 2011], and to model the geographic diffusion of new lexis [Eisenstein et al., 2012]. Moreover, 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 linguistics, 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 2017; Laitinen et al. 2017]. At the same time, it has been suggested that using social media as a source of data is still in its infancy [Davies, 2015].

We present the first results from the Nordic Tweet Stream initiative (NTS), which is downloading tweets in real time from the Nordic region. It was initiated in April 2016, and we plan to continue for several years, thus creating a large and dynamic data base for sociolinguistic research. 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 to study the social challenges posed by its expansion in the Nordic region [Laitinen and Levin, 2016].

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.

The specific aims of the present paper are to (a) introduce the technical details of our material collection process, (b) to present the basic statistics of the pilot stage of the first four months of data streaming, and (c) to illustrate the potential uses of our corpus through two case studies that focus on English as part of the multilingual repertoire in the tweet stream in the Nordic region. In the last part, we make use of not only the tweet content but also the rich metadata repertoire in tweets. This big and rich data approach of our cross-disciplinary collaboration leads to increased accuracy

in sociolinguistic descriptions, especially when the results are contextualized with traditional survey data.

The paper is structured in the following way: section 2 presents background information on Twitter, NTS and the sampling frame used. Section 3 details the basic statistics of the pilot phase of the data collections. Section 4 presents the first results of language choice in tweet data, its stratification according to the metadata parameters, and a case study of one discourse-related phenomenon. Lastly, we will detail the future prospects of our approach.

2 Twitter and the Nordic Tweet Stream

Twitter is a microblogging platform allowing users to exchange short messages called tweets [www.twitter.com]. Since its launch in 2006, it has expanded rapidly, and in February 2018 it was ranked as the 12th most popular website in the world by the Alexa ranking [http://www.alexa.com/topsites] with an estimated 310 million users publishing 500 million 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 (repost) the tweet.

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.[1]
As an illustration of the tweet-specific information, Twitter's own language identification tool, which is based on a machine-learning algorithm, is used to classify the languages in the tweets [https://blog.twitter.com/2015/evaluating-language-identification-performance].

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). Note that we are not referring to the user provided home location which often is misleading or missing [Graham et al. 2013], but rather refer to the geolocation information provided by Twitter. Depending on users' 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. For example, RIPE NCC

[1] The service extended the length of a message from 140 characters to 280, excluding url-links, re-tweets, etc in November 2017.

(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). Bounding boxes, however, can result from privacy settings applied to GPS data or from GeoIP data. 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.

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. Throughout the article we show the text-level mark-up around repetitive items, such as hashtags, retweets, etc.

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

Another main reason why Twitter has been tapped into in various scientific projects (apart from its widespread use and rich metadata) 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 is called the Twitter Streaming API and it 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 rather good 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 taps into the tweet stream from five Nordic countries, excluding the Svalbard Islands. Figure 1 visualizes the region on a global scale.

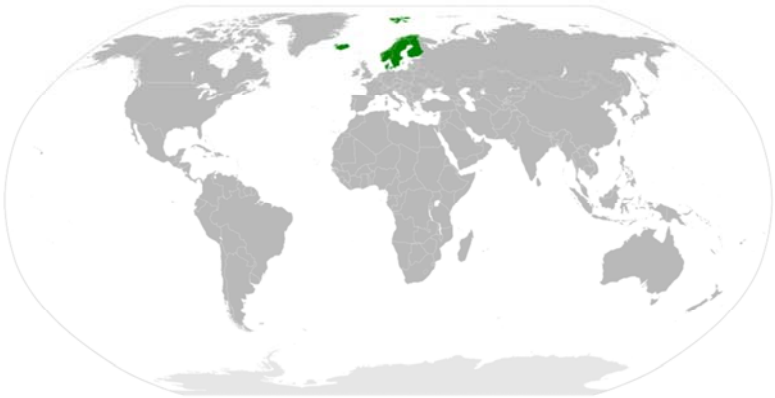

Figure 1: The approximate geographic scope of the NTS (the map is available for public domain through the Creative Commons license)

The data collection makes use of the free Twitter Streaming API. As our downloading mechanism we used 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 (Figure 1). 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. It is difficult to determine how many Nordic tweets are missed due to this group of users, but previous studies suggest that in other geographic contexts, the proportion in general is low [Barbaresi, 2016]. Figure 2 visualizes the corpus creation pipeline.

