Gender and grammatical Frequencies in social media English from the Nordic countries

Steven Coats, *University of Oulu*

Abstract

English has become firmly established as a primary vehicle for global communication, and is thus also increasingly used in online contexts for local communicative purposes, for example in the Nordic societies. This paper investigates the extent to which English is used on Twitter in the Nordic countries and builds on previous research by investigating the link between gender and grammatical or part-of-speech frequencies, a link which has hitherto been considered mainly in the context of data collected in L1 Anglophone contexts.
The Twitter Streaming API was used to create a corpus of English-language messages originating from the Nordic countries. Automatic methods were used to disambiguate author gender and apply part-of-speech tags, and the relative frequencies of grammatical types by gender were determined for each country. Principal components analysis shows that Nordic English-language discourse on Twitter diverges according to gender for a number of grammatical features. The analysis supports L1 findings pertaining to gendered differences in feature frequencies in English.

Keywords: Twitter, CMC, sociolinguistics, gender, corpus linguistics

1 INTRODUCTION

Recent shifts in communication behavior towards online social media platforms provide opportunities for the study of variation in English as it is used worldwide. While the status of English, as the world's principal lingua franca, continues to consolidate in many global contexts of use, it is hardly a monolithic entity: English as it is used in global computer-mediated communication (CMC) exhibits a great variety of features in orthography, lexis, grammar, and style, especially in non-L1 environments. Such diversity has been characterized by Blommaert (2012) as a "supervernacular".

CMC and social media such as Twitter have become important sites of interaction for many, and in recent years a number of studies have investigated various properties of Twitter language (for an overview of the communicative and discourse functions of Twitter language, see Page 2012, Zappavigna 2011, and Squires 2015). The ubiquity and volume of Twitter data, its public availability through a well-maintained set of APIs (*Application Programming Interfaces*), and the extensiveness of the associated tweet metadata fields allow for a rich variety of analyses. As a significant proportion of tweets are associated with metadata detailing the physical location of their authors, geographical analyses of language use and linguistic diversity have been a natural focus of research interest (e.g. Leetaru et al. 2013, Mocanu et al. 2014). Twitter data has also been used to investigate dialectological (Eisenstein et al. 2014) and sociolinguistic aspects of American English, including the relationship between gender and language variation (Bamann, Eisenstein and Schnoebelen 2014).

Differences between the genders in the relative frequency of lexical types or word classes have been investigated in a number of studies. A large, corpus-based study of lexical type frequencies based on writing samples submitted to a website found significant differences between males and females in the relative frequencies of pronouns, numbers, negators, articles, and prepositions, among other world classes (Newman et al. 2008). Corpus-based research using language data extracted from instant messaging or blog posts has also found that some differences in feature frequency can be associated with gender. For example, it has been found in online writing that females may use more personal pronouns, modal verbs, and emoticons, while males use more determiners such as articles or demonstrative pronouns and more numbers or numerals (Baron 2004, Herring and Paolillo 2006, Argamon et al. 2007). Similar findings have resulted from a large-scale investigation of word frequencies and gender on Twitter, although gender-based associations with particular features are typically less strong than associations based on local networks (Bamann, Eisenstein and Schnoebelen 2014). For the most part, however, analysis of type frequencies

in English has been conducted on data from Anglophone contexts, mainly in the United States, and relatively little corpus-based research has looked into relative frequencies in non-L1 contexts.[1] Frequency-based analyses of variation in global Englishes as they are manifest in aggregate online media such as Twitter have not yet been undertaken on a large scale, although some studies exist.[2] Given the global nature of social media and the ever-increasing importance of English, variation in English in global contexts represents an important site of language variation and change.

Knowledge of English is extensive in the Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland, nations with well-developed economies and high levels of educational attainment. With populations that are to a large degree bilingual in a national language and English, the Nordic countries are perhaps the societies in which English is most extensively used without being an official language: English is so prevalent in the Nordics that it has been suggested that the national languages are becoming linguistic systems with "restricted functional range" (Görlach 2002: 16). Although much research has addressed various aspects of English use in the Nordic countries (for Sweden, e.g., see Bolton and Meierkord 2013; for Finland see the extensive survey study of Leppänen et al. 2011), and some preliminary work on language use on Twitter by country has also provided data for the Nordics (Mocanu et al. 2013), linguistic diversity on social media in Northern Europe has not been investigated in detail. Likewise, although some work exists on grammatical feature frequencies in Nordic non-CMC genres (e.g. for Swedish in Allwood 1998), there are few studies of feature frequencies in English in non-L1 environments, and the relationship between author gender and feature frequency in CMC or social media language varieties such as Twitter has not yet been explored in Nordic contexts, whether in local languages or English.[3]

This study adopts an approach based in part on multidimensional analysis (Biber 1988, 1995). After establishing the extent to which English is used on Twitter in the Nordic national contexts, relative grammatical feature frequencies are calculated and the features most strongly associated with gender identified. Using principal components analysis, the underlying associations among feature frequencies, gender, and communicative function are established.

1 See, however, Xiao 2009 for a corpus-based investigation of world English varieties as represented in the International Corpus of English.

2 E.g. Coats (2016).

3 For an analysis of feature frequencies in English as it is used in various Asian contexts see Xiao (2009). Baron (2004) analyses a small corpus of Instant Messenger data in English from American and Swedish university students.

