Language Choice and Gender in a Nordic Social Media Corpus

Abstract

This study analyzes language choice, bi- and multilingualism, and gender in a corpus of more than 22 million Twitter messages by almost 36,000 authors from the Nordic countries and territories. Author location, gender, and tweet language are disambiguated using a novel method. Three principal findings are discussed: First, gendered preference for particular languages in the Nordics can be explained in part by patterns of gendered migration. Second, a distinct geographical pattern of female/male preference for the national languages of the region and for English is evident for users who are likely L1 users of a Nordic language: Females are more likely to use English, while males are more likely to use a Nordic language. Third, while high rates of bi- and multilingualism are found, males are more likely to be bi- or multilingual in all the Nordic countries/territories. The latter two findings are interpreted in light of sociolinguistic considerations as evidence for incipient language shift towards English for Nordic users on the Twitter platform.

Keywords: Linguistic diversity, Computer-mediated communication, Social media, Language and gender, Twitter, Bilingualism, Multilingualism

1. INTRODUCTION AND BACKGROUND

Recent years have witnessed increased use in Nordic societies of computer-mediated communication modalities such as instant messaging apps, pseudo-anonymous image boards, or social media platforms such as Facebook and Twitter (Audience Project 2016, Christensen 2017, NRK 2015, Sperstad 2018). At the same time, local environments in the Nordics have become more bilingual with English and/or multilingual, due to educational policies, the availability of foreign-language media, international mobility, and demographic shifts, among other reasons (Linn 2016, Björklund et al. 2013, OECD 2018). The shift towards increased use of English evident in most domains in Nordic societies (Görlach 2002, Linn 2016) necessitates investigation into its prevalence and extent as well as the contexts in which English will be used rather than a local language, but to date, few or no studies have collected large samples of naturally occurring multilingual language data from the Nordics in order to analyse language choice. In situations in which language shift is underway, gender-based differences are frequently evident (Labov 1990, 2001; Trudgill 1998); thus, an analysis of language choice according to gender may shed light on incipient language shift.

In this study, data from the social media platform Twitter is used to analyze language choice, bilingualism with English, and multilingualism in the Nordic countries and territories according to national location and user gender. Language on a social media platform is not necessarily representative of overall patterns of use, and the national languages in the Nordics are not in danger of disappearing, whether in online or in other contexts. Nevertheless, a high degree of linguistic diversity can be demonstrated for Twitter in the Nordics, and bilingualism with English is the norm. In terms of male and female language choice, differences are apparent, particularly with regard to use of English by presumable L1 Nordic speakers.

Work in the sociolinguistic tradition has proposed a link between gender identity and the frequency of use of phonological, lexical, or grammatical features, with females found to make more use of features that index standardness and prestige (e.g. Trudgill 1974; 1998, Labov 1990; 2001, Cheshire 2002). A similar gender-based pattern is evident for some types of language shift: sociolinguistic fieldwork and survey data have been used to show that males and females can have different attitudes to languages in diglossic environments, with females typically showing a more positive orientation towards the more prestigious, supra-local language (Gal 1979, Bilaniuk 2003, Lai 2007, Smith-Hefner 2009). To date, there have been relatively few large-scale corpus-linguistic studies in which gender, extent of bi-or multilingualism, and language choice have been analyzed, and none for the Nordic countries. Access to online language use via the APIs (*Application Programming Interfaces* – for websites, endpoints from which users can consume data) of social media sites can remedy this deficit. This study investigates online bi- and multilingualism in the Nordic countries by means of a quantitative analysis of Twitter messages. Three principal research questions are posed: First, how does the use of national languages, English, and other languages vary according to national context in the Nordics? Second, which languages are more used by males and females? And third, to what extent are females and males in the Nordic countries/territories bi- or multilingual on the platform? It can be demonstrated that the official languages of the region and English play central roles in connecting persons within the Nordic countries and territories on the Twitter platform, but gender-based differences in language preference and bi- or multilingualism are evident. These can be interpreted in light of demographic and sociolinguistic factors.

The paper is organized as follows: In Section 2, some previous work pertaining to the language situation of the Nordic countries (especially in regards to English), language choice and gender, and language diversity, bi- and multilingualism on Twitter is reviewed. In Section 3, a description of the methods used for the collection and filtering of Twitter data is provided: Following an outline of the metadata structure of a tweet, methods for source filtering, location inference, gender inference, and language detection are described. In Section 4, the results are presented: First, aggregate language use by country/territory and gender is considered. Second, the tendency for females and males to use a particular language is quantified using an odds ratio. A subset of the data provides insight into active bilingualism with English by users whose presumable L1 is a Nordic language. Third, the proportion of users who are bi- or multilingual is calculated by country/territory and gender. Section 5 interprets the findings for the

Nordic languages and for English in terms of sociolinguistic factors, and for other languages in terms of migration statistics, using data from the Organization for Economic Co-operation and Development (OECD). A moderately strong correlation between gendered migration and gendered language preference is found. In the conclusion (Section 6), some caveats are offered, as well as suggestions for future work.

2. PREVIOUS WORK

Contemporary Nordic societies are multilingual, a result of historical factors, national educational policies, and immigration (Björklund et al. 2013). Many inhabitants of the Nordics are essentially bilingual with English, and research into English use in the Nordics has investigated its status as a lingua franca, the extent of its use in various media or communicative contexts, and attitudes of speakers towards its use, mainly on the basis of survey data (an overview of some of the literature is provided in Linn 2016). For example, in Iceland, the majority of Icelanders are exposed to English every day, while 21% of Icelanders report speaking English daily (Arnbjörnsdóttir 2011). Younger Icelanders and Faroese have developed a bilingual identity (Jeeves 2011, Mortensen 2011). Norwegians are reported to be essentially diglossic (Rindal 2010, Rindal and Piercy 2013). By 1999, 47% of Danes reported using English at least once a month, and in 2003, 90% of Danes aged 15–21 reported using English at least once a week (Preisler 1999, Andersen 2003, both as cited in Lønsmann 2009:1139). For Sweden, Bolton and Meierkord attest that while Swedish remains the "preferred language... in most domains" (2013:93), English is dominant in academia and business. Similar findings are reported for Finland in the results of an extensive survey into the use of English in the country: English has become "a language used in many domains and settings within Finnish society" (Leppänen et al. 2011:16).

Insight into the role of English in the Nordics is provided by the MIN project (*Moderne importord i språka i Norden*, ‘Modern import words in the languages of the Nordic countries’), a series of qualitative and quantitative studies that investigated the prevalence of (mostly English) loanwords in Nordic languages as well as attitudes of language users (Sandøy 2003, Vikør 2003, Kristiansen and Sandøy 2010a). A survey-based study found that compared to Swedes, Danes, and Finns, Icelanders and Norwegians have slightly more negative attitudes towards the “influx of English” and the use of English words in local languages, although higher-status social groups have more positive attitudes towards English in all the Nordic countries (Thøgersen 2004, see also Graedler 2014). In terms of attitudes towards the local national languages, western and eastern peripheries of the Nordics (Iceland, the Faroes, Swedish-speaking Finland and Finland) exhibit more linguistic purism, while Norway, Denmark, and Sweden show more openness to integration of foreign-language material (Kristiansen and Sandøy 2010b).

2.1 *Language and gender*

Relationships between language use and gender identity have been extensively researched in the fields of sociolinguistics, education, sociology, and computational linguistics. Early studies, for example Lakoff (1973), posited a “deficit hypothesis” in which gender differences in language use could be understood as reflecting the different social status of females and males. Some early studies in this tradition were based not on empirically collected data, but researcher intuition. Much research has been conducted upon gender-based differences in language use in the tradition of the classic variationist sociolinguistic paradigm (for an overview, see Cheshire 2002), but research into language choice according to gender has not been as extensive.

Measures of use and attitudes towards local and non-local languages or language varieties have shown gender differences, with females typically more open to non-local languages. Gal (1979) reported higher male bilingualism in a study of a traditionally Hungarian-speaking community in Austria as a result of females having already shifted towards the language with higher prestige in the supra-local national environment (in this context, German). Woolard (1997) studied social networks and language choice among bilingual (Spanish-Catalan) high school students in Barcelona, and found that females tended to have more homogenous networks than males, and hence lower rates of active bilingualism. Bilaniuk (2003) found that females in Ukraine exhibited more positive attitudes towards the use of Russian and English compared to males, who exhibited more positive attitudes towards the use of Ukrainian. The finding was interpreted as indexing the languages' relative usefulness in terms of economic and social advancement. In a similar manner, Smith-Hefner (2009) reported young women in Java, Indonesia, to be more likely than young men to use the standardized national Indonesian language rather than Javanese. In a study of attitudes towards local and foreign languages among 1,048 pupils in Hong Kong, Lai (2007) found females to have more positive attitudes towards the supra-local languages Mandarin and English, and males to have more positive attitudes towards the local language Cantonese.

Survey-based research provides some insight into knowledge and use of foreign languages according to gender in the Nordic countries. The Adult Education Survey of the European Union was conducted in 2007, 2011, and 2016 (European Commission 2018). In 2016, survey respondents were asked to report how many foreign languages they knew. The percentage of males and females reporting knowledge of at least one foreign language was similar according to gender in the Nordic countries (Norway males 92.1%, females 92.1%; Denmark males 95.3%, females 96.3%; Sweden males 96.4%, females 96.8%; Finland males 89.4%, females 94.7%). In an additional question, respondents were asked to rate their proficiency in their best-known foreign language. Possible responses were "basic", "good", and "proficient". For males, the percentages of those reporting "proficient" knowledge of the best-known foreign language were 47.6% for Norway, 44.4% for Denmark, 61.8 % for Sweden, and 37.4% for Finland. For females, the corresponding percentages were 44.8%, 37.9%, 57.5%, and 31.0%.