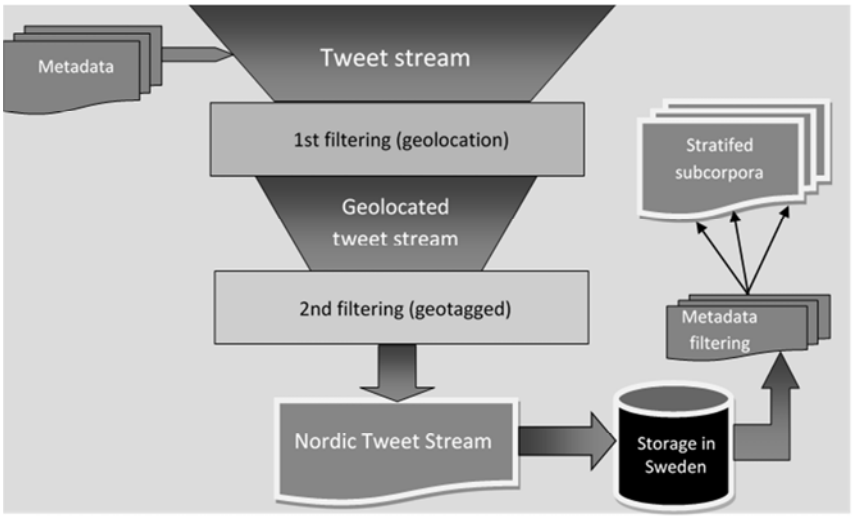

Figure 2: The pipeline for creating the NTS

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 1608 tweets, 1606 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.

A late-breaking addition to this article is that we have implemented the bot-filtering algorithm in a spin-off project. This project aims at increasing accuracy of using tweets in sociolinguistic research. 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 [Huang et al. 2016]. We currently use machine learning algorithms to recognize suspected bot accounts and use the method developed by [Lundberg et al. 2018]. Their 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, but the results reported here are based on data that contain English and Swedish bots.

3 Basic statistics

As of August 8, 2016, the NTS had downloaded 6,639,648 tweets from 183,210 user accounts.[2] Note that the user accounts do not represent the exact number of

[2] After implementing the bot-recognition algorithm in autumn 2017, we have considerably increased the accuracy of our data streaming, and the tweet count in 31

individuals, as any user is free to register a new account. The character count in this raw data excluding emojis is over 628 million orthographic units, and there are hundreds of millions of points of metadata. In the future, we plan on making parts of the data available for academic purposes via a website, which is possible according to Twitter's Terms of Service.

The Tweet count statistics are based on the number of tweets each user published since we started downloading tweets (April 26, 180 days).

Quantity	Median	Average	St. deviation	Range
Tweets	4	36	530	[1, 149582]
Friends	234	522	2906	[0, 834528]
Followers	194	1116	10653	[0,1883395]

Table 2: Per user count of tweets, friends, and followers for Nordic users

The median value 4 indicates that most Twitter users are publishing less than one tweet each month. This suggests that users mainly use Twitter as a one-way source of information. They follow a number of interesting users (people, organizations) without themselves actually participating or contributing markedly to the exchange of information. The large difference between the median (4) and average (36) values, and the very high standard deviation (530), indicate that there are a few power users who tweet a lot. A closer inspection of the most extreme power users (> 100 tweets a day) indicates that these are machine generated. The Nordic top tweeter (#EveryFinnishNo) is simply a program tweeting a new number (in text) about 840 times a day. As pointed out in the previous section, we are currently working on blacklisting bots and other unwanted accounts [cf. Barbaresi, 2016].

The number of friends and followers for a user varies over time as they get more contacts. The friends and followers statistics are therefore based on the last tweet downloaded for each user. Furthermore, the so-called verified users have been removed from the numbers in Table 2. Verified users are often celebrities or user accounts that represent a company or an organization. For example, the clothing manufacturer H&M has a verified user account with more than 8 million followers.

The median values (234 friends, 194 followers) indicate that the average user interacts with around 200 users. The very high standard deviation indicates once again that we have a number of power users with many more friends and followers than the average user. For instance, #SciencePorn, a user publishing light science-related URLs, has 1.8 million followers in the data. In our sociolinguistics research, we have made use of friends and followers information to test the social network model in sociolinguistics using Twitter data [Laitinen et al. 2017].

December 2017 is 10,325,217 messages from over 200,000 accounts, with English and Swedish bots cleaned with an accuracy rate of 96% [Lundberg et al. 2018].