2 METHODS

The methods used in the study include the collection of data from Twitter's Streaming API, the filtering of this data to remove tweets sent by bots or other non-human agents, the disambiguation of tweet author gender and assignation of tweets to gendered subcorpora, the assignation of exact location and language to each tweet, the tokenization of tweets, part-of-speech tagging of the English-language tweets, and the statistical analysis of the resulting subcorpora. Data collection, filtering, and statistical analysis were done in Python and in R.

2.1 Data collection

Data was collected in .json format from Twitter's Streaming API from 9 November 2016 until 18 February 2017 by utilizing the *Tweepy* library in Python (Roesslein 2015).[4] The data collection script saved only tweets with a populated *place* field.

2.2 Filtering for automatic tweets

A substantial proportion of messages on Twitter are automatically generated texts created by bots or scripts, some of which automatically generate English text. The *Foursquare* app, for example, can automatically tweet short English-language sentences about a user's GPS-determined location. In an effort to reduce the potential error that such messages could introduce into the analysis (such users may not necessarily author any English-language tweets), an initial filtering step selected from the metadata *source* field those sources that are likely to be used by human agents.[5]

2.3 Geolocation

When composing a tweet, users often select a *place* from a list automatically generated by Twitter. These place suggestions are based on a user's IP address, with the coordinates automatically assigned by Twitter as a bounding box of latitude-longitude coordinates in the tweet's metadata. Some users (those using smartphones or

4 https://github.com/tweepy/tweepy.

5 The sources selected were *Twitter Web Client*, *Twitter for iPhone*, *Twitter for Android*, *Twitter for iPad*, *Twitter for Windows Phone*, *Twitter for Android Tablet*, *Tweetbot for Mac*, and *Instagram*. Although there were over 1,500 sources in the initial data, these eight accounted for 91% of all the tweets collected from the Streaming API.

other GPS-enabled devices) additionally opt to broadcast exact latitude-longitude coordinates with each status update; these appear in the *geo* metadata field.

Each tweet in the data was assigned exact latitude-longitude coordinates: either the exact coordinates from the *geo* field, or (if no GPS coordinates were available), a set of latitude-longitude values calculated as the center of the bounding box circumscribing the *place* field. Although users can manually enter a *place* that does not correspond to their physical location, this does not seem to occur on a large scale. For tweets that contained both *place* and *geo* objects, the product-moment correlation of the coordinate values in the Nordic data was 0.989 (for longitude) and 0.960 (for latitude).[6]

Filtering for the *country_code* field selected only tweets with geo-coordinates within the territorial boundaries of the Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland. Of the 310.7 million tweets collected globally in the initial dataset, 1.76m were from the Nordic countries.

Subcorpora were prepared for each country by filtering the data according to the *language* field: tweets in the principal national language(s), and tweets in English.[7] Tweets originating from outside the Nordic countries and in other languages were not further considered. The English-language data comprised in total 460,260 tweets and 6,360,835 tokens.

2.4 Gender disambiguation

Unlike some social media platforms, Twitter does not provide users with a profile field where gender is reported; nor are users required to otherwise supply gender information. In the absence of self-reported gender information, an automatic procedure for gender disambiguation based on values in the *author_name* field was employed. Disambiguation of tweet author gender based on gender-name associations has been employed for data from the United States (Rao et al. 2010; Mislove et al. 2011),[8] but, to the best of our knowledge, not for the Nordic countries.

6 Some *place* values in the data were obviously not accurate, such as over 1,000 tweets with a *place* value for Bouvet Island, a small, uninhabited sub-Antarctic island. Twitter uses an internal database of *places* that includes places with ISO-3166 codes; these place names (and others) are then automatically suggested to users based on their IP address and keyboard input when they are selecting a *place* for a tweet. The *location* field in the Twitter user profile utilizes the same Twitter-internal database of locations from which users can select the appropriate one.

7 Based on the value in the *language* field. For Norway, both *Nynorsk* and *Riksmål* were categorized as "Norwegian". For Finland, corpora were also created for the country's second official language, Swedish.

8 Latent attribute inference using Twitter data manually tagged for gender is a popular topic in machine learning (cf. Pennacchiotti and Popescu 2011; Ciot, Sonderegger and Ruths 2013). The approach used here relies on the association between given name and author gender, rather than using machine learning to infer gender based on the content of messages whose authors' gender has been manually tagged.

In order to assign tweets to male or female gender categories, lists of the most frequent given names in the Nordic countries were obtained from the national statistical offices. The *author_name* field for each tweet was then filtered via regular expressions for strings that either begin with or include as a discrete element the most common male and female given names in the corresponding Nordic country.[9] While extensive name information was available for Denmark, Sweden, and Finland, it was less available for Iceland and Norway. In total, 13,506 unique male and 15,497 unique female given names from the lists were matched with the value of the *author_name* attribute for each unique user in the dataset. Users matching both male and female names were discarded. The method assigned gender to 61.5% of Nordic tweets (25% of Iceland, 57% of Norway, 60% of Denmark, 63% of Sweden, and 70% of Finland tweets).[10]

2.5 Additional text filtering

Before tokenization and part-of-speech tagging was undertaken, HTML escape characters in the *text* field were replaced with the corresponding characters. The following subcorpora were created for further analysis: First, from the gender-disambiguated data, for each country a subcorpus of tweets in all languages, in order to gauge the relative representation of different languages in the Nordics. Second, for each Nordic country a male subcorpus and a female subcorpus consisting of English-language messages geo-located to those countries whose *author_name* values matched the corresponding list of frequent male and female given names.