In Finland, detailed information about reported knowledge and use of English and other languages is available in Leppänen et al. (2011). Finnish males are more likely than females to consider themselves bi- or multilingual, while Finnish females are more likely to consider themselves monolingual (2011:Table 6a.1. The p-values for these differences, however, suggest that they are not significant). Female Finns are more likely than male Finns to consider themselves fluent speakers, writers and readers of English, as well as report being able to understand spoken English without problems (2011:Tables 22a.1, 22b.1, 22c.1, 22d.1). However, males are more likely to report reading various genres of English-language texts, such as newspapers, magazines, comics, professional literature, web pages, manuals, and emails. Females report more reading of fiction and other literature in English (2011:Table 28.1).

2.2 *Twitter language and multilingualism*

Studies of computer-mediated communication (CMC), for example language use on Twitter, have investigated phenomena such as the discourse functions of hashtags (Wikström 2014, Squires 2015), lexical innovation (Eisenstein et al. 2014), African-American Vernacular English (Jørgensen et al. 2015), grammatical variation in English-language Twitter from Finland and the other Nordic countries (Coats 2016; 2017a; 2017b), or the interaction between demographic parameters such as gender with lexical and grammatical features in American English (Bamman et al. 2014).

Gender can usually be assigned in a straightforward manner when directly observing speakers or working with survey data in which respondents report their gender, but for much CMC data, the self-identified gender of a particular user is not necessarily made explicit. The Twitter platform does not provide a metadata field in which users are asked to specify their gender. In such cases, algorithms may be applied in an attempt to automatically identify author gender (or other identity parameters). Tweets from authors who have been manually annotated for gender can be used to train classifiers (Rao et al. 2011, Kokkos and Tzouramanis 2014, Volkova et al. 2015). Burger et al. (2011) classified user gender based on text from blogs linked in tweet metadata. Gender can also be disambiguated by comparing a Twitter user's "author name" metadata with name frequencies according to gender in governmental records (Mislove et al. 2011) – such an approach is applied in this study. Some recent work has acknowledged social constructionist approaches by noting that while gender can be disambiguated as a binary, category membership based on shared topics may better model the ways in which users represent themselves online (Bamman et al. 2014).

In the context of the study of online language variation and linguistic diversity, many studies have examined individual practices or those of a handful of users, but large-scale approaches are relatively few. As Lee noted, many studies "tend to follow the discourse analytic or interactional sociolinguistic approach" (2016:128). Examples of Nordic-oriented studies along these lines include Leppänen et al. (2009), Stæhr and Madsen (2014), or Kytölä and Westinen (2015).

While few corpus-based studies of multilingual social media have been undertaken using Nordic data, multilingualism on Twitter has been addressed in some other studies. Some of these bear upon the present research in terms of their methodology, primarily in that they analyse patterns in large corpora of Twitter language. Such studies can give insight into the linguistic behavior of groups of users at a macro level, or demonstrate how languages themselves are utilized as resources in multilingual, global contexts. Hong et al. (2010) found different patterns of hashtag, username, and URL use, as well as retweets and responses to others according to language community. They showed that language communities on Twitter differ in their aggregate behavior, with, for example, Korean language users much more likely than Indonesian language users to reply to tweets by others. Ronen et al. (2014) compared the worldwide influence of languages by analyzing networks of bi- and multilingual book translations, Wikipedia author editors, and Twitter users, and found that English plays an important central role. Hale (2014) investigated global multilingual networks on Twitter, including the network associations of retweets and user mentions, and found that while most interaction networks are language-based, and English is the most important single mediating language, other languages collectively represent a larger bridging force. Graham et al. (2014) found that for tweets containing GPS metadata collected globally, automatic language detection is not always accurate, and user location as indicated in the user profile does not always correspond to GPS metadata. Eleta and Golbeck (2014) examined the tweets of 92 multilingual Twitter users and showed that their language choice on Twitter reflects the predominant language of their social networks. While it has been found that users of less represented languages are more likely to switch languages and that English has become the central mediating language, the interaction of multilingualism with gender in a large social media data set has not yet been subject to research attention.

Establishing the bi- or multilingualism of a social media user can be difficult: Online, users may make use of different languages than they do in face-to-face spoken interaction, or they may consume social media content in a second or third language, but not actively author posts in that language. In this study a bi- or multilingual user is defined by setting a quantitative cutoff based on an active-use criterion that is in line with Grosjean's broad definition of of a bilingual as someone who "uses two (or more) languages (or dialects)" (2008:34). The criterion is described in Section 4.

3. DATA AND METHODS

Twitter provides various endpoints to its APIs. Default access to the Streaming API, which returns tweets as they are broadcast in real time, is 1% of the data stream volume, although commercial partners can gain access to higher data volumes. In order to create a corpus of "seed" data from which Nordic users could be identified, tweets with “place” metadata were collected globally from the default Streaming API from November 2016–June 2017 using the Python module *Tweepy* (Roesslein 2015). Retweets were excluded. The metadata of these tweets (approximately 653 million) were matched with regular expressions for Nordic locations and male and female names (see below). User timelines (up to 3,250 tweets) of matching users were then downloaded from the Twitter REST API in November 2017 for users who could be located within the Nordic countries and assigned gender.

3.1 *Filtering and localization*

Some tweets are generated automatically by apps or bots that interact with the Twitter APIs (Haustein et al. 2015). Messages sent by apps and bots, which often include automatically generated text content, were filtered using the “source” metadata field to exclude tweets not sent by the eight following apps: *Twitter Web Client*, *Twitter for iOS*, *Twitter for iPhone*, *Twitter for Android*, *Twitter for Windows Phone*, *Twitter for Instagram*, *Tweetbot for iOS*, and *Tweetbot for iPhone*. Tweets with these sources collectively comprised over 87% of the Nordic tweets overall.

Tweet metadata can include three types of location information. Many Twitter users provide their home location in the “location” field within the “user” entity upon registering their profile. The value can be any sequence of Unicode characters, and can be changed at any time. Tweets composed on GPS-enabled devices such as smartphones can contain “geo” metadata: latitude-longitude coordinates corresponding to the exact location of the device when the tweet was broadcast. In order to broadcast exact location with tweets, a user must activate this option. The third type of location metadata is the “place” field. When composing a tweet, a user can optionally choose to add a “place” to a tweet by clicking on a button adjacent to the text input window. The button returns a list of place names based on the IP address of the service being used to access Twitter, or the user can select a place from Twitter’s internal place dictionary, which is accessed via text input. “Place” metadata is a type of point-of-interest metadata used by many online services (for more information, see Hochmair et al. 2018). Twitter "place" metadata contains a place name, a country code and country name, and an array of the four latitude-longitude coordinates which form the boundaries of the box that encloses the place in geographical space.

Relying on a single type of place metadata can pose methodological problems: Relatively few tweets contain the "geo" or "place" metadata (Leetaru et al. 2013, Laylavi et al. 2016), and the values of these metadata fields, when present, may not correspond to a user's home location, for example because the user is travelling or has tagged a tweet with a "place" to indicate tweet topic, not user location (e.g., "There has been an earthquake in Japan", "The Eurovision song contest was in Stockholm"). Users can also select "place" values for humorous or other reasons.[1] Inducing geolocation using more than one type of location metadata can achieve higher precision compared to relying on the "place" attribute or exact geolocation alone (Schulz et al. 2013, Ajao, Hong, and Liu 2015).

For these reasons, a multi-step method was used to infer user location. First, the location field in the "user" entity of each tweet in the seed data was matched using a dictionary of 1,627 place names in the Nordic nations and territories of Greenland, Iceland, the Faroe Islands, Norway, Denmark, Sweden, Åland, and Finland (available at https://github.com/stcoats/Nordic-Place-Names/). For each country or territory, a list of all the municipalities of the country (e.g. "Oslo", "Helsinki") and sub-country-level units of administration according to ISO 3166 (e.g. "Norðoyar", "Skåne"), were combined with the names of the countries/territories in English and the principal Nordic languages. For Finnish places with more than one official language, all official place names for that locality were collected (i.e. in Finnish, Swedish, and Sámi).

Some places in the Nordics have names that can also refer to other places: For example, *Alta* is the name of a town in Norway, but also a place in Utah, United States. *Gran* is likewise a place in Norway, but also an element in the name of the Spanish island *Gran Canaria*. To prevent false positives, such items were modified in the list by specifying the Nordic country or territory in the regex (e.g. *Gran* was changed to *Gran, Norway*; *Gran, Norge; Gran, Noreg;* and *Gran, Norja*). Tweets producing matches were then disambiguated for author gender. Gendered tweets with consistent "location" and "place" metadata were retained.

3.2 *Gender disambiguation*

In the next step, the gender of users with a Nordic place location in the "user" metadata was disambiguated on the basis of name frequency information provided by the statistical offices of the Nordic countries (Statistics Greenland 2017; Statistics Iceland 2017; Danmarks Statistic 2015a, 2015b; Statistics Norway 2017a, 2017b; Statistics Sweden 2016a, 2016b; Avoindata.fi. 2017). Name and gender frequency information was obtained for 58,874 given names from Greenland, Norway, Denmark, Sweden, and Finland. For Greenland, the list consisted of all names assigned at least 5 times in total to male or female newborns in the years 1910–2011. For Iceland, name lists of the most frequent single and double names in 2017, 2016, and 2008 were aggregated. For Norway, the list consisted of all names assigned to male or female newborns at least 4 times in the years 2008–2017 or 2006–2017, respectively. For Denmark, the list consisted of all names occurring 4 or more times in the registered population of Denmark as of 1.1.2016. For Sweden, all names occurring at least 10 times in the resident population for each of the years 1999–2015 were available. For Finland, the list consisted of all names occurring at least 10 times in the resident population in September 2017. Due to differences in collecting and publishing name information, the number of names available from each country or territory (summarized in Table 1) is quite variable. For example, there are fewer Icelandic names than Greenlandic names, despite Iceland's larger population – this is because the Greenland list is more comprehensive, covering 100 years compared to three select years. In addition, the freely available name data from Iceland consists only of the most frequent names – access to more extensive data requires payment. Similarly, there are far fewer Norwegian names than Danish, Swedish, or Finnish names. This is because the Norwegian data considers only names with a minimum frequency among newborns per year, for a 10-year period, while the Danish, Swedish, and Finnish data consist of names with a minimum frequency in the entire populations – a rare name may not be assigned to a newborn more than 4 times a year in Norway, but the sum of all living persons who have a rare name can easily be much more than 4.