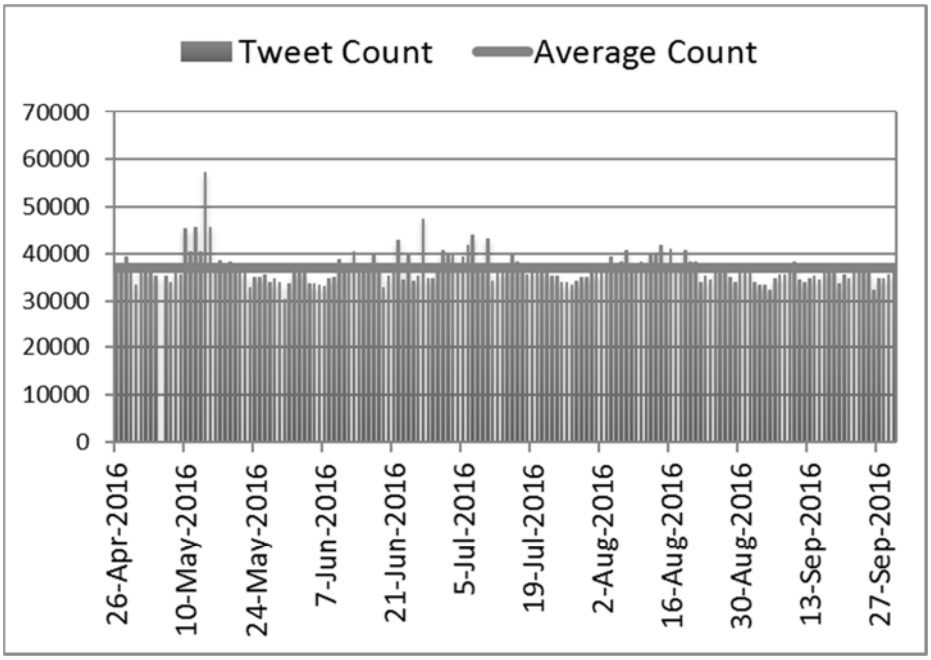

Figure 3: Distribution of tweets in the five Nordic countries from 180 days during 2016

Figure 3 presents the number of tweets per day in the material. 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, while the two semi-finals were held on May 10 and 12. On May 14, almost 9,000 of 57,000 Nordic tweets contained the word *Eurovision*, either in the running text or as a hashtag (e.g., *The points are all over the place so far. <HT>#Eurovision</HT> <HT>#escse</HT>* (tweeted by a Swede)), but the focus on the competition is obvious also in many other tweets (e.g., *Ge Inte så låga poäng* 😩 ('don't give such low scores' (Swedish))). 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)). Here too large numbers of other tweets in different languages relate to this specific media event (e.g., *Roy reiser.. hjem.....* ('[England coach] Roy [Hodgson] is going home' (Norwegian)). Immediately after the game, more than ten Nordic tweets per second were registered that discussed the game.

In contrast to the distributions across the days during the time span of data collection, distributions across weekdays provided only little variation. The average

number of tweets produced is the lowest on Mondays (35,500) from where it increases steadily to its peak on Fridays (37,800) to then drop off slightly towards the next week.

Figure 4 presents the frequencies of tweets per country.

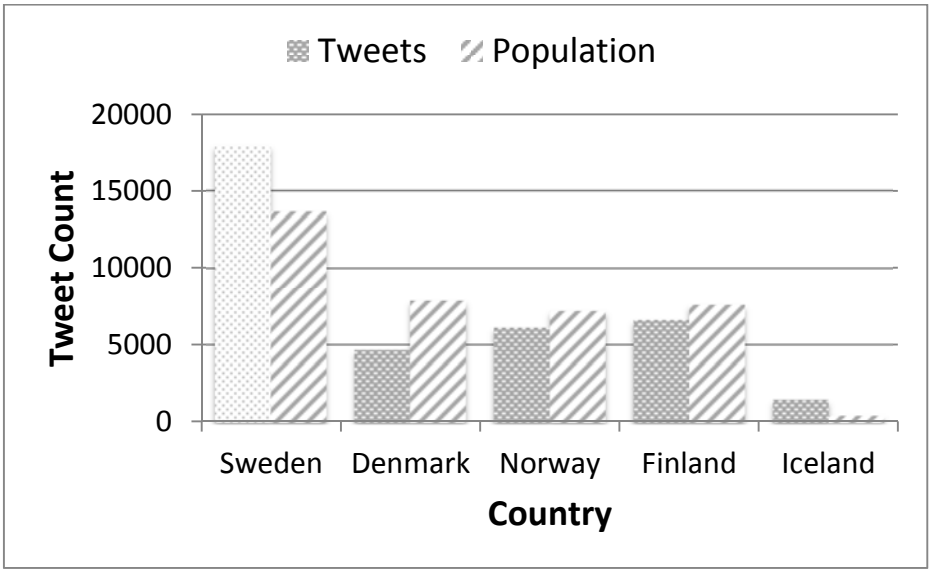

Figure 4: The average number of tweets per day as compared to the expected proportion based on the population in each country

The figure shows that the use of Twitter is more widespread in Iceland and Sweden than in Denmark, Norway and Finland. Almost half of all tweets (48.7%) were written in Sweden although only 37% of the Nordic population lives there. These differences cannot be explained by different levels of internet access, since all five countries rank among the top twelve in the 2015 ICT Development Index [International Telecommunications Union, 2015] with Denmark in the second place worldwide.