2.6 Tokenization and part-of-speech tagging

The Carnegie-Mellon University Twitter Tagger (Gimpel et al. 2011, Owoputi et al. 2013) was used to tokenize the gendered English-language subcorpora and apply part-of-speech tags using a subset of the Penn Treebank tagset (Marcus, Marcinkiewicz and Santorini 1993), with additional tags for the Twitter-specific features *username*, *hashtag*, and *retweet*. The tool was trained on Twitter data and is somewhat tolerant of the non-standard orthography typical of Twitter messages.

9 http://www.statice.is, http://www.ssb.no/befolkning, http://www.scb.se/sv_/Hitta-statistik, and the open data portal for Finland https://www.avoindata.fi.

10 The differences are due in part to the somewhat different name frequency information obtained from the national statistical offices. For example, only 402 given names were obtained from Iceland, but 1741 from Norway, 5,382 from Denmark, 25,226 from Sweden, and 7,899 from Finland. For a dataset of American tweets disambiguated for gender using name data from the U.S. Census Bureau, Mislove et al. report 64.5% gender disambiguation and a similar overrepresentation of males (2011: 556). The reason for the male overrepresentation in the data is unknown: Males may be more active on Twitter, or for whatever reason, may be more likely to use their legal name in the *author_name* field.

3 ANALYSIS AND DISCUSSION

The linguistic profiles of the national subcorpora were determined, and the relationship between gender and grammatical features in English-language messages assessed using Student's t-tests of population means. Principal components analysis was used to investigate underlying variability and so gauge the extent to which males and females from the Nordic countries may utilize different communicative styles in English on Twitter.

3.1 Language profile

English is extensively used in Twitter user messages originating from the Nordic countries. Table 1 shows the proportions of tweets in the national language(s), English, and other languages for tweets that were assigned gender based on the *author_name* values.[11]

Table 1: Percent tweets by country and language.

	Nat. Lang.	English	Other
Iceland	74.4	13.7	11.9
Norway	43.5	27.1	29.3
Denmark	38.3	41.5	20.2
Sweden	57.5	23.3	19.2
Finland	63.2	22.6	14.2

Use of English on Twitter is most extensive in Denmark, followed by Norway, Sweden, Finland, and Iceland. For the combined male and female data, the proportion of tweets in English by province is shown in Figure 1.[12] Although clear patterns of English use within the individual Nordic countries are not evident, there is a trend towards higher rates of English use in capital regions and more urbanized areas: For example, the territories of the national capitals

11 For Finland, the percentage shown includes messages in the national languages of Finnish and Swedish (Finnish = 62.0% of tweets, Swedish = 1.2%). "Other" includes tweets classified as in other languages, as well as (typically short) tweets whose language could not be automatically detected.

12 As of early 2017, the Twitter-internal library of places which are prompted to users when they compose tweets does not contain any province or city names for Iceland. Only the place "Iceland" can be given. As such, tweets from Iceland with a *place* value but without exact GPS coordinates are located in the center of the latitude-longitude bounding box around the country. For this data, this falls within the province of Norðurland vestra, which in Figure 1 has an English density of 12.4%. Because relatively few of the gendered tweets contain GPS coordinates (for Iceland 5.7%) and far more tweets have *place* coordinates, the overall percentage of English tweets in the gendered data from Iceland is 13.7%.

of Oslo, Copenhagen, Stockholm, and Helsinki show a higher proportion of tweets in English than do their respective countries overall. In a sociolinguistic context, such a pattern may demonstrate the fact that residents of capitals and larger cities typically have above-average levels of income and educational attainment, and that English may serve as a high-prestige language associated with internationality.

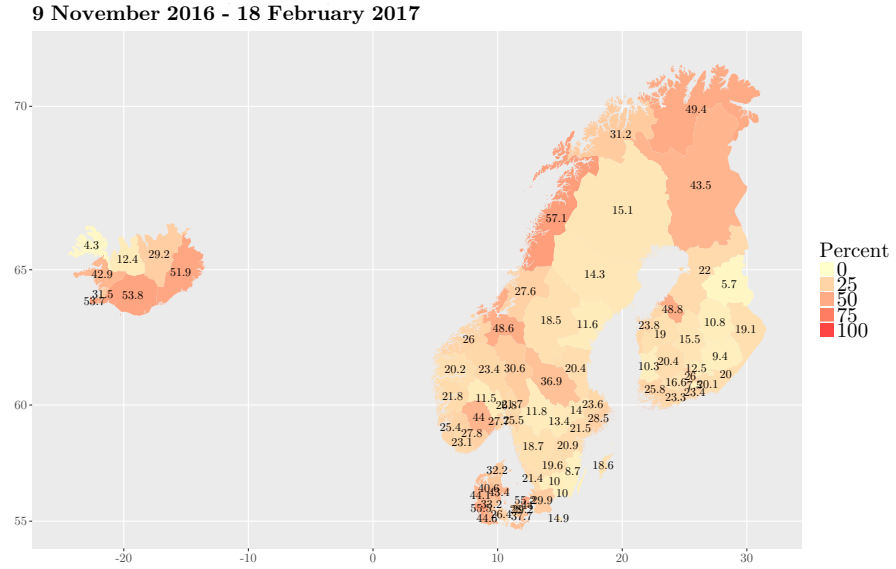

Figure 1: Percent of gendered tweets in English.

Males use the national language on Twitter more than females do in all five Nordic countries; females use English more in all countries except for Iceland (Table 2).

Table 2: Percentage of tweets by country, gender and language.