Table 1. Summary of name disambiguation data

Country/ territory	**Male types**	**Female types**
Greenland	648	664
Iceland	252	150
Norway	836	941
Denmark	9,071	11,399
Sweden	11,861	13,446
Åland	(included in Finland totals)	
Finland	5,098	4,610
Totals	27,766	31,210

For each country, the probability that a name in the records is male or female was calculated by dividing the number of times the name was assigned to one gender by the total number of occurrences of the name in the statistical data.[2] Names that were female or male with a probability of ≥ 0.8 were retained in the name set for that country. To create the name lists for the Nordic region, the country name lists were aggregated by gender. Duplicates and names occurring in both male and female lists were removed. In total, this resulted in non-overlapping lists of 17,856 names given to females and 15,406 names given to males. This resource is available at github.com/xxx/xxx.

For each unique user from the seed data who matched a Nordic user location, the value in the "author name" metadata field was matched against the aggregate Nordic name lists using a case-insensitive regular expression.[3] Author names that did not match the regex (e.g. "Acme Customer Service" or "sverigetjej2018") were filtered out. For the matches, user timelines (up to 3,250 tweets) were downloaded from the Twitter REST API. In a final geographical filtering step, each user's tweets were aggregated according to the country of the "place" metadata. Only users whose "place" metadata matched the “location” metadata in the “user” entity at the country level were retained for the analysis. Particularly for the Nordic countries/territories with smaller populations, this filtering step may be necessary in order to counteract the influence of short-term visitors such as tourists in the signal.

3.3 *Language determination*

A consideration of bi- and multilingualism on the Twitter platform critically depends on accurate language detection, but automatic methods present difficulties due to short message length, non-standard orthography, and language mixture. Character sequences in URL addresses, usernames, and hashtags, as well as emoji characters, can create problems for automatic language detection algorithms, as they rarely correspond to character sequences in the lexicons of natural languages. Since March 2013 tweets contain automatically detected language metadata in the “lang” metadata (Twitter 2013). Some languages are inaccurately detected with high frequency: For example, the Twitter algorithm labels a large number of messages from Nordic data as having been written in *kreyól ayisien*, or Haitian Creole, including messages such as *Hlaupabòlustelpan ad horfa à Dòru* or *Oh yes Griezmann I love you ?? #FRAGER*. (cf. Zubiaga et al. 2016). One method for reducing inaccurate language assignation is to compare the results of two different algorithms – accurate identification is likelier when the algorithms are in agreement (Twitter 2015). For an analysis of language use in Nordic contexts, Twitter’s native language detection algorithm is also unsuitable as it does not detect Faroese or Kalaallisut (Greenlandic), as of late 2017. For these reasons, a multi-step approach was taken: First, tweet texts were stripped of usernames, hashtags, emoji characters, and URL addresses. Language was then detected using an implementation of *cld2 (compact language detector 2)*, an algorithm developed for Google's Chrome browser (Sites 2013, see also Lui & Baldwin 2014). Tweets for which Twitter’s algorithm agreed with *cld2* were retained; others were discarded. Tweets detected as Faroese or Kalaallisut (Greenlandic) by *cld2* were assigned that language. Tweets identified by *cld2* as Norwegian Nynorsk were assigned the code for Norwegian.

A manual test of a random selection of tweets showed that the method results in highly accurate language detection (see Appendix A). Tokenization was undertaken using the NLTK Twitter Tokenizer (Bird et al. 2009), the Jieba tokenizer for Mandarin (Sun 2014) and the Tiny Segmenter for Japanese (Hagiwara 2014). A summary of the Nordic data collected using the methods described above is shown in Table 2.

Table 2. Summary statistics

Country	Gender	Users	Tweets	Mean	S.D.	Tokens
Greenland	f	14	3,300	235.71	399.23	51,776
	m	27	6,293	233.07	357.94	90,908
Iceland	f	320	214,067	668.96	751.46	3,233,608
	m	553	445,758	806.07	799.89	6,667,414
Faroe Islands	f	11	4,587	417.00	548.91	74,888
	m	14	6,971	497.93	648.45	110,023
Norway	f	1,981	1,069,848	540.05	599.79	17,549,027
	m	4,084	2,401,746	588.09	616.02	40,269,450
Denmark	f	2,014	1,049,581	521.14	664.38	17,421,923
	m	3,436	1,922,858	559.62	635.72	32,678,226
Sweden	f	5,159	3,955,137	766.65	734.92	62,804,891
	m	9,657	7,590,100	785.97	732.51	120,583,939
Åland	f	2	2,012	1,006.00	1210.57	24,791
	m	9	4,678	519.78	344.63	75,470
Finland	f	3,550	1,859,276	523.74	643.46	26,993,930
	m	5,046	3,300,180	654.02	728.45	48,580,663
Totals		**35,877**	**23,836,392**	**664.39**	**700.87**	**377,210,927**

4. RESULTS

The data were analyzed according to country or territory and gender in terms of overall language use, preference for particular languages, and prevalence of bi- and multilingualism.

4.1 *Aggregate language use by country and gender*

In the first set of results, the proportion of tweets by language was calculated according to country/territory and gender. Aggregate language use according to gender is summarized for the four most-used languages per country/territory in Table 3.

Table 3. Percentage of tweets in the four most-used languages by country/territory and gender

Country	Gender	Most-used languages				
		English	**Danish**	**Kalaallisut**	**Norwegian**	**Other**
Greenland	f	39.4	47.0	13.0	0.5	0.1
	m	65.4	26.2	7.9	0.4	0.1
		Icelandic	**English**	**Spanish**	**Faroese**	**Other**
Iceland	f	69.7	28.5	0.8	0.3	0.7
	m	72.1	26.6	0.1	0.4	0.8
		English	**Danish**	**Faroese**	**French**	**Other**
Faroe Islands	f	68.5	2.4	5.0	24.0	0.1
	m	89.1	9.2	1.4	0.1	0.2
		Norwegian	**English**	**Spanish**	**Swedish**	**Other**
Norway	f	48.9	46.0	1.7	0.8	2.6
	m	58.2	37.4	0.7	0.7	3.0
		Danish	**English**	**Spanish**	**Norwegian**	**Other**
Denmark	f	45.3	50.3	1.0	0.7	2.7
	m	50.7	45.4	0.8	0.7	2.4
		Swedish	**English**	**Spanish**	**Turkish**	**Other**
Sweden	f	72.3	25.3	0.4	0.4	1.6
	m	71.4	26.1	0.5	0.3	1.7
		Swedish	**English**	**Norwegian**	**Spanish**	**Other**
Åland	f	98.2	1.5	0.1	>0.1	0.1
	m	87.2	12.2	0.2	0.2	0.2
		Finnish	**English**	**Swedish**	**Russian**	**Other**
Finland	f	69.3	26.5	2.4	0.7	1.1
	m	68.2	28.2	2.0	0.1	1.5
		Swedish	**English**	**Finnish**	**Norwegian**	**Other**
Totals	f	35.7	31.6	15.9	6.6	10.2
	m	35.1	30.6	14.4	9.1	10.8

In Figures 1 – 4, the y-axis shows the proportion of tweets in a language, while the x-axis shows two-character codes for five official languages of the Nordics (Swedish = sv, Finnish = fi, Norwegian = no, Danish = da, Icelandic = is), and for the five additional languages with the highest overall use in the data set (English = en, Spanish = es, Turkish = tr, Arabic = ar, Russian = ru), ordered according to overall frequency in the data for all the countries/territories.