Figure 5 below shows a clear pattern over the average day. It visualizes the hourly averages of all the material captured by the stream and the share of English (lang=en) material. The vertical bars for standard error of the mean help to estimate how significant the shifts are.

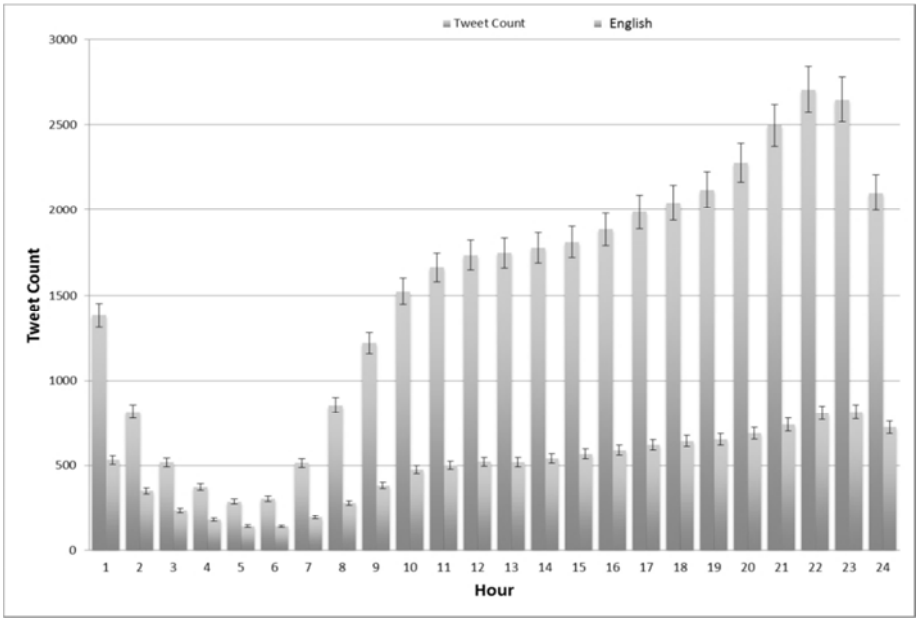

Figure 5: Tweet stream and the share of English per hour compensated for time zones in NTS

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.

If there is considerable variation in the proportion of English during the day, there is much variation less across weekdays. The proportion is at its highest (32.7%) on Thursdays and its lowest (31.6%) on Saturdays.

4 Results

Our empirical part demonstrates the potential uses of this corpus and presents two studies that focus on sociolinguistic aspects. The first provides broad cross-country data on language choice in the Nordic region and makes use of the automatically generated metadata parameter of a tweet language in the data (see 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 second one focuses on the texts keyed in by the tweeter and looks into discourse properties of tweets. It investigates to what extent the Twitter messages written in the most frequently used language in data, English, exhibit oral style. Previous studies have suggested that digital communication has blurred the traditional distinctions between written and spoken communication and have shown that elements from spoken language are very frequent in e-language in general [Knight et al., 2014].

We approach both of these topics with a sociolinguistic focus, 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]. Alternative forms exist on all levels of language, starting from the basic question of what language one uses to minute phonetic alternations of sounds that make up syllables. As an illustration, a bi/multilingual tweeter has to make a choice of what language to use when tweeting and decide between oral vs. literate variant forms in the text.

The results provide empirical evidence to two theoretically relevant questions. On the one hand, our data adds a big data perspective to the globalization of English in the expanding circle context. The perspective is novel, since the previous approaches on English in the expanding circle are based on small sets of data. It is beyond doubt that English has spread considerably in recent decades. This expansion has been brought about by mobility and the emergence of the internet, and it is fair to say that English today serves as a symbol of modernization and globalization [Schneider, 2014]. One specific angle of this debate is the theoretical notion that there might be several regional and social "centers of action" developing globally [Kohnen and Mair, 2012]; these "centers" are socially and culturally strong areas in which speakers play a considerable role in shaping English. The Nordic region is a pilot case and the objective is to extend this to other geographic regions. We propose that quantitative evidence from big data sources can reveal the existence of such centers. On the other hand, there is need to generate empirical data of how much English is used in daily life in one possible center, the Nordic region. As noted above, this region is technologically advanced, socially and culturally relatively homogenous, and the role of English in the educational systems is equally strong throughout. English has no official legal position in the five countries, but its de facto role is important. There is therefore need to provide comparative empirical evidence of how English is actually used in the region.

As mentioned in Section 3, our results make use of a sample of over 183,000 informants. This informant figure can be contrasted with the sample populations 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 in 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, and we wish to highlight the need to combine methods from both traditional methodologies and studies making use of big and rich data.