		Nat. Lang.	English	Other
Iceland	males	74.6	14.0	11.4
	females	74.0	13.3	12.7
Norway	males	46.0	24.1	29.9
	females	38.9	32.8	28.3
Denmark	males	45.8	37.6	16.6
	females	27.5	47.2	25.3
Sweden	males	58.8	22.9	18.3
	females	55.4	24.0	20.6
Finland	males	64.2	21.4	14.4
	females	61.4	24.5	14.1

The difference is most pronounced for Denmark and Norway, and less pronounced for Sweden, Finland, and Iceland. The differences in English use by gender were significant at $p < 0.05$ for all countries but Iceland (Fisher's Exact Test).[13]

3.2 Relationships among grammatical features, country and gender

Thirty-eight of the PoS tags were applied at least once in all of the ten gendered subcorpora. For each subcorpus, the relative frequency of each tag per 1,000 tokens was calculated (Table 3).

Table 3: Frequencies of grammatical features per 1,000 tokens.

	Iceland		Norway		Denmark		Sweden		Finland	
	m	f	m	f	m	f	m	f	m	f
Left bracket (()	1.03	1.22	1.59	1.03	1.85	1.18	1.64	1.41	2.07	1.16
Right bracket ())	1.09	1.03	1.47	0.88	1.74	1.16	1.59	1.34	2.25	1.07
Comma	16.87	12.41	21.81	14.66	19.72	15.91	24.25	16.68	20.25	16.97
Other punctuation (: ; ... + - = < > [ ])	19.47	30.56	20.77	17.18	26.02	20	17.89	19.14	27.6	20.87
Sentence-ending punctuation (. ? !)	57.56	49.09	55.96	49.41	54.12	44.31	66.26	54.88	56.75	52.06
Quotation marks (»)	8.77	5.92	7.85	6.56	7.29	6.74	8.83	8.54	9.26	7.22
Coordinating conjunction	17.9	17.59	18.19	19.27	19.57	20.34	20.82	21.26	19.41	21.8
Number	13.3	10.72	14.29	9.52	13.21	10.21	13.71	11.49	15.77	11.44
Determiner	65.97	62.44	61.75	67.12	60.24	53.43	63.68	60.34	54.53	53.84
Existential *there*	0.42	0.38	0.48	0.34	0.43	0.34	0.45	0.39	0.62	0.52
Foreign word	0.06	0.09	0.03	0	0.03	0.02	0.04	0.02	0.06	0.03
Hashtag	36.28	59.71	39.56	34.39	36.98	38.74	32.69	34.59	61.26	59.39
Preposition or subordinating conjunction	73.23	72.79	76.78	55.47	78.25	65.39	76.68	70.42	75.73	69.39
Adjective	50.85	42.13	48.07	65.93	50.99	50.4	53.19	52.86	52.75	52.72
Comparative adjective	1.75	1.5	1.73	1.18	1.83	1.4	1.83	1.52	1.82	1.77

13 Iceland: $p = 0.188$, odds ratio = 0.94; Norway: $p < 2.2e{-}16$, odds ratio = 1.54; Denmark: $p < 2.2e{-}16$, odds ratio = 1.48; Sweden: $p = 1.05e{-}16$, odds ratio = 1.06; Finland: $p < 2.2e{-}16$, odds ratio = 1.19.

	Iceland		Norway		Denmark		Sweden		Finland	
	m	f	m	f	m	f	m	f	m	f
Superlative adjective	3.57	3.01	2.4	1.72	2.27	2.4	2.5	2.59	2.46	2.77
Modal verb	11.19	8.65	9.77	7.27	10.93	10.32	11.88	9.98	8.92	9.41
Noun, singular or mass	118.51	109.55	109.15	119.79	114.41	99.38	109.84	108.27	112.37	105.37
Proper noun	74.5	64.23	80.85	85.14	76.04	55.15	64.91	64.46	74.76	56.95
Plural noun	29.08	25.3	28.04	35.65	29.33	23.45	33.04	27.66	31.34	27.73
Personal pronoun	59.26	60.28	50.62	53.61	55.41	80.04	63.16	72.71	44.03	68.76
Possessive pronoun	14.21	16.36	10.83	13.13	12.13	15.98	11.89	17.87	9.86	14.46
Adverb	42.27	35.55	39.61	37.5	43.39	48.44	48.44	47.17	39.45	49.53
Comparative adverb	2.12	1.5	1.4	1.06	1.61	1.18	1.58	1.4	1.39	1.41
Phrasal particle	4.41	4.61	4.3	4.1	4.17	4.04	4.23	4.09	3.26	3.28
Retweet	0.06	0.09	0.3	0.13	0.09	0.1	0.06	0.2	0.07	0.22
to	15.72	17.02	16.95	14.01	17.15	17.11	18.06	17.16	18.92	17.86
Interjection/ emoticon/ emoji	30.29	60.65	36.51	70.44	35.08	63.5	25.08	46.45	28.34	43.34
URL	34.95	47.49	29.03	29.91	31.45	29.91	28.34	31.61	37.1	33.97
Username (preceded by @)	55.15	41	79.08	54.51	58.55	75.72	49.35	45.81	59.27	53.44
Verb, base form	40.15	38.65	36.05	31.81	38.89	38.94	40.64	42.69	34.28	38.75
Verb, past tense	17.96	15.23	17.53	20.12	16.64	19.13	18.24	17.59	16.08	18.01
Verb, gerund or present particle	18.68	17.59	16.92	16.58	18.36	18.3	16.8	18.48	18.36	18.89
Verb, past participle	5.5	7.24	6.89	4.67	7.79	6.03	8.18	6.79	7.2	6.25
Verb, non-3rd person singular present	26.79	26.52	23.36	34.09	24.67	32.94	28.3	31.42	21.11	28.93
Verb, 3rd person singular present	20.5	19	19.72	13.51	19.73	17.61	20.62	19.18	21.15	19.29
Wh-determiner	0.67	0.38	0.58	0.41	0.62	0.53	0.84	0.7	0.69	0.71
Wh-pronoun	4.54	4.89	4.26	2.92	3.57	4.02	4.34	4.24	3.88	4.23
Wh-adverb	5.32	7.43	5.47	4.96	5.33	6.19	6.09	6.57	5.5	6.11