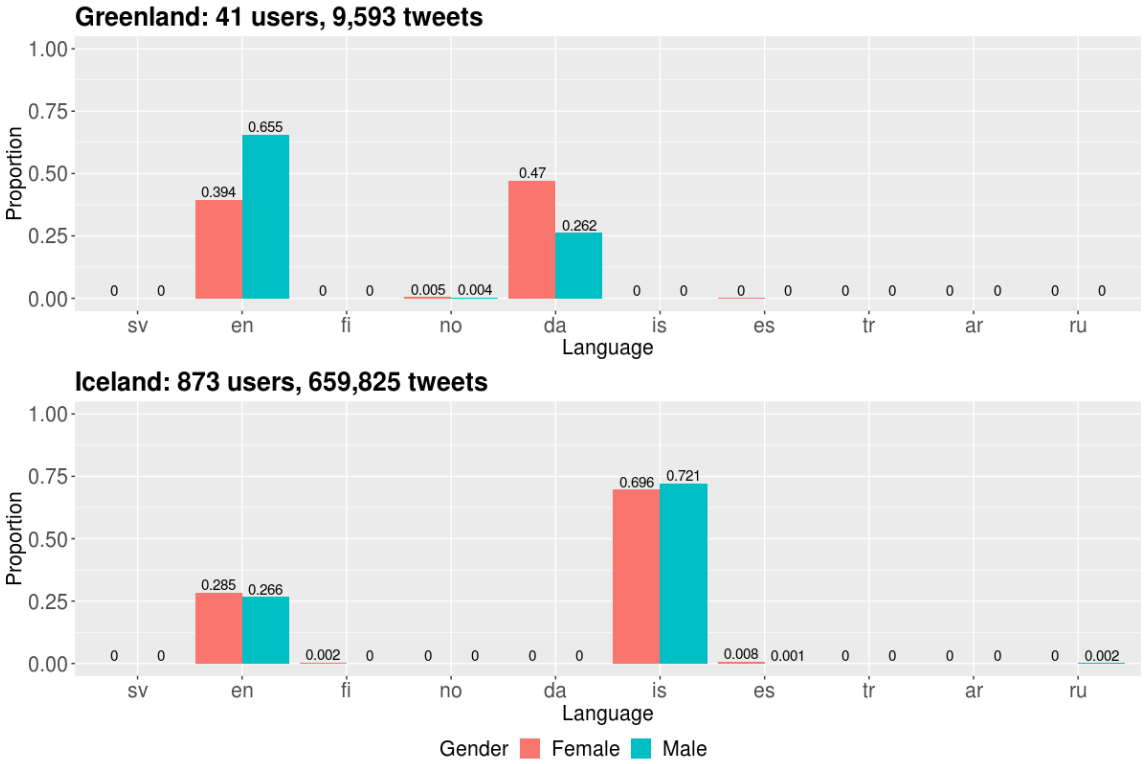

Figure 1: Language use by gender, Greenland and Iceland

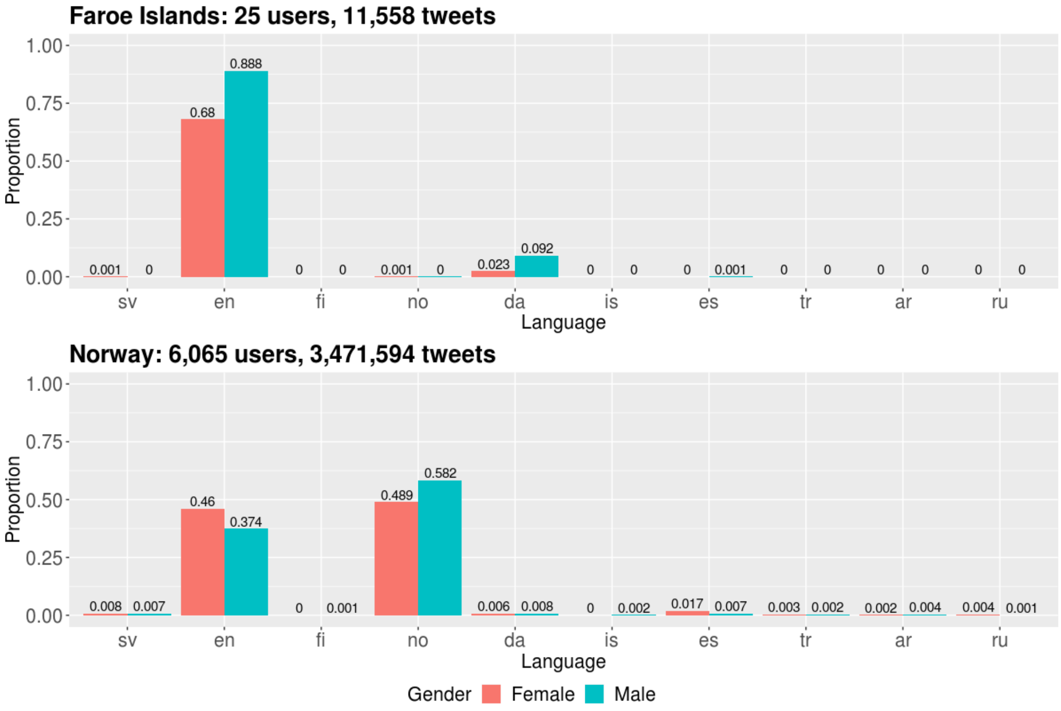

Figure 2: Language use by gender, Faroe Islands and Norway

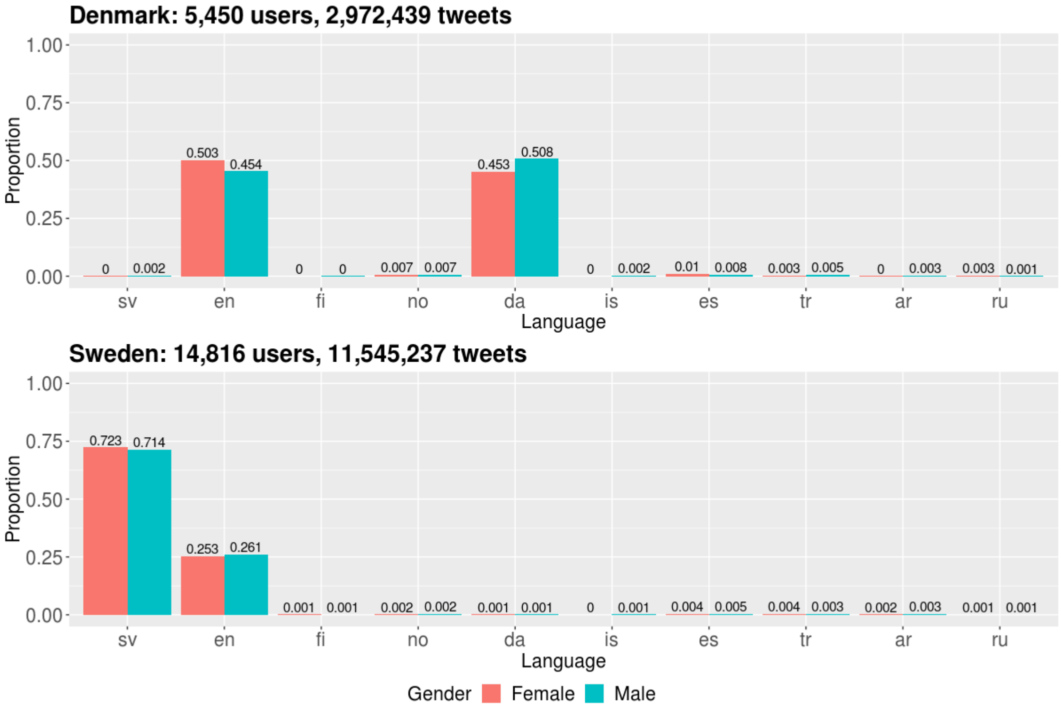

Figure 3: Language use by gender, Denmark and Sweden

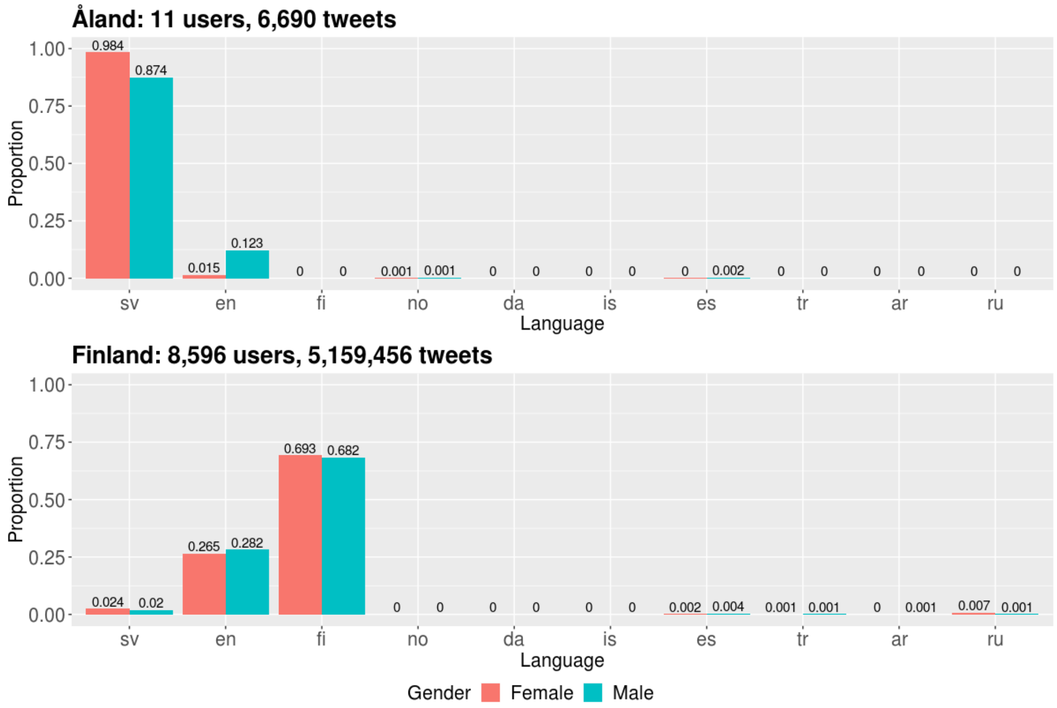

Figure 4: Language use by gender, Åland and Finland

Overall, the official languages of the Nordics and English are the most-used languages. Swedish is the most prevalent language in the data, followed by English, Finnish, and Norwegian. This corresponds to the relative sizes of the user samples: Significantly more users from Sweden were sampled than from the other countries/territories.

Countries/territories with official languages that have only a small population of users, i.e. Greenland and the Faroe Islands, tend to use much more English (or in Greenland, also Danish) than official languages: In these samples, the local official language is used in 10% of tweets or less, while English is much more widely used. The high rate of use of French among Faroes females is due to one user (a French exchange student) with a large number of tweets. In Norway Norwegian is used slightly more than English, while in Denmark Danish and English are used at approximately the same rate. In

Iceland, Sweden and Finland, Twitter users are more likely to write in their local languages: the proportion of tweets in Icelandic, Swedish, and Finnish for these countries ranges from 0.69 to 0.72.

4.2 *Quantifying language preference by country and gender*

In the second set of results, gender-based preference for using a particular language was quantified by calculating the female-male odds ratio for the proportion of tweets in that language. Odds ratio values greater than 1 indicate a language is more likely to be used by females, while values less than one show male preference. Results are shown on a logarithmic scale in Figures 5 and 6, with 95% confidence intervals calculated using Fisher's exact test. In the figures, orange-colored bars indicate an odds ratio > 1 and blue bars < 1. The numerical values immediately above or below the bars along the x-axis indicate the odds ratio (in larger typeface) and the upper and lower bounds of the 95% confidence interval (in smaller typeface). Confidence intervals that do not contain the value of 1 correspond to significance at $p = 0.05$; these are indicated by the language name in black type along the x-axis, while red typeface indicates no significant association. The bars and confidence intervals are not graphically depicted on the plots if they fall outside the plot range of 0.01 to 100.

