Grammatical Frequencies and Gender in Nordic Twitter Englishes

Steven Coats
University of Oulu, Finland
English Philology, Faculty of Humanities, 90014 University of Oulu, Finland
Email: steven.coats@oulu.fi

Abstract
English is increasingly used for online communication in many contexts in which it is not the primary local language, particularly on social media platforms with global extent such as Twitter. The grammatical properties of online and Twitter Englishes, however, have mainly been considered in L1 contexts, as have correlations between gender and some grammatical features. In this study, the correlation of grammatical types (parts of speech) and gender is undertaken for English-language Twitter messages originating from the Nordic countries. A corpus of geo-located English-language Twitter messages was created by accessing the Twitter Streaming API. After disambiguating author gender and applying part-of-speech tags, the relative frequencies of grammatical items were determined and those with significant gender divergence identified. Principal components analysis shows some gender-based separation of discourse in the Nordic countries in terms of grammatical features. The analysis supports previous findings pertaining to gendered differences in English and sheds light on how English continues to evolve in online environments.

Keywords: corpus linguistics, Twitter, world Englishes, language and gender

1. Introduction and Background
Technological developments can affect the way we interact with one another, and the recent shift towards mediated, text-based communication in online environments provides opportunities for the study of English varieties in global contexts. Although the status of English as the world’s principal lingua franca continues to consolidate, its use in global computer-mediated communication (CMC), especially in non-L1 environments, exhibits a diversity of orthography, lexis, and grammar that 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 sought to characterize the communicative and discourse functions of Twitter language (Page 2012; Zappavigna 2011; Squires 2015 for an overview). The extensiveness of Twitter data, its public availability, and the richness of the associated metadata have allowed for geographical analyses (Leetaru et al. 2013; Mocanu et al. 2014) and dialectological and sociolinguistic analyses of English (Eisenstein et al. 2014; Bamann, Eisenstein and Schnoebelen 2014).

Some previous studies of English-language CMC and Twitter have found different rates of use of particular word classes by males and females. For example, it has been found that females use more personal pronouns, more modal verbs, and more 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; Bamann, Eisenstein and Schnoebelen 2014). For the most part, however, analysis of Twitter English has been conducted on data without consideration of its geographical provenance, or on data gathered from Anglophone national contexts, mostly in the United States. Knowledge of English is extensive in the Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland, countries with well-developed economies and high levels of educational attainment, to such an extent that it has been suggested that their national languages are becoming linguistic systems with “restricted functional range” (Görlach 2002: 16). Although research has addressed language use on Twitter by country (e.g. Mocanu et al. 2013), and work exists on grammatical feature frequencies in Nordic non-CMC genres (e.g. for Swedish in Allwood 1998), studies of feature frequencies in English from non-L1 environments have been few, and the relationship between author gender and feature frequency in CMC language has not yet been investigated in detail in Nordic contexts, whether in local languages or English.1

In this study an approach based in part on multidimensional analysis (Biber 1988; 1995) is taken. 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. With a principal components analysis, the underlying association between feature frequencies and gender is established.

2. Data Collection and Processing
Data was collected in .json format from Twitter’s Streaming API during May 2016 by utilizing a scripting library in Python.2 The raw .json data was filtered for the tweet text (the “status update”) and the metadata fields author_name, screen_name, time, id, lang (language), country, and the latitude and longitude coordinates.

---

1 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 The Tweepy library (Roesslein 2015) was used (https://github.com/tweepy/tweepy).
2.1. Geolocation

The collection script selected only tweets with a populated place object that originated within a bounding box circumscribing the territorial boundaries of the Nordic countries (longitude -26 to 32, latitude 53 to 72; see Figure 1).

![Figure 1: Area from within tweets with geographical coordinates were collected from the API.](image)

Each tweet was assigned exact latitude/longitude coordinates. From the 2.155 million tweets collected by the script, 302,737 were retained to create subcorpora from the Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland, based on the country values within the place field. For further analysis, two subcorpora were prepared for each country by filtering the data according to the lang field in the tweet object: one consisting of tweets in the principal national language, and one of tweets in English. Tweets originating from outside the Nordic countries and in other languages were not further considered. The English data comprised 101,956 tweets and 1,475,553 tokens.