While the distributions of feature frequencies for frequent features such as pronouns or verbal forms approach normality, infrequent features such as Wh-determiners

are not normally distributed in the data. Thus, to determine whether differences in feature use by gender exist, Mann-Whitney U tests were conducted for each feature on the basis of the mean standardized values for males and for females in the gendered subcorpora. Of the 39 features, eleven exhibited significant ($p < 0.05$) differences in use between males and females: Right brackets, commas, sentence-ending punctuation, quotation marks, numbers/ numerals, prepositions or subordinating conjunctions, comparative adjectives, and 3rd-person singular present verb forms were significantly more likely to be utilized by males, while possessive pronouns, interjections/emoticons/emoji, and non-3rd-person singular present verb forms were significantly more likely to be used by females (Table 4).

Table 4: Grammatical features by gender.

	Feature	Gender	p-value		Feature	Gender	p-value
1	Left bracket (()	m	0.151	21	Personal pronoun	f	0.095
2	**Right bracket ())**	**m**	**0.032**	22	**Possessive pronoun**	**f**	**0.016**
3	**Comma**	**m**	**0.016**	23	Adverb	f	1.000
4	Other punctuation (: ; ... + - = < > [ ])	m	0.841	24	Comparative adverb	m	0.151
5	**Sentence-ending punctuation (. ? !)**	**m**	**0.016**	25	Phrasal particle	m	0.548
6	**Quotation marks (»)**	**m**	**0.032**	26	Retweet	f	0.151
7	Coordinating conjunction	f	0.548	27	*to*	m	0.690
8	**Number**	**m**	**0.008**	28	**Interjection/ emoticon/emoji**	**f**	**0.008**
9	Determiner	m	0.690	29	URL	f	0.690
10	Existential *there*	m	0.095	30	Username (preceded by @)	m	0.222
11	Foreign word	m	0.310	31	Verb, base form	f	1.000
12	Hashtag	f	1.000	32	Verb, past tense	f	0.421
13	**Preposition or subordinating conjunction**	**m**	**0.008**	33	Verb, gerund or present particle	f	1.000
14	Adjective	f	1.000	34	Verb, past participle	m	0.222
15	**Comparative adjective**	**m**	**0.032**	35	**Verb, non-3rd person singular present**	**f**	**0.032**
16	Superlative adjective	m	0.841	36	**Verb, 3rd person singular present**	**m**	**0.008**
17	Modal verb	m	0.151	37	Wh-determiner	m	0.421
18	Noun, singular or mass	m	0.222	38	Wh-pronoun	m	0.841
19	Proper noun	m	0.151	39	Wh-adverb	f	0.151
20	Plural noun	m	0.151				

Significant differences by gender at $p < 0.05$ for features in bold (Mann-Whitney U test)

A t-test of population means conducted on the same data gave similar results (Figure 2).

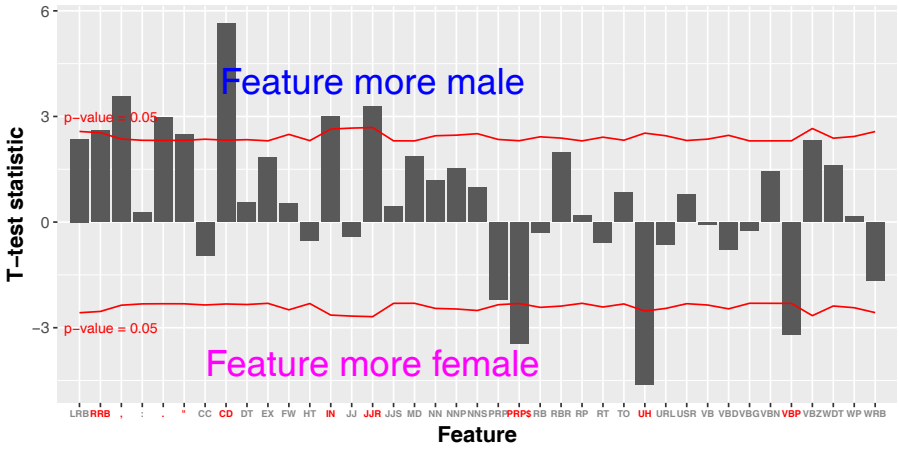

Figure 2: Grammatical features by gender (t-test).

Gendered differences were also considered by country and feature on the basis of the aggregate feature frequencies per unique user in the data. While differences in sample size make the results of t-tests for infrequent and non-normally-distributed features somewhat unreliable, particularly for Iceland due to the small number of users in the sample, many of the differences in feature frequencies between males and females were found for most or all of the Nordic countries.

3.3 Principal components analysis

In order to explore the underlying patterning of the variance in the data, a principal components analysis was conducted on a covariance matrix of the normalized frequencies of the 39 variables for the ten English subcorpora (the male and female subcorpora for each of the five Nordic countries). The first two components capture 58.21% of the variance in the data. The strongest loadings (> 0.2) on the first two components are shown in Table 5.