4.2.1 *All users in the Nordics*

Aggregating the data for all the Nordic countries/territories, the female-male odds ratios are summarized in Figure 5.

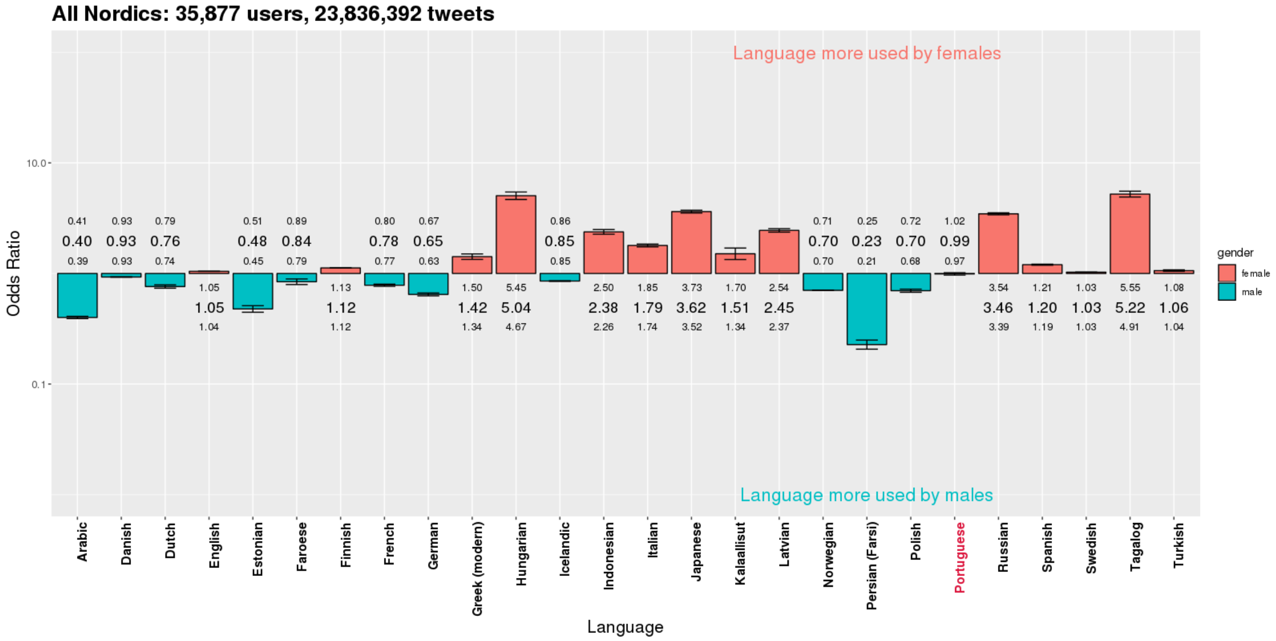

Figure 5: Odds ratios female-male for 26 most common languages, all Nordics

In the aggregated data, the statistically significant "male" languages, in order of their overrepresentation among male users, are Persian, Arabic, Estonian, German, Polish, Norwegian, Dutch, French, Faroese, Icelandic, and Danish. The "female" languages are Tagalog, Hungarian, Japanese, Russian, Latvian, Indonesian, Italian, Kalaallisut, Greek, Spanish, Finnish, Turkish, English, and Swedish. The discussion in Section 5 proposes an interpretation of these results in light of migration statistics and sociolinguistic patterns.

For less widely-used languages in the data, the obtained odds ratio values may reflect language use by relatively few individuals with a large variance in the number of tweets sampled, and hence be unrepresentative of the populations sampled. In addition, as the data represents use of a particular language both by presumable L1 users (i.e. of a Nordic language for most of the sample) as well as by L2 or LX users (e.g. of English for much of the sample), drawing inferences from patterns in the data is complicated by the lack of information about the linguistic identities of the individual users sampled. In the following section, therefore, only English and the principal national language(s) and users who are presumable L1 users of a Nordic language are considered.

4.2.2 *Presumable L1 Nordic language users*

Two additional filtering steps were undertaken in order to analyze language choice by presumable L1 Nordic-language users. In addition to information about the automatically-detected language of the text, the data of a single tweet also contains a field identifying the language in which the Twitter platform is presented to the user (i.e. the language of tabs, buttons, notices, drop-down menus, etc.). The language can be selected by the user; the default setting is that of the service used to access Twitter (e.g. the browser). In a first filtering step, users were selected for whom this field ("user:lang") matched the principal language of the country. Only Norway-, Denmark-, Sweden-, and Finland-based users were filtered, as the Twitter interface is not available in Kalaallisut/Greenlandic or Faroese, and just two of the 1,249 sampled users from Iceland had selected "Icelandic" as the language of their Twitter interface. In a second step, users from each country were matched by name with smaller lists of the 1,000 most frequent male or female name types from that specific country (i.e. not with the aggregate list of 17,856 female and 15,406 male Nordic names).[4] Applying these two filtering steps resulted in a smaller sample of 15,127 users (5,849 females and 9,278 males) who are, presumably, more likely to be L1 users of the principal language of the country in which they are located and more likely to have been born or reside there, compared to the larger sample.

As noted above, except for English and the principal national languages (and Swedish for Finland), less widely-used languages are not evenly dispersed in the sample. For example, the odds ratio value in Figure 5 shows greater female than male use of Hungarian, but only a small fraction of users in the sample are responsible for tweets in Hungarian (0.5% of females and 0.7% of males). The resulting odds ratio reflects the influence of a single Sweden-based female user who broadcast several thousand Hungarian-language tweets.

The tendency for an attribute to be equally spread within a population, dispersion, can be quantified with (among other measures), Julliand's D, which normalizes the coefficient of variation into the range (0,1): smaller values indicate an item is not evenly distributed across categories, while higher values indicate the an item is evenly distributed (Gries 2010). For each of the larger Nordic countries, D values greater than 0.9 were found only for English and for the national language(s). For this reason, the focus in the following is on use of the principal national language(s) and English in the larger Nordic countries of Norway, Denmark, Sweden and Finland.

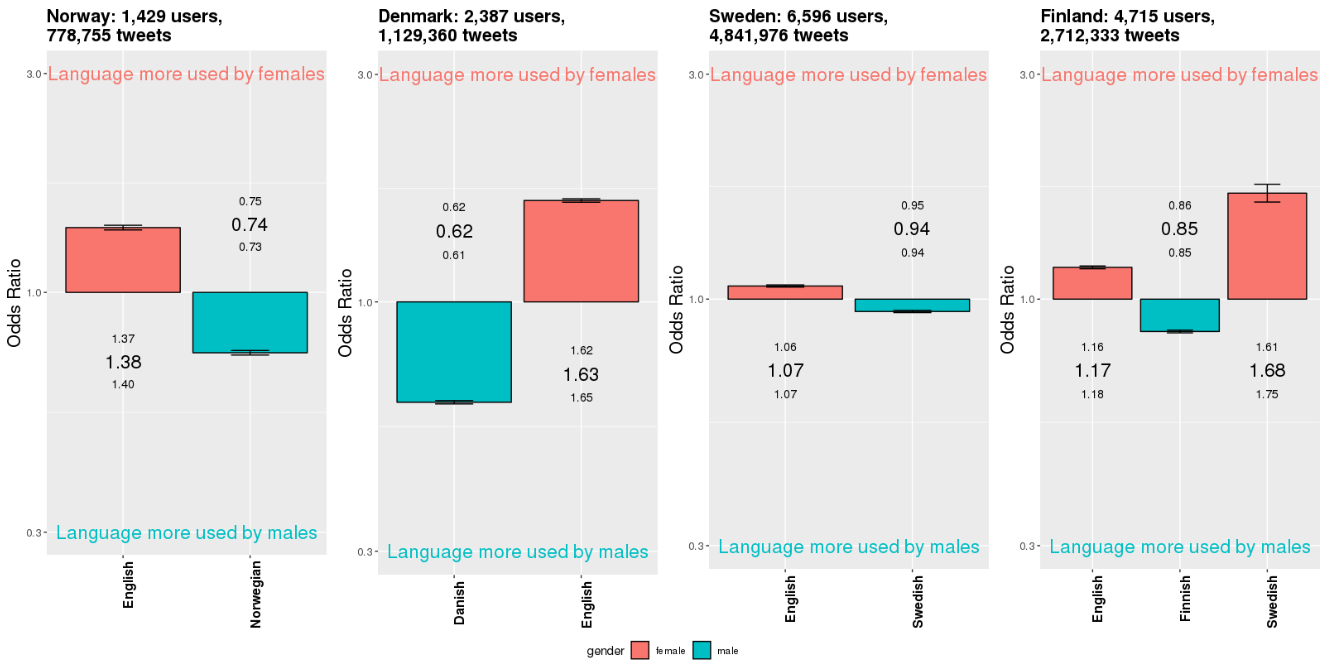

Figure 6: Odds ratios female-male for English and principal national language(s), presumable L1 speakers of the main national language

Figure 6 shows the female-male odds ratios for English and the principal national language for the four larger Nordic countries. In each country, English is more likely to be used by females, whereas the principal national language is more likely to be used by males. In Finland, Swedish is more likely to be used by females. For English, the effect size is largest amongst the presumable L1 Danish users – females are 63% more likely than males to author a tweet in English. A large effect size is also found for Norway users (38%), but less pronounced for Finland (17%) and Sweden (7%). For the national languages, male L1 Danish users are 61% more likely than females to author a tweet in Danish (1/0.62). For Norwegian, the effect size is 35%, for Swedish 6%, and for Finnish 17%. In Finland, female presumable L1 Finnish users are 68% more likely than male users to post a tweet in Swedish.

4.3 *Extent of bi- and multilingualism by country and gender*

In the third set of results, active bi- and multilingualism on Twitter were quantified by country and gender for the presumable L1 Nordic speakers. To counter the possibility of inaccurate language detection, users were determined to have command of a language if they authored at least 10 tweets in that language. By this criterion, approximately 74% of female and 81% of male L1 Nordic-language users were bi- or multilingual. The breakdown by country and gender is shown in Table 4.