2.2. Gender Disambiguation

Unlike some social media platforms, Twitter does not provide a profile entry where gender is to be identified nor require users to otherwise supply gender information. Therefore, gender was disambiguated for tweets based on gender-name associations (Rao et al. 2010; Mislove et al. 2011).\(^5\)

Lists of the most frequent given names in the Nordic countries were obtained from the corresponding national statistical offices. The author_name field for each user was then filtered 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. Users matching both male and female names were discarded. The method assigned gender to 39% of Iceland, 50% of Norway, 61% of Denmark, 47% of Sweden, and 62% of Finland tweets.\(^6\)

2.3. 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 subcorpora and apply part-of-speech tags using a subset of the Penn Treebank tagset (Marcus, Santorini and Marcinkiewicz 1993) and additional tags for the Twitter-specific features username, hashtag, and retweet. The tool is somewhat tolerant of the non-standard orthography typical of Twitter messages.

3. Analysis and Discussion

The linguistic profiles of the subcorpora were determined and the relationship between gender and individual grammatical features assessed using t-tests. Principal components analysis was used to gauge the extent to which males and females 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).\(^7\)

In Iceland, Norway and Denmark, males use the national language on Twitter more than do females; Females use English more. This difference is most pronounced for Denmark. In Sweden and Finland the rates of language use by gender are similar, with males using slightly more English and females the national languages.

3.2. Correlation of Grammatical Features, Country and Gender

34 of the PoS tags were applied at least once in all of the ten gendered subcorpora. For each subcorpus, the relates...
in English on Twitter are personal pronouns, compared to 4.28% by Swedish males.

![Figure 2: Percent of all tokens by feature for features that differ significantly by gender from Sweden.](image)

### Table 2: Grammatical features by gender

Gendered differences were also considered by country and feature. For Sweden, for example, the distribution of those features for which a significant difference by gender was detected is depicted in Figure 3. The differences between males and females are not large (Cohen's $d \leq 0.24$), but statistically significant according to a t-test of population means: E.g. 5.83% of all words used by Swedish females

![Figure 3: Percent of all tokens by feature for features that differ significantly by gender from Sweden.](image)

### Table 3: Loadings $\geq |0.2|$ on first two principal components

For the features with the strongest loadings on the first principal component, grammatical types with interpersonal interaction and stance orientation functions (personal pronouns, 1st- and 2nd-person singular present verb forms, and interjections\(^6\)) have the strongest positive loadings, while

---

\(^6\) The Carnegie-Mellon Twitter tagger also assigns the interjection tag to emoticons, word types that are often associated with the expression of emotional affect (Vandergriff 2013).
features with informational and text-organizational functions (nouns, proper nouns, and sentence-ending punctuation) have the strongest negative loadings.

![PCA of Gendered Subcorpora, Components 1 and 2](image)

The positions of the gendered subcorpora along the first two principal components are shown in Figure 4. The analysis suggests some functional separation between males and females in Nordic Twitter Englishes as they are manifest in terms of grammatical feature frequencies: The male corpora all have negative values in the first principal component, while the female corpora have positive values. Gender separation along the second principal component is also manifest, although not as pronounced. In terms of the individual Nordic countries, the distance between males and females is larger for Iceland and Norway, while it is somewhat smaller for Denmark, Sweden and Finland.

4. Conclusion and Summary

Geographically specified and gender-induced corpora of online Englishes compiled from social media sites such as Twitter 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. The results of this study bear upon research into online English varieties and the relationship between language and gender.

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 (cf. Mocanu et al. 2013). Overall, persons in Denmark and Norway send more tweets in English, the Nordic countries (cf. Mocanu et al. 2013). Overall, per-

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). In this study, Nordic Twitter data that have been induced to reflect author gender exhibit differentiation by gender along a first principal component, explaining the majority of variance in the data (58.9%). 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 (1988: 107). Most work on differences in feature frequencies by gender has been conducted on L1 English data, but there is some evidence for differential use of word classes by gender in other languages. This study shows that similar differences exist for (presumable) non-L1 English users on Twitter. It has been suggested that the small differences in aggregate grammatical 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 for Nordic Twitter corpora with induced author gender, functional separation of English-language feature frequencies by gender can be observed. A tentative confirmation of some of the trends observed in CMC and Twitter data from L1 Anglophone contexts raises interesting questions as to the possible causes. Future work could further investigate this topic by exploring the extent to which gender differentiation is present in Twitter material in the Nordic languages, and whether language transfer phenomena may influence the large-scale patterning of linguistic elements in non-L1 online Englishes.

5. References