Table 5: Loadings > 0.2 on first two principal components.

Feature	PC 1	PC 2
Interjection/emoticon/emoji	0.76	-0.27
Personal pronoun	0.36	0.32
Proper noun	-0.22	-0.54
Sentence-ending punctuation	-0.25	
Preposition	-0.27	0.20
Hashtag		0.41
Noun		-0.26
Adjective		-0.25
Determiner		-0.20

The strongest positive loadings on the first principal component are for two features with interpersonal interaction and stance orientation functions: Interjections/emoticons/emoji and the use of personal pronouns. Negative loadings are associated with features that typically relate to the presentation of information (proper nouns) and the organization of discourse (sentence-ending punctuation and prepositions).

The second principal component also shows a positive loading for personal pronouns and a negative loading (somewhat greater in magnitude than for the first component) on proper nouns, but positive loadings for prepositions and hashtags and negative loadings for nouns, adjectives, and determiners. Tokens tagged as interjections have a negative loading on the second principal component.

Both principal components seem to index interactive discourse, but with somewhat different focuses. It may be the case that the first principal component captures affect expression and stance orientation (for example, in tweets expressing affective content that include emoticons or emojis), while the second principal component may capture interactions that make reference to discourse external to the tweet messages themselves, such as through the use of hashtags.

The positions of the gendered subcorpora along the first two principal components are shown in Figure 3. The analysis shows clear functional separation between males and females along the first principal component: The male subcorpora all have negative values, while the female subcorpora have positive values. Gender separation along the second principal component is less distinct. Although the female subcorpora from Iceland, Denmark, Sweden and Finland exhibit higher values than the male subcorpora, the Norwegian female subcorpus is an outlier, with a negative value much lower than any those for the male subcorpora. An examination of the data reveals that the values for Norwegian females are strongly influenced by the extremely high Twitter activity of a single author whose posts tend to consist mainly of sequences of hashtags.

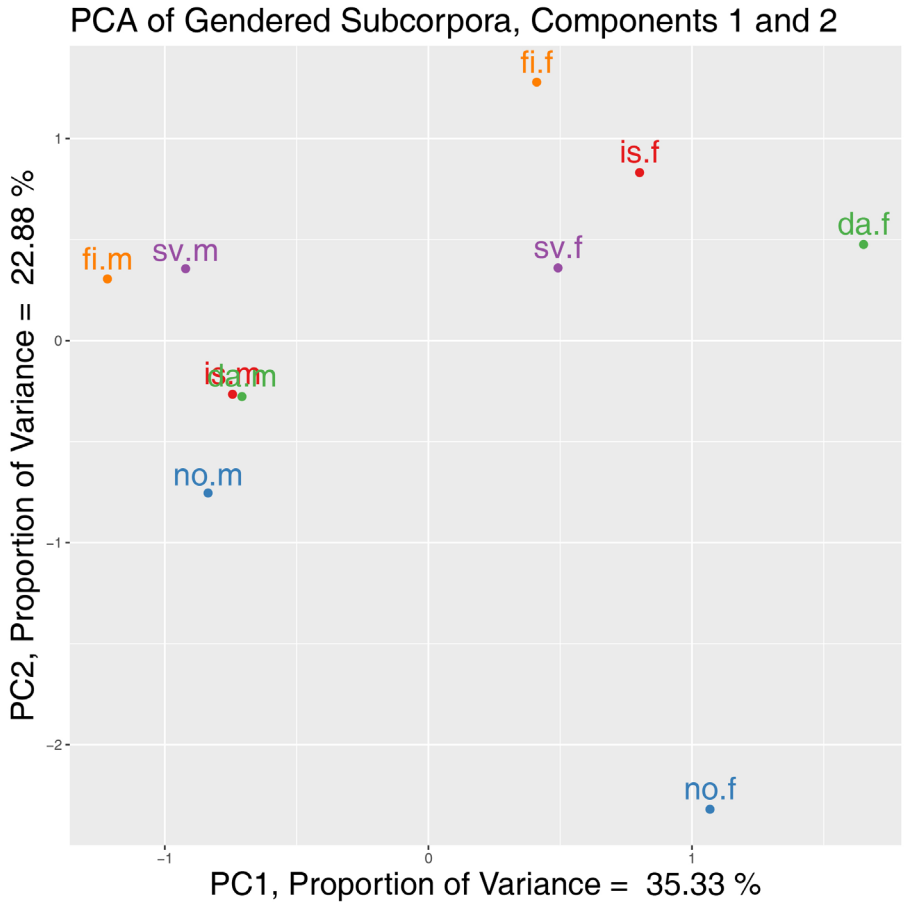

Figure 3: Loadings on components 1 and 2 of PCA for English subcorpora.

The distance between male and female subcorpora for the same country are also notable, and the Euclidean distance for the first two principal components is comparable for the individual Nordic countries. The genders are closer in Sweden and Finland and somewhat further apart in Iceland, Denmark, and Norway.

Component scores for the gendered subcorpora were calculated by summing the scaled frequencies (expressed in terms of standard deviation distance from the mean value for all ten subcorpora) of those components with weights > 0.2 on the first two components (see Biber 1988: 93—97).

Table 6: Component Scores for PC 1 and PC2.