Table 4. Percentage bi- and multilingualism by country/territory and gender

Country	Gender	1 Language	2 Languages	3 Languages	4+ Languages
Norway	f	19.97	55.15	22.81	2.06
	m	12.53	50.98	33.64	2.84
Denmark	f	25.61	55.99	17.53	0.87
	m	17.17	59.93	20.92	1.98
Sweden	f	24.39	71.81	3.33	0.47
	m	18.49	76.95	4.15	0.42
Finland	f	28.26	64.74	6.45	0.55
	m	21.96	72.12	5.50	0.41

Unsurprisingly, considering the global nature of social media and the educational attainment level of people in Nordic countries, the majority of users from the Nordics qualify as bi- or multilingual on the Twitter platform. Users from Norway and Denmark were more likely to write in three or more languages on Twitter, compared to Finland- and Sweden-based users. However, this may simply reflect the fact that because Danish and Norwegian (Bokmål) are quite similar in orthography, more than 10 tweets are likely to have been detected in both languages for many users. To test this, Danish and Norwegian were collapsed into a single language; the proportions of bi- or multilinguals from Denmark and Norway were reduced by approximately 2% for both males and females, but otherwise the multilingualism distribution as seen in Table 4 was not greatly affected. Finland's slight edge in Twitter multilingualism over Sweden probably reflects the status of Swedish as an official language in Finland and the fact that a sizeable number of persons in Finland are trilingual in Finnish, Swedish, and English. The higher rate of multilingualism for females in Finland supports this interpretation, given that Swedish is highly favored by females in Finland (Figure 6).

5. DISCUSSION

Language attitudes towards the use of English in the Nordics, as measured in 2002 by the MIN project, correspond to patterns of use in the Twitter data, at least in some respects. For example, Danes and Norwegians hold more positive attitudes towards the use of English as a workplace language compared to Finns, Swedes, Finland Swedes, Faroese, or Icelanders (Vikør 2003:50). For the five larger Nordic countries/territories, Danish and Norwegian users in this study have the highest percentages of tweets in English. On the other hand, the percentage of survey respondents reporting frequent use of English in the previous week was highest in Iceland, followed by Sweden, Denmark, Finland, Norway, the Swedish-majority portion of Finland, and the Faroes (Vikør 2003:46).

For many of the languages in this data, the demographics of immigration may in part explain female-male discrepancies: OECD statistics for migration to the Nordics by gender and country of origin for the years 2000–2016 are largely congruent with the female-male language balance. For example, the number of female migrants to the Nordics from the Philippines (658,410) outstrips the number of male migrants to the Nordics (176,168) in the years 2000–2016 by a factor of 3.73 to 1 (OECD 2018). Similar female overrepresentation is true for migration from many of the countries whose main languages are also more "female" in the Twitter data set, such as Japan (1.61 to 1 female/male migration ratio), the Russophone countries of Russia, Belarus and Ukraine (1.52 to 1), Indonesia (1.42 to 1), or Latvia (1.07 to 1).

Likewise for the languages more used by males: The female-male migrant ratio to the Nordics from countries of the Arab world[5] is 0.71 to 1. For countries that use Dutch (Netherlands and Belgium) the ratio is 0.74 to 1, for Poland 0.87 to 1, for Iran and Afghanistan 0.74 to 1, and for Greece 0.6 to 1. The ratio for the German-speaking countries of Germany, Austria, and Switzerland is also slightly male (0.96 to 1). French is "male" in the Twitter data, and migration to the Nordics from Francophone countries[6] is also more male, with a ratio of 0.72 to 1. On the other hand, Spanish is more used by females in the Twitter data, while migration from Hispanophone[7] countries to the Nordics has been slightly more male from 2000–2016, with a ratio of 0.93 to 1.

Despite some outliers in the data, the Pearson product-moment correlation between the female-male odds ratio for non-Nordic languages in the Twitter data and the OECD female-male migration ratio from the principal countries in which the language has official status is moderately strong at 0.68 (p = 0.0013). In Figure 7, on the y-axis, values greater than one indicate more female than male migrants to the Nordics from the country/countries in which the language has official status. On the x-axis, values greater than one indicate that the language is more likely to be used by females in the Twitter data in this study. The correlation between gendered language use and migration suggests not only that migrants to the Nordics are using their L1 on Twitter, but further strengthens previous findings demonstrating that social media signals can be interpreted as a proxy for global movement and migration patterns (Hawelka et al. 2014).

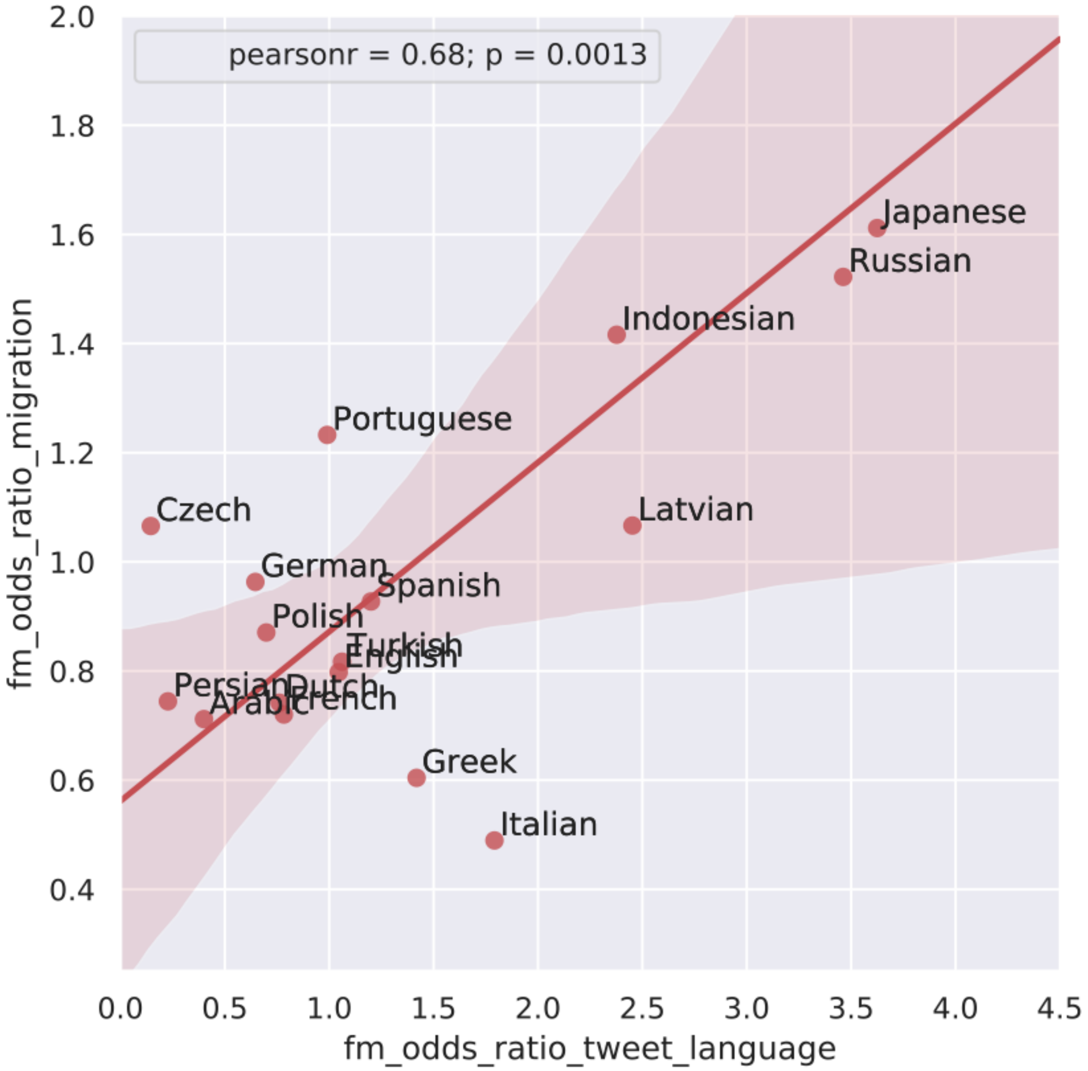

Figure 7: OECD female-male migration ratio vs. Twitter female-male language use ratio.

For the larger Nordic societies of Norway, Denmark, Sweden, and Finland, the principal national languages are more likely to be used by males and English by females. Male preference for local languages compared to females may be the result of gender differences in orientation towards extra-local norms, where females, particularly those who are better-educated, are quicker to adopt prestigious language features or language use patterns from outside the local communities (Gal 1979; Labov 1990, 2001; Cheshire 2002; Bilaniuk 2003; Lai 2007; Smith-Hefner 2009). Male use of local languages, as opposed to English, may represent allegiance to local identities during a period of increasing globalization and practical language shift, in some domains and types of interaction, away from the Nordic languages and towards English.

In line with findings from some previous studies and surveys (Gal 1979, Woolard 1997, Leppänen et al. 2011, European Commission 2018), males are found to exhibit higher overall rates of bi- and multilingualism in the Twitter data. In the sociolinguistics literature it has been noted that higher male bi- or multilingualism can reflect a gender-based disparity in rates of participation in work environments in which bilingualism is the norm (Labov 2011:274). More detailed information about the identities of the users who comprise the data sample would be necessary in order to test this hypothesis. In this data, higher male bi- and multilingualism may be a result of some females being more likely to tweet in English rather than in a local language.

6. SUMMARY AND CONCLUSION

In this study a large corpus of Twitter communication authored by users from the Nordic countries and territories was analyzed in terms of country/territory, gender, and language choice. Users were located within the Nordics and disambiguated for gender using list-based approaches.

