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.

Each tweet was assigned exact latitude/longitude coordinates.[3] 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.[4] 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 rela-

[3] Most Twitter users select a `place` when registering with the service; the coordinates of the `place` are then automatically assigned by Twitter as a lat-long bounding box in tweet metadata. Some users additionally opt to broadcast precise GPS coordinates with each status update. For tweets without precise geographical coordinates, location was induced by calculating the center of the bounding box circumscribing the `place` field. Correlation of the precise GPS coordinates and the induced coordinates based on centering the `place` entity was 0.993, as the `place` entity is almost always populated by a bounding box circumscribing a small area such as a city. See also Leetaru et al. (2013).

[4] For Finland, corpora were also created for the country's second official language, Swedish.

[5] Latent attribute inference using Twitter data manually tagged for gender is a popular topic in machine learning (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, but both approaches can be used to investigate links between language use and gender.

[6] The differences are due in part to the somewhat different name frequency information obtained from the national statistical offices. For example, only 395 unique given names were obtained from Iceland, but 1190 from Norway, 5382 from Denmark, 1704 from Sweden, and 7899 from Finland.

[7] The Twitter automatic language detection algorithm classifies both *Riksmål* and *Nynorsk* with the language code `no`, "Norwegian". For Finland, the percentage shown includes messages messages in the national languages of Finnish and Swedish.

		Nat. lang.	English	Other
Iceland	males	80.8	9.8	9.4
	females	71.5	17.6	10.9
Norway	males	46.6	28.9	24.5
	females	37.3	40.0	22.7
Denmark	males	45.4	40.0	14.6
	females	25.7	52.5	21.8
Sweden	males	61.9	24.5	13.6
	females	63.8	23.8	12.4
Finland	males	57.2	28.8	14.0
	females	58.5	25.0	16.5

Table 1: Percent tweets by country, gender and language.

tive frequency of each tag was calculated. To determine whether features were preferred by males or females, a t-test of population means was conducted on the basis of the mean standardized value for males and for females in all subcorpora. Of the 34 features, ten exhibited significant ($p < 0.05$) differences in use between males and females: Sentence-ending punctuation, numbers or numerals, proper nouns, and gerund or present participle forms were more frequently utilized by males, while personal pronouns, possessive pronouns, adverbs, interjections, usernames, and past particples were more likely to be used by females (Table 2).

	Feature	Gender	p-value	Signif.
1	Quotation marks (")	m	0.320	
2	Left bracket (()	m	0.080	
3	Right bracket ())	m	0.089	
4	Comma	m	0.098	
5	Period (. ? !)	m	0.010	*
6	Other punctuation (: ; ... + - = <> [ ])	m	0.245	
7	Coordinating conjunction	f	0.269	
8	Number	m	0.040	*
9	Determiner	m	0.416	
10	Hashtag	f	0.758	
11	Preposition or subordinating conjunction	m	0.502	
12	Adjective	m	0.405	
13	Comparative adjective	f	0.848	
14	Superlative adjective	f	0.213	
15	Modal verb	f	0.695	
16	Noun, singular or mass	m	0.275	
17	Proper noun	m	0.014	*
18	Plural noun	m	0.596	
19	Personal pronoun	f	0.005	*
20	Possessive pronoun	f	0.005	*
21	Adverb	f	0.036	*
22	Phrasal particle	m	0.449	
23	*to*	f	0.596	
24	Interjection	f	0.018	*
25	Username (preceded by)	m	0.168	
26	Verb, base form	f	0.007	*
27	Verb, past tense	f	0.441	
28	Verb, gerund or present particle	f	0.866	
29	Verb, past participle	m	0.022	*
30	Verb, non-3rd person singular present	f	0.001	*
31	Verb, 3rd person singular present	f	0.292	
32	Wh-determiner	m	0.094	
33	Wh-pronoun	f	0.934	
34	Wh-adverb	f	0.106	

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 in English on Twitter are personal pronouns, compared to 4.28% by Swedish males.