4.1 Language choice in the Nordic Tweet Stream

Figure 6 shows the language distribution in our data. English is the main language, and its share is 32.3%. 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 Nordic region are the next most frequent. The highest share is Swedish (26.2%), followed by Finnish (10.6%), Norwegian (5.9%), Danish (5.1%), and Icelandic (2.1%). The fact that so many of the tweets are written in Sweden (as seen in Figure 4 above) at least partly explains why Swedish is so much more frequent than any other language except English. The relatively low proportions of Danish and Norwegian as compared to Swedish and Finnish will be discussed further below.

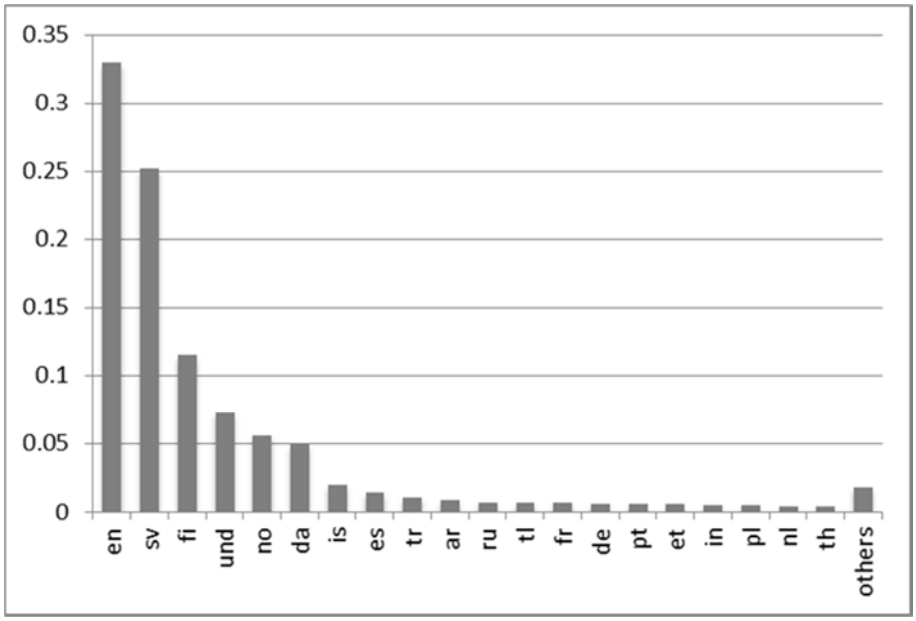

Figure 6: Language choice in the NTS data.

One possible factor in language choice is the cognitive factor of how salient one language is for an individual tweeter. English is the main language on Twitter in general, so it can be assumed that if a user re-sends a tweet, it could influence language choice. The data show, however, that retweeting is a very infrequent phenomenon, only 0.02% of all the NTS tweets contain retweets, and that has no major impact on our numbers.

Some brief comments are needed about the some of the other classifications. The category "undefined", 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.

To provide a snapshot of what the language-choice situation looks like in relation to English, we sampled 20 users of English by picking the first English tweet from randomly selected hours from randomly selected days (excluding automatically generated tweets). Of these users, 12 tweeted in English only (or, in some cases, also produced some "undefined" tweets) during one day. Among these, at least eight could be identified as most likely being native speakers of Scandinavian languages but nevertheless choosing to use English (writing, e.g., *Going to Comic Con Copenhagen tomorrow 😄 I'm so excited!*). Two were apparently English speakers (from Ireland and the US) visiting Iceland (tweeting, e.g., *Hi vikings*). Of those using English in parallel with other languages, one tweeter mostly used Serbo-Croatian and another Indonesian, while five of the others switched between English and the Nordic languages. Language choice here seems to depend on the topic and who the tweeter is addressing. For instance, a Swede tweeted another Swede in Swedish, asking *<AT>@...</AT> sett senaste the Purge?* ('seen the latest the Purge?') and later commented on a hashtag on American politics in English (*Who in their right minds would vote for this?? <HT>#dumptrump</HT>*). A similar example was produced in Denmark, where a football fan in Danish noted their satisfaction that a player would not be available against their team the next day (*godt tilfreds med at Højbjerg ikke et med mod <AT>@...</AT> imorgen*), while expressing their support to the team's injured Serbian goalkeeper in English (*I wish all the best for <AT>@...</AT> what an unfortunate injury, get well soon.*). The goalkeeper himself chooses to tweet in English. More detailed analyses of language choice in individual speakers will be carried out in the future.

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 3). Its share varies between the lowest share (26%) in Finland and the highest (46%) in the Danish data. Table 3 shows the five most frequently used languages. It excludes the tweets with unidentified language codes.