In the first set of results, it was demonstrated that while the main Nordic languages are widely used by Nordic-based persons on Twitter, English use is also quite extensive. In the second set of results, gender differences in language use were considered by calculating the female-male odds ratio for 26 most widely-used languages on Twitter for the entire Nordic region. The data suggest that a language's use by males or females in the Twitter data corresponds approximately to migration rates to the Nordics by persons from places where that language has official status. A third set of results showed that presumable L1 users of Nordic languages, identified on the basis of their language choice for the Twitter interface and their user name, show a consistent gender-based pattern in use of the principal national language and of English: Males use the former more, whereas females use more English. In addition, males exhibit slightly higher rates of bi- and multilingualism.

Gender differences in language use have been seen as evidence for language shift in progress in some sociolinguistic studies, with bilingual females typically leading the change by making more use of the language or language variety perceived to have higher prestige or offer more opportunities for social advancement. In Nordic contexts (as well as in other geographical contexts) using English on Twitter may be seen as a means for connecting with a prestigious globalized culture, rather than a local culture with more limited prestige. The data suggest that females in the Nordic societies are leading a shift towards English on the Twitter platform.

While the procedures used in this study give insight into differential language use on Twitter in the Nordics according to location and gender, some methodological improvements are possible. As has been noted above, the sample sizes for Greenland, the Faroes, and Åland are too small to give reliable results. Inducing author gender based on name frequencies from the Nordic statistical offices restricts the data to names written using the Latin alphabet and derived glyphs: Twitter users from the Nordics with "author name" metadata written in Cyrillic, Asian, or Arabic (among other scripts), were not matched. This group, however, represents a small proportion of the overall number of Nordic Twitter users.

More information about the demographic parameters of Nordic-based Twitter users would allow a more nuanced interpretation of the findings pertaining to gendered differences in use of the Nordic languages and English. Of particular interest would be the manner in which demographic parameters such as age, educational level, occupation or language attitudes pattern with language choice and extent of bilingualism. While some information about these parameters could be gleaned from Twitter metadata (for example by scraping the "user" metadata for age- or work-related terms), a crowd-sourced survey method is also conceivable. For example, Nordic-based Twitter users willing to provide demographic details (which would be anonymized) could be solicited on the Twitter platform itself; confidence intervals for the relationship between a particular demographic trait and language use would depend on the number of survey responses. Finally, more detailed information about the migrant populations in Nordic countries could shed light on the observed patterns of gendered language choice that do not seem to correspond to migration patterns, for example for Spanish.

Widespread global use of social media platforms has made studies of language use possible on a scale that previously was not logistically feasible. For the Nordics – prosperous and innovative societies on the cusp of many types of global change – understanding the changing ways in which languages are used online necessitates a data-intensive approach, an approach that allows consideration of the complex interaction between parameters of personal identity such as gender and the broad currents of language evolution.

7. ACKNOWLEDGEMENTS

The author would like to thank three anonymous reviewers and the editors of the *NJL* for useful suggestions and Finland's Center for Scientific Computing (csc.fi) for providing computational resources and data storage facilities.

NOTES

1. For example, the seed data for this study included several thousand tweets with the place value “Bouvet Island”, a remote, uninhabited Antarctic island under Norwegian sovereignty. None of the texts in these tweets made reference to the island.

2. Except for Iceland, for which frequency data were not freely available.

3. The regex used was “\b^[^a-zA-ZÀ-ÿ]*(?:%s'%')[^a-zA-ZÀ-ÿ]+”, where %s'%' represents a name with ≥ 0.8 probability of being assigned to one gender.

4. Due to inconsistency in the way Nordic statistical agencies collect and report name frequency data, the Norwegian name filter for this step used 935 female and 831 male names, which were all of the name types with a probability of $\geq$ 0.8 of being assigned to single gender.

5. Morocco, Mauritania, Mali, Algeria, Tunisia, Libya, Egypt, the Sudan, Saudi Arabia, Yemen, Oman, the UAE, Bahrain, Qatar, Kuwait, Iraq, Lebanon, and Syria.

6. Algeria, Belgium, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Comoros, Democratic Republic of Congo, Djibouti, Equatorial Guinea, France, Gabon, Guinea, Haiti, Ivory Coast, Luxembourg, Madagascar, Mali, Monaco, Morocco, Niger, Republic of the Congo, Rwanda, Senegal, Seychelles, Switzerland, Togo, and Tunisia.

7. Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Spain, Uruguay, and Venezuela.

REFERENCES

Ajao, Oluwaseun, Jun Hong, & Weiru Liu. 2015. A survey of location inference techniques on Twitter. *Journal of Information Science* 41(6), 855–864. https://doi.org/10.1177/0165551515602847

Arnbjörnsdóttir, Birna. 2011. Exposure to English in Iceland: A quantitative and qualitative study. In Ingvar Sigurgeirsson, Ingólfur Ásgeir Jóhannesson, & Gretar L. Marinósson (eds.), *Ráðstefnurit Netlu: Menntakvika 2011*. Reykjavík: Menntavísindasvið Háskóla Íslands.

Audience Project. 2016. Audience Project device study 2016: Social media across the Nordics. https://www.audienceproject.com/wp-content/uploads/study_social_media_across_the_nordics.pdf (accessed 10 October 2018).

Avoindata.fi. 2017. Etunimitilasto 2017-09-04 VRK ("Given name statistics 2017-09-04 date") [Data set]. https://www.avoindata.fi/data/dataset (accessed 10 October 2018).

Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2), 135–160.

Bilaniuk, Laada. 2003. Gender, language attitudes, and language status in Ukraine. *Language in Society* 32, 47–78.

Bird, Steven, Edward Loper & Ewan Klein. 2009. *Natural Language Processing with Python*. Newton, MA: O'Reilly Media.

Björklund, Mikaela, Siv Björklund & Kaj Sjöholm. 2013. Multilingual policies and multilingual education in the Nordic countries. *International Electronic Journal of Elementary Education* 6(1), 1–22.

Bolton, Kingsley & Christiane Meierkord. 2013. English in contemporary Sweden: Perceptions, policies, and narrated practices. *Journal of Sociolinguistics* 17, 93–117.

Burger, John, John Henderson, George Kim & Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1301–1309. New York, NY: Association for Computing Machinery. http://aclweb.org/anthology//D/D11/D11-1120.pdf

Cheshire, Jenny. 2002. Sex and gender in variationist research. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change*, 423–443. Oxford: Blackwell.

Christensen, Dennis. 2017. *Media Development 2017: DR Audience Research Department's Annual Report on the Development of Use of Electronic Media in Denmark*. Copenhagen: Danmarks Radio. https://www.dr.dk/om-dr/about-dr/media-development-2010-2017 (accessed 10 October 2018).

Coats, Steven. 2016. Grammatical feature frequencies of English on Twitter in Finland. In Lauren Squires (ed.), *English in Computer-mediated Communication: Variation, Representation, and Change*, 179–210. Berlin: De Gruyter.

Coats, Steven. 2017a. Gender and lexical type frequencies in Finland Twitter English. In Säily, Tanja, Turo Hiltunen & Joseph McVeigh (eds.), *Big and Rich data in English Corpus Linguistics: Methods and Explorations* (= Studies in Variation, Contacts and Change in English 19). Helsinki, Finland: Varieng.

Coats, Steven. 2017b. Gender and grammatical frequencies in social media English from the Nordic countries. In Darja Fišer and Michael Beißwenger (eds.), Investigating social media corpora, 102–121. Ljubljana, Slovenia: U. of Ljubljana Academic Publishing.

Danmarks Statistic. 2015a. *Fornavne 2015 – Kvinder* ("Given names 2015 – Women") [Data set].

Danmarks Statistic. 2015b. *Fornavne 2015 – Mænd* ("Given names 2015 – Men") [Data set].

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(1).

Eleta, Irene & Jennifer Golbeck. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior* 41, 424–432.

European Commission. 2018. *Education and Training Database*. https://ec.europa.eu/eurostat/web/education-and-training/data/database (accessed 10 October 2018).

Gal, Susan. 1979. *Language Shift: Social Determinants of Linguistic Change in Bilingual Austria*. New York: Academic Press.

Graedler, Anne-Line. 2014. Attitudes towards English in Norway: A corpus-based study of attitudinal expressions in newspaper discourse. *Multilingua* 33(3-4), 291–312.

Graham, Mark, Scott A. Hale & Devin Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer* 66(4), 568–578. http://dx.doi.org/10.1080/00330124.2014.907699

Gries, Stefan. 2010. Dispersions and adjusted frequencies in corpora: Further explorations. In Stefan Gries, Stefanie Wulff, & Mark Davies (eds.), *Corpus Linguistic Applications: Current Studies, New Directions*, 197–212. Amsterdam: Rodopi.

Grosjean, François. 2008. Studying bilinguals: Methodological and conceptual issues. In Tej K. Bhatia & William C. Ritchie (eds.), *Handbook of Bilingualism*, 32–63. Malden, MA: Wiley-Blackwell.

Görlach, Manfred. 2002. *Still More Englishes*. Amsterdam: John Benjamins.

Hagiwara, Masato. 2014. *Tinysegmenter: Tokenizer Specified for Japanese*. https://github.com/SamuraiT/tinysegmenter (accessed 10 October 2018).

Hale, Scott. 2014. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 833–842. New York, NY: Association for Computing Machinery.

Haustein, Stefanie, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto & Vincent Larivière. 2015. Tweets as impact indicators: Examining the implications of automated 'bot' accounts on Twitter. *Journal of the Association for Information Science and Technology* 67(1), 232–238. https://dx.doi.org/10.1002/asi.23456

Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos & Carlo Ratti. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Hochmair, Hartwig H., Levente Juhász & Sreten Cvetojevic. 2018. Data quality of points of interest in selected mapping and social media platforms. In Peter Kiefer, Haosheng Huang, Nico Van de Weghe, & Martin Raubal (eds.), *Progress in Location Based Services 2018*, 293–313. Cham: Springer.

Hong, Lichan. Gregorio Convertino & Ed H. Chi. 2010. Language matters in Twitter: A large scale study. In *International AAAI Conference on Weblogs and Social Media*, 518–521. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Jeeves, Anna. 2011. Learning English in contemporary Iceland – the attitudes and perceptions of Icelandic youth. In Andrew Linn, Neil Bermel & Gibson Ferguson (eds.), *Attitudes towards English in Europe: English in Europe, Volume 1*, 271–296. Berlin/Boston: De Gruyter Mouton.