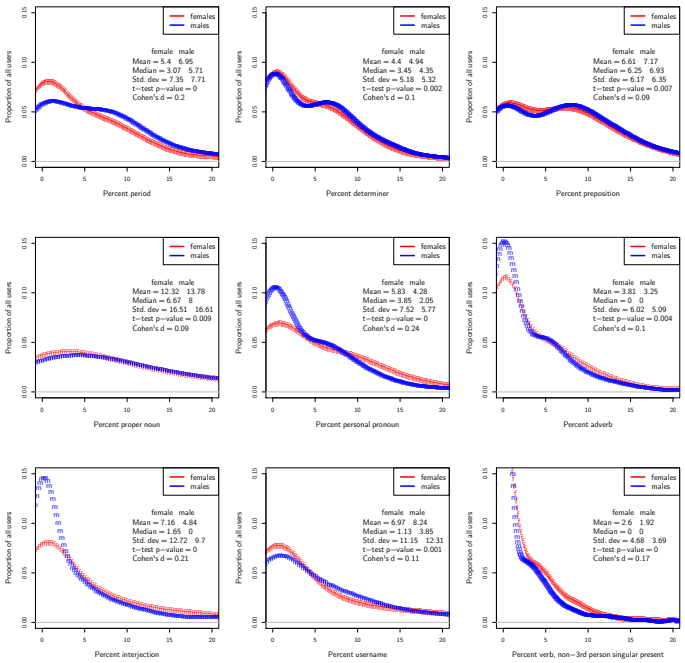

Figure 2: Percent of all tokens by feature for features that differ significantly by gender from Sweden.

3.3. Principal Components Analysis

In order to explore underlying patterning of the variance in the data, a principal components analysis was conducted on a covariance matrix of the normalized frequencies of the 34 variables for the ten English subcorpora (a male and a female subcorpus for each of the five Nordic countries). The first two components capture 70.8% of the variance in the data. The strongest loadings ($\geq |0.2|$) on the first two components are shown in Table 3.

Feature	PC1	PC2
Personal pronoun	0.60	-0.21
Interjection	0.31	0.34
Verb, non-3rd person singular present	0.21	
Period (. ? !)	-0.28	0.28
Noun, singular or mass	-0.25	-0.51
Proper noun	-0.45	
Comma		0.38
Number		0.34
Username		0.23

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[8]) have the strongest positive loadings, while

[8]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.

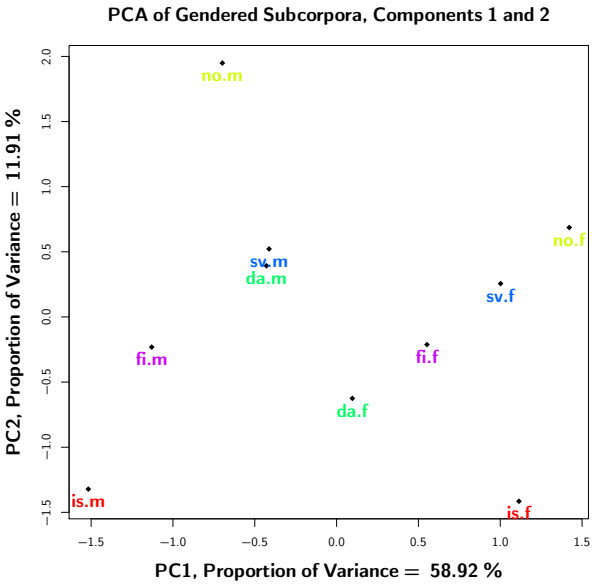

Figure 3: Loadings on components 1 and 2 of PCA for English subcorpora.

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 complied 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, and females more than males.

In the present work, gender analysis reinforces findings from previous corpus studies and research into L1 Twitter or CMC English: Females tend to use features such as personal pronouns, possessive pronouns or affect markers more than males, whereas males use features such as determiners, numbers/numerals, and nouns more than do females (Bamann, Eisenstein and Schnoebelen 2014). This patterning holds true for English used on Twitter in the Nordic countries by persons with common Nordic names, many of whom are likely non-L1 English users.

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.[9] 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

Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. In *Proceedings from the XVI:th Scandinavian Conference of Linguistics*, Turku, Finland. Department of Linguistics, University of Turku.

Argamon, S