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). At this pilot

stage of data streaming, we do not know if these differences are reflections of real variability or whether they have been brought about by technical factors.

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%)

Table 3: The top-5 language per five countries in the NTS

It is noteworthy how the presence of a language in this list can give insights to regionally-relevant languages. In addition to English (among the top-5 in all five countries), Swedish (4/5) and Spanish (4/5) appear among the most used languages. We strongly feel that the material can in the future be used to study a range of topics not limited to English but foresee that the study of regionally-relevant languages could also be developed.

Figure 7 visualizes the shares of English relative to the main language in each of the five countries. At this stage it is not possible to determine why there are such considerable differences between the language choices in the Nordic countries with, for instance, Finnish being twice as frequent as English in Finland, and English being used substantially more than Danish in Denmark.

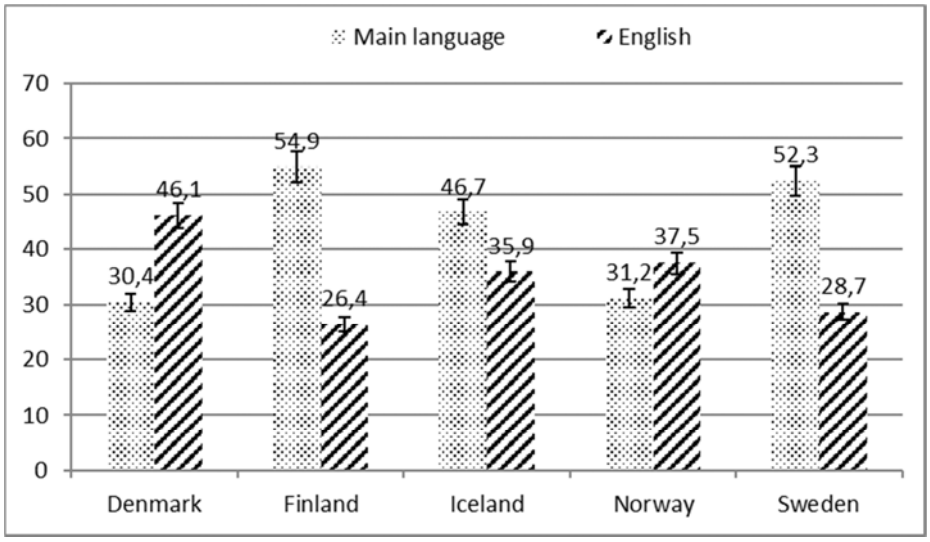

Figure 7: The proportions of the main language (L1) and English in the data.

The marked differences in the proportions of English and the main L1 are conceivably due to either more foreign nationals residing – and tweeting in English – in Denmark and Norway than in Finland, Iceland and Sweden, or to Danes and Norwegians choosing to use English more often than Finns, Icelanders and Swedes do. In order to test these hypotheses, we randomly selected 50 English tweets from Denmark, which has the highest proportion of English and the lowest of the main L1, and Finland with the most L1 tweets and the least English tweets. Needless to say, it is precarious to try to identify the users behind Twitter accounts and to pinpoint their linguistic backgrounds. However, it is possible to establish the identities and first languages of a sizeable proportion of the users with some degree of certainty. Some users try to hide their identities, but for many users this would be counter-productive, since the accounts are at least partly used for (semi-)professional purposes. It has been noted that identity-play is still a part of internet culture, but that the possibilities of uploading videos and images has increased users' willingness to portray their real offline lives, rather than fantasies [Sloan et al., 2015].

In order to identify the users, we combined the names given by them, their self-descriptions and the information gleaned from their blogs, homepages and other sources. The names given include many stereotypical Danish and Finnish names, but also some clearly foreign ones. Many self-descriptions hint at the first languages of the users (e.g., one user presenting himself as a Danish storyteller, two define themselves as expats, another as a videoblogger from Finland) and a homepage shows that one account belongs to a Finnish rock band.

As in [Sloan et al., 2015] the occupations given would seem to suggest that Nordic Twitter users of English are often employed in the creative industries (e.g., journalists or graphic designers) or as professionals (e.g., researchers or sales directors). However, as [Sloan et al., 2015] point out, there is no foolproof way of ascertaining whether the information given is correct, but such problems are also encountered in offline collections of sociological variables.

Keeping the caveats in mind, it is nevertheless striking that the 50 English tweets from Denmark and the 50 from Finland seem to stem from similar language backgrounds: about half are produced by people who appear to be L1 speakers of Danish and Finnish respectively, a quarter are produced by foreign nationals residing in or temporarily visiting the countries and a quarter remain unidentified. Based on this small sample, admittedly with methodological weaknesses, there is no reason to assume that the populations tweeting in English in Denmark and Finland are very different in their compositions. There just seem to be more of them in Denmark.