		PC 1	PC 2
Iceland	male	6.19	11.28
	female	6.83	12.48
Norway	male	6.37	11.02
	female	6.57	12.25
Denmark	male	6.42	11.37
	female	6.89	11.51
Sweden	male	6.42	10.86
	female	7.00	11.81
Finland	male	5.74	11.33
	female	6.35	11.93

Here as well, a modest but clear functional separation is observable in the differences between male and female scores.

4 CONCLUSION

Corpora consisting of messages in English posted online collected from social media sites such as Twitter can shed light on the ways in which English continues to develop and diversify globally, especially in contexts where it has not traditionally been a language of daily communication. Data that has been appended metadata tags for location and disambiguated for author gender can provide insight into global English varieties and the relationships between language and gender in different geographical and social contexts.

While it is not surprising that English is extensively used on a global internet platform such as Twitter, the present research confirms high rates of use of English on Twitter in the Nordic countries attested in previous research. Overall, people in Denmark and Norway send more tweets in English than do those in Iceland, Sweden and Finland, and females more than males. It may be the case that the proportion of messages from the Nordic countries written in English on Twitter is increasing over time: For example, Mocanu et al. (2013) report rates of use for English in the Nordics in GPS-enabled tweets collected from 2010—2012. They find Iceland has 45%, Norway 24.6%, Denmark 40%, Sweden 18.1%, and Finland 27.1% English tweets.[14] This study finds similar values (slightly higher for Norway, Denmark and Sweden; slightly lower for Iceland and Finland), but considers not only GPS-tagged tweets (i.e. those with a populated *geo* field) but also those with a

14 http://www.twitterofbabel.org/

place value. Considering the fact that GPS-tagged tweets are typically sent on smartphones by users who are, on average, younger than the overall population and tend to use more English (see Pavalanathan and Eisenstein 2015), the data from the present study suggests and increase in English use in the Nordics over the past six years.

The results of the gender analysis in the present work complement those from previous corpus studies on English-language data collected from CMC or Twitter in Anglophone societies such as the United States: Females tend to use features such as personal pronouns, possessive pronouns or affect markers more often than males, whereas males use features such as punctuation, numbers/numerals, and nouns more than do females (Bamann, Eisenstein and Schnoebelen 2014). The same general pattern can be found in the present data set for English used on Twitter in the Nordic countries by persons with common Nordic names.

Multidimensional approaches based on factor analysis or principal components analysis have shown that differences in aggregate grammatical feature frequencies for national varieties of English can be interpreted in terms of communicative or discourse-functional dimensions (Biber 1988; 1995; Xiao 2009). The Nordic Twitter data used in this study was induced to reflect author gender, and the results show differentiation by gender along a first principal component, explaining a large proportion of variance in the data. The loadings on this component correspond to grammatical features whose discourse or communicative functions may contrast interactive stance orientation and affective content with informational and discourse organization functions – a finding comparable to the proposed "involved versus informational production" dimension found by Biber in a corpus of print media texts (1988: 107).

Although most work on differences in feature frequencies by gender has been conducted on L1 English data, there is some evidence for differential use of word classes by gender in other languages as well.[15] This study shows that gender-based differences in feature frequency in Twitter data from the Nordics matches up well with differences found in CMC and non-CMC data from Anglophone and non-Anglophone contexts.

It has been suggested that the small differences in aggregate Anglophone and non-Anglophone feature frequencies between males and females may reflect different orientations towards the use of communicative or discourse functions for the negotiation of affect maintenance or solidarity (Holmes 1998). Exploratory data analysis suggests that functional separation of English-language feature frequencies by gender can be observed for Nordic Twitter corpora with induced author gender. This tentative confirmation of some of the trends observed in CMC and Twitter data from L1 Anglophone contexts raises interesting questions as to

15 For French, see Schenk-van Witsen (1981). For French, Turkish, Indonesian and Japanese, see Ciot, Sonderegger and Ruths (2013).

the possible causes: Have cultural attitudes found in Anglophone contexts such as the United States been transmitted through the internet and other media to Northern Europe and become manifest in the patterning of grammatical features by Nordic people using English? Or is it the case that there may be underlying differences in interaction and communication style between the genders that are rooted not in cultural specifics, but aspects of human biology?

One interesting prospect for future investigation could thus be to investigate the extent to which the gender differentiation in grammatical type frequencies found in English-language data are also present in language data in the Nordic languages. Another possibility for future research, suggested by the presence of metadata fields in tweets that indicate direct responses to others, would be to combine aggregate feature frequency information by gender with user network information in order to gauge the relative contribution of each to differences in language. As English continues to evolve in diverse geographical as well as ever-more specialized technological contexts of CMC, the investigation of the relationship between language use and factors of demographic identity such as gender will continue to provide insights into our shared experience.

References

Allwood, Jens, 1998: Some frequency based differences between spoken and written Swedish. *Proceedings from the XVI:th Scandinavian conference of linguistics*. Turku, Finland. Department of Linguistics, University of Turku. http://sskkii.gu.se/jens/publications/docs076-100/084.pdf. (Last accessed 1 March 2017.)

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker and Jonathan Schler, 2007: Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday*, 12/9. http://pear.accc.uic.edu/ojs/index.php/fm/article/view/2003/1878. (Last accessed 1 March 2017.)

Bamann, David, Jacob Eisenstein and Tyler Schnoebelen, 2014: Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18/2. 135–160. http://onlinelibrary.wiley.com/doi/10.1111/josl.12080/full. (Last accessed 1 March 2017.)

Baron, Naomi S., 2004: See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology*, 23/4. 397–423.

Biber, Douglas, 1988: *Variation across speech and writing*. Cambridge University Press: Cambridge, UK.