Jørgensen, Anna Katrine, Dirk Hovy & Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, 9–18. Stroudsburg, PA: Association for Computational Linguistics. http://aclweb.org/anthology/W15-4302 (accessed 10 October 2018).

Kokkos, Athanasios & Theodoros Tzouramanis. 2014. A robust gender inference model for online social net-works and its application to LinkedIn and Twitter. *First Monday* 19(9).

Kristiansen, Tore & Helge Sandøy. 2010a. The linguistic consequences of globalization: the Nordic laboratory. [Special issue]. *International Journal of the Sociology of Language* 204.

Kristiansen, Tore & Helge Sandøy. 2010b. Conclusion. Globalization and language in the Nordic countries: Conditions and consequences. *International Journal of the Sociology of Language* 204, 151–159.

Kytölä, Samu & Elina Westinen. 2015. 'I be da reel gansta' — A Finnish footballer's Twitter writing and metapragmatic evaluations of authenticity. *Discourse, Context & Media* 8, 6–19.

Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2, 205–254.

Labov, William. 2001. *Principles of Linguistic Change, vol. 2: Social Factors*. Oxford: Blackwell.

Lai, Mee-Ling. 2009. Gender and language attitudes: A case of postcolonial Hong Kong. *International Journal of Multilingualism* 4(2), 83–116.

Lakoff , Robin. 1973. Language and woman's place. *Language in Society* 2(1), 45–80.

Laylavi, Farhad, Abbas Rajabifard & Moshen Kalantari. 2016. A multi-element approach to location inference of Twitter: A case for emergency response. *International Journal of Geo-Information* 5(56).

Lee, Carmen. 2016. Multilingual resources and practices in digital communication. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), *The Routledge Handbook of Language and Digital Communication*, 118–132. London and New York: Routledge.

Leetaru, Kalev H., Shaowen Wang, Guofeng Cao, Anand Padmanabhan & Eric Shook. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18(5/6).

Leppänen, Sirpa, Anne Pitkänen-Huhta, Arja Piirainen-Marsh, Tarja Nikula & Saija Peuronen. 2009. Young people's translocal new media uses: A multiperspective analysis of language choice and heteroglossia. *Journal of Computer-Mediated Communication* 14(4): 1080–1107.

Leppänen, Sirpa, Anne Pitkänen-Huhta, Tarja Nikula, Samu Kytölä, Timo Törmäkangas, Kari Nissinen, Leila Kääntä, Tiina Räisänen, Mikko Laitinen, Heidi Koskela, Salla Lähdesmäki & Henna Jousmäki. 2011. *National Survey on the English Language in Finland: Uses, Meanings and Attitudes* (= Studies in Variation, Contacts and Change in English, Volume 5). Helsinki: Varieng.

Linn, Andrew. 2016. The Nordic experience. In Andrew Linn (ed.), *Investigating English in Europe: Contexts and Agendas*, 201–258. Berlin/Boston: De Gruyter Mouton.

Lønsmann, Dorte. 2009. From subculture to mainstream: The spread of English in Denmark. *Journal of Pragmatics* 41(6): 1139–1151.

Lui, Marco & Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) EACL 2014*, 17–25. Stroudsburg, PA: Association for Computational Linguistics.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela & J. Niels Rosenquist. 2011. Understanding the demographics of Twitter users. In *Proceedings of ICWSM*, 554–557. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang & Alessandro Vespignani. 2013. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE* 8(4).

Mortensen, Bjarma. 2011. Policies and attitudes towards English in the Faroes today. In Linn, Andrew, Neil Bermel & Gibson Ferguson (eds.), *Attitudes towards English in Europe: English in Europe, volume 1*, 71–96. Berlin/Boston: De Gruyter Mouton.

Norsk Rikskringkasting. 2015. Oppsummeringen 2015: NRK Analyse ("Summary 2015: NRK Analysis"). Oslo: NRK. https://fido.nrk.no/3059e4aff03749086d752a93b64cee618921d5c7bc51bd87b2e07bd8703fef69/medier_norge_2015_nrk_analyse.pdf (accessed 10 October 2018).

OECD. 2018. *International Migration Database*. http://dx.doi.org/10.1787/data-00342-en (accessed 10 October 2018).

Rao, Delip, David Yarowsky, Abhishek Shreevats & Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, 37–44. New York, NY: Association for Computing Machinery.

Rindal, Ulrikke. 2010. Constructing identity with L2: Pronunciation and attitudes among Norwegian learners of English. *Journal of Sociolinguistics* 14, 240–261.

Rindal, Ulrikke & Caroline Piercy. 2013. Being 'neutral'? English pronunciation among Norwegian learners. *World Englishes* 32(2), 211–229.

Roesslein, Joshua. 2015. *Tweepy*. Python programming language module. https://github.com/tweepy/tweepy (accessed 10 October 2018).

Romaine, Suzanne. 2008. The bilingual and multilingual community. In Tej K. Bhatia & William C. Ritchie (eds.), *Handbook of Bilingualism*, 385–405. Malden, MA: Wiley-Blackwell.

Ronen, Shahar, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker & César A. Hidalgo. 2014. Links that speak: The global language network and its association with global fame. *PNAS* 111(52), E5616–E5622.

Sandøy, Helge (ed.), *Med 'bil' i Norden i 100 år. Ordlaging og tilpassing av utalandske ord* ("100 years with 'car' in the North: Construction and adaptation of foreign words"). Moderne importord i språka i Norden 1 ("Modern import words in the Nordics 1"). Oslo: Novus.

Sandøy, Helge. 2003. Moderne importord i Norden. Ei gransking av bruk, normer og språkholdningar ("Modern import words in the North. A review of usage, norms, and language attitudes"). In Sandøy, Helge (ed.), 73–100.

Schulz, Axel, Aristoteles Hadjakos, Heiko Paulheim, Johannes Nachtwey & Max Mühlhäuser. 2013. A multi-indicator approach for geolocalization of tweets. In *Proceedings of ICWSM*, 573–582. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Sites, Dick. 2013. *Compact Language Detector 2*. https://github.com/CLD2Owners/cld2 (accessed 10 October 2018).

Smith-Hefner, Nancy. 2009. Language shift, gender, and ideologies of modernity in central Java, Indonesia. *Journal of Linguistic Anthropology*, 19(1), 57–77.

Sperstad, Tormod. 2018. Oppdatert sosiale medier-statistikk fra Norge ("Updated social media statistics from Norway"). https://www.tormodsperstad.no/oppdatert-sosiale-medier-statistikk-norge/ (accessed 10 October 2018).

Squires, Lauren. 2015. Twitter: Design, discourse, and implications of public text. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), *The Routledge Handbook of Language and Digital Communication*, 239–256. London and New York: Routledge.

Stæhr, Andreas & Lian M. Madsen. 2014. Standard language in urban rap: Social media, linguistic practice and ethnographic context . *Tilburg Papers in Culture Studies*, Paper 94. Tilburg: Tilburg University.

Statistics Greenland. 2017. *De hyppigst anvendte (fem eller flere bærere) fornavne i Grønland. 1. juli 2011* ("The most frequent given names (five or more bearers) in Greenland. July 1 2011") [Data set]. http://www.stat.gl/dialog/main.asp?lang=da&version=201102&sc=NA&colcode=b. (accessed 10 October 2018).

Statistics Iceland. 2017. *Population and elecions* [Data sets]. http://px.hagstofa.is/pxen/pxweb/en/Ibuar/Ibuar__Faeddirdanir__Nofn__Nofnkk/ (accessed 10 October 2018).

Statistics Norway. 2017b. *Guttenavn alfabetisk 2008-2017* ("Boy names, alphabetized 2008–2017") [Data set]. https://www.ssb.no/befolkning/statistikker/navn/aar (accessed 10 October 2018).

Statistics Norway. 2017a. *Jentenavn, alfabetisk 2006-2017* ("Girl names, alphabetized 2006–2017") [Data set]. https://www.ssb.no/befolkning/statistikker/navn/aar (accessed 10 October 2018).

Sun, Junyi. 2014. *Jieba: Chinese word segmentation* module. https://github.com/fxsjy/jieba (accessed 10 October 2018).

Thøgersen, Jacob. 2004. Attitudes towards the English influx in the Nordic countries: A quantitative investigation. *Nordic Journal of English Studies* 3(2), 23–38.

Trudgill, Peter. 1974. *The Social Differentiation of English in Norwich*. London: Cambridge University Press.

Trudgill, Peter. 1998. Sex and covert prestige. In Jennifer Coates, (ed.), *Language and Gender: A Reader*, 21–28. Oxford, UK and Malden, MA: Blackwell.

Twitter. 2013. Introducing new metadata for Tweets. https://blog.twitter.com/2013/introducing-new-metadata-for-tweets (accessed 10 October 2018).

Twitter. 2015. Evaluating language identification performance. https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.html (accessed 10 October 2018).

Vikør, Lars. 2003. Nordiske språkhaldningar: Presentasjon av ei meiningsmåling ("Nordic language attitudes. Presentation of survey results"). In Sandøy, Helge (ed.), 42–51.

Volkova, Svitlana, Yoram Bachrach, Michael Armstrong & Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. *Proceedings of the Twenty-Ninth AAAI Conference* *on Artificial Intelligence*, 4296–4297. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Wikström, Peter. 2014. #srynotfunny: Communicative functions of hashtags on Twitter. *SKY Journal of Linguistics* 27, 127–152.

Woolard, Kathryn A. 1997. Between friends: Gender, peer group structure, and bilingualism in urban Catalonia. *Language in Society* 26(4), 533–560.

Zubiaga, Arkaitz, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, & Víctor Fresno. 2016. TweetLID: a benchmark for tweet language identification. *Language Resources and Evaluation* 50(4), 729–766.