4.2 Using the NTS for studying discourse phenomena

Lastly, we take a brief look at one subsection of the material, i.e. the tweets written in English, and investigate its discourse properties, operationalizing them using one feature that is used to investigate oral discourse style [cf. Scheffler, 2014; Knight et al., 2014]. Our illustration focuses on English modal auxiliaries and their idiomatic alternatives, known as semi-modals. For many English core modal elements, like *must* or *will* (*You must do this*, etc.), there is an equivalent alternative, HAVE *to* (*You have to do it*) and BE *going to* (*He will do it* vs. *He's going to do it*). Note that the items with a capital letter indicate lemmas in which all possible inflected forms are included. The story of these semi-modals is such that they are latecomers to English,

and they are typically more frequent in spoken styles. Four of the semi-modals also have even more spoken-like variant forms, which represent the latest additions to the feature pool. So, HAVE *to* can be further contracted to *hafta*. For WANT *to* one can write *wanna*; BE *going to* + verb can be replaced by *gonna*, and HAVE *got to* by *gotta*.

In a previous study [Laitinen, 2016], we have shown that tweets as a genre behave unexpectedly when it comes to the uses of core and semi-modals. They do not show high frequencies of more informal semi-modals only, but writers often opt for the contracted forms (i.e. *gonna*, *gotta*, etc.). Our initial explanation for this result is that the 140-character limit in tweets imposes a restriction so that often longer semi-modal idioms are not used as frequently as in spoken or other spoken-like written texts. This economy restriction is also reflected in the higher share of univerbated forms, *viz*. *wanna*, *gonna*, etc. Table 4 compares the proportions of four semi-modals in 2,142,861 tweets containing circa 20 million words in the subsection of NTS tagged as English (lang=en). It shows the total frequencies of the search words and their univerbated forms. It also includes the absolute frequencies of these shortened forms and their proportional share. The results show that with the exception of HAVE *to* (*hafta*) the shares of these univerbated forms are remarkably high, over 50% for *gonna* and *gotta*.

Type	**Total frequency**	**Univerbated forms**	**Univerbated (%)**
HAVE to	13,242	Hafta	3 (0%)
WANT to	29,848	Wanna	9,184 (31%)
BE going to	26,974	Gonna	14,063 (52%)
(HAVE) got to	6,068	Gotta	3,805 (63%)

Table 4: Total frequencies of four modal types and their contracted forms in the Twitter subcorpus

The results indicate that tweeters not only use overtly speech-like forms, but they skip over one natural stage in the development of language. A meaningful point of comparison comes from the *Corpus of Contemporary American English* (COCA) (see http://corpus.byu.edu/coca/) and its spoken component for 2010–2015. COCA is a mega corpus of circa 520 million tokens, and we selected the spoken component since spoken AmE is the most advanced variety for many ongoing grammatical changes in English. In addition, its size is comparable with the tweet corpus, as it contains circa 20 million words of spoken American English.

The proportions of the univerbated forms in COCA confirm our initial explanations based on the tweet data. In COHA, the highest share is with *gonna* + any verb 10% (1,825 instances out of 18,625), and the shares of *wanna* and *gotta* are substantially lower (4% and 3% respectively), and *hafta*, according to COCA is non-existent.

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 four months of the streaming process, which started in spring 2016. The objective is to continue the streaming for several years, thus updating the corpus with over 35,000 tweets per day. The data collection is taking place on a two-layered model in which we limit ourselves to geotagged tweets in a specified geographic region, and we hope to expand the method to new regions.

This results in a multilingual corpus, and it is the outcome of a cross-disciplinary collaboration of a computer scientist and a group of sociolinguists. Our collaboration aims at better methodological accuracy in collecting new types of data for social sciences and humanities, and it builds on making the best use of the big and rich data for research with high theoretical relevance in sociolinguistics, as illustrated for instance in [Laitinen et al. 2017]. We have illustrated how new forms of computer-mediated communication data can lead to new insights in humanities and increasingly also in sociolinguistics and its applications. The corpus will be made available in a format that complies with the terms of service, and we foresee that it can be used as a source of empirical evidence in a range of subfields in sociolinguistics.

References

[Barbaresi, 2016] Barbaresi, A.: Collection and indexation of Tweets with a geographical focus. Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 2016. *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC)*, (2016) 24–27. <hal-01323274>

[Bollen et al., 2011] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market. *Journal of Computational Science*, 2, 1 (2011), 1–8.

[Bolton and Meierkord, 2013] Bolton, K. and Meierkord, C.: English in contemporary Sweden: Perceptions, policies, and narrated practices. *Journal of Sociolinguistics* 17, 1 (2013), 93–117.