Biber, Douglas, 1995: *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press: Cambridge, UK.

Blommaert, Jan, 2012: Supervernaculars and their dialects. *Dutch Journal of Applied Linguistics*, 1/1. 1–14.

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

Ciot, Morgane, Morgan Sonderegger and Derek Ruths, 2013: Gender inference of Twitter users in non-English contexts. *Proceedings of the 2013 conference on empirical methods in natural language processing*. Stroudsburg, PA: Association for Computational Linguistics. 1136–1145. http://www.aclweb.org/anthology/D13-1114. (Last accessed 1 March 2017.)

Coats, Steven, 2016: Grammatical feature frequencies of English on Twitter in Finland. Squires, Lauren (ed.): *English in Computer-mediated Communication: Variation, Representation, and Change.* Berlin: De Gruyter. 179–210. https://doi.org/10.1515/9783110490817-009. (Last accessed 1 March 2017.)

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing, 2014: Diffusion of lexical change in social media. *PLoS ONE* 9/1. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0113114. (Last accessed 1 March 2017.)

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan and Noah A. Smith, 2011: Part-of-speech tagging for Twitter: Annotation, features, and experiments. *Proceedings of the 49th Annual meeting of the association for computational linguistics: human language technologies*. Stroudsburg, PA: Association for Computational Linguistics. 42–47. www.ark.cs.cmu.edu/TweetNLP/gimpel+etal.acl11.pdf. (Last accessed 1 March 2017.)

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

Gustafson-Capková, Sofia and Britt Hartmann, 2008: *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University. https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf. (Last accessed 1 March 2017.)

Herring, Susan and John Paolillo, 2006: Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10/4. 439–459.

Holmes, Janet, 1998: Women's talk: The question of sociolinguistic universals. *Australian Journal of Communications* 20. 125–149.

Leetaru, Kalev H., Shaowen Wang, Guofeng Cao, Anand Padmanabhan and Eric Shook, 2013: Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18/5–6. http://firstmonday.org/article/view/4366/3654. (Last accessed 1 March 2017.)

Leppänen, Sirpa, Anne Pitkänen-Huhta, Tarja Nikula, Samu Kytölä, Timo Törmäkangas, Kari Nissinen, Leila Kääntä, Tiina Räisänen, Mikko Laitinen, Heidi Koskela, Salla Lähdesmäki and Henna Jousmäki, 2011: National Survey on the English Language in Finland: Uses, meanings and attitudes. *Studies in Variation, Contacts and Change in English* 5. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/05/evarieng-vol5.pdf. (Last accessed 1 March 2017.)

Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santorini, 1993: Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19/2. 313–330. http://dl.acm.org/citation.cfm?id=972475. (Last accessed 1 March 2017.)

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and J. Niels Rosenquist, 2011: Understanding the demographics of Twitter users. *Proceedings of the fifth international AAAI conference on weblogs and social media*. Menlo Park, CA: AAAI. 554–557. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816/3234. (Last accessed 1 March 2017.)

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang and Alessandro Vespignani, 2013: The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE* 8/4. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061981. (Last accessed 1 March 2017)

Newman, Matthew L., Carla J. Groom, Lori D. Handelman and James W. Pennebaker, 2008: Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45/3. 211–236. http://dx.doi.org/10.1080/01638530802073712. (Last accessed 1 March 2017)

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneid and Noah A. Smith, 2013: Improved part-of-speech tagging for online conversational text with word clusters. *Proceedings of NAACL-HLT*. Stroudsburg, PA: Association for Computational Linguistics. 380–390. http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.naacl13.pdf. (Last accessed 1 March 2017.)

Page, Ruth, 2012: The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. *Discourse & Communication* 6/2. 181–201.

Pavalanathan, Umashanthi and Jacob Eisenstein, 2015: Confounds and consequences in geotagged Twitter data. http://arxiv.org/pdf/1506.02275v2.pdf. (Last accessed 1 March 2017,)

Pennacchiotti, Marco and Ana-Maria Popescu, 2011: A machine learning approach to Twitter user classification. *Proceedings of the fifth international AAAI conference on weblogs and social media*. Menlo Park, CA: Association for the Advancement of Artificial Intelligence. 281–288. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2886/3262. (Last accessed 1 March 2017.)

Rao, Delip, David Yarowsky, Abhishek Shreevats and Manaswi Gupta, 2010: Classifying latent user attributes in Twitter. *Proceedings of the 2nd international workshop on search and mining user-generated contents*. New York, NY: Association for Computing Machinery. 37–44. http://dl.acm.org/citation.cfm?doid=1871985.1871993. (Last accessed 1 March 2017.)

Roesslein, Josh, 2015. *Tweepy*. Python programming language module. http://www.tweepy.org. (Last accessed 1 March 2017)

Schenk-van Witsen, Rosalien, 1981. Les différences sexuelles dans le français parlé: Une étude-pilote des différences lexicales entre hommes et femmes. *Langage et Societé*, 17/1. 59–78. http://www.persee.fr/doc/lsoc_0181-4095_1981_num_17_1_1328. (Last accessed 1 March 2017.)

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

Vandergriff, Ilona, 2013: Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics* 51. 1–12. http://www.sciencedirect.com/science/article/pii/S037821661300057X. (Last accessed 1 March 2017.)

Xiao, Richard, 2009: Multidimensional analysis and the study of world Englishes. *World Englishes* 28/4. 421–450.

Zappavigna, Michele, 2011: Ambient affiliation: A linguistic perspective on Twitter. *New Media and Society* 13/5. 788–806.