[Bradley, 2016] Bradley, L.: The Mobile Language Learner – Use of Technology in Language Learning. *Journal of Universal Computer Science* 21, 10 (2016), 1269–1282. doi: 10.3217/jucs-021-10-1269.

[Campbell, 2011] Campbell, D. G.: *Egypt Unsh@ckled: Using Social Media to @ # :) the System: how 140 Characters Can Remove a Dictator in 18 Days* (2011). Llyfrau Cambria/Cambria Books.

[Coats, 2017] 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. (2017), (Studies in Variation, Contacts and Change in English 19). http://www.helsinki.fi/varieng/series/volumes/19/. (Accessed 15 Jan 2018).

[Davies, 2015] Davies, M.: Corpora: an introduction. In: Biber, D. and Reppen, R. (eds.), *The Cambridge Handbook of English Corpus Linguistics* (2015), 11–31. Cambridge: Cambridge University Press.

[Eisenstein et al., 2012] Eisenstein, J, O'Connor, B., Smith, N.A. and Xing, E.P.: Mapping the geographical diffusion of new words. arXiv:1210.5268v3 [cs.CL].

[García Laborda et al., 2015] 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 (2015), 114–123.

[Gayo Avello, et al., 2011] 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* (2011). Association for the Advancement of Artificial Intelligence.

[Graham et al., 2013] Graham, M., Hale, S.A., Gaffney. D. Where in the world are you? Geolocation and language identification in twitter. The Professional Geographer 66, (2013), 568–578. doi 10.1080/00330124.2014.907699

[Huang et al., 2016] Huang, Y., Guo, D. Kasakoff, A. and Grieve, J.: Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59 (2016) 244–255. doi:10.1016/j.compenvurbsys.2015.12.003.

[International Telecommunications Union, 2015] http://www.itu.int/net4/ITU-D/idi/2015/

[Knight et al., 2014] Knight, D., Adolphs, S. and Carter, R.: CANELC: Constructing an e-language corpus. *Corpora* 9, 1 (2014), 29–56.

[Kohnen and Mair, 2012] Kohnen, T. and Mair, C.: 2012. Technologies of communication. In Nevalainen, T. and Traugott, E. C. (eds.), *The Oxford History of the English Language* (2012), 261–284. Oxford: Oxford University Press.

[Laitinen, 2016] Laitinen, M.: 2016. Ongoing changes in English modals: On the developments in ELF. In Timofeeva, O., Chevalier, S., Gardner A.-C. and Honkapohja A. (eds.), *New Approaches in English Linguistics: Building Bridges* (2016), 175–196. Amsterdam: John Benjamins. doi: 10.1075/slcs.177.07lai.

[Laitinen and Levin, 2016] Laitinen M. and Levin M.: On the globalization of English: Observations of subjective progressives in present-day Englishes. In: Seoane, E. and Suárez-Gómez, C. (eds.), *World Englishes: New Theoretical and Methodological Considerations* (2016), 229–252. John Benjamins, Amsterdam. doi 10.1075/veaw.g57.10lai.

[Laitinen et al. 2017] 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, (2017), 303–325. Amsterdam: John Benjamins. doi 10.1075/ahs.7.12lai.

[Leppänen, et al., 2011] 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 (*2011). <http://www.helsinki.fi/varieng/series/volumes/05/>

[Lundberg et al., 2018] Lundberg, J., Nordqvist, J. Matosevic , A. On-the-fly Detection of Autogenerated Tweets, arXix preprint.

[Mair, 2013] Mair, C.: The World System of Englishes: Accounting for the Transnational Importance of Mobile and Mediated Vernaculars. *English World-Wide* 34 (2013), 253–278.

[Morstatter, et al., 2013] 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, (2013), 400–408.

[Scheffler, 2014] Scheffler, T.: A German Twitter Snapshot. *Proceedings of LREC*, (2014), 2284–2289.

[Schneider, 2014] Schneider, E.W.: New reflections on the evolutionary dynamics of world Englishes. *World Englishes* 33, 1 (2014), 9–32.

[Sloan et al., 2015] Sloan L., Morgan J., Burnap P. and Williams M.: Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLoS ONE* 10, 3, (2015), 1–20. e0115545. doi:10.1371/journal.pone.0115545

[Tagliamonte, 2012] Tagliamonte, S.: *Variationist Sociolinguistics: Change, Observation, Interpretation*. (2012), London: Blackwell.

[Tumasjan et al., 2010] 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 (2010), 178–185.

[Wang et al., 2016] Wang, H.-W., Wu, Y.-C. J. and Dong, T.-P. Exploring the Impacts of Social Networking on Brand Image and Purchase Intention in Cyberspace. *Journal of Universal Computer Science* 21, 11 (2016), 1425–1438. doi: 10.3217/jucs-021-11-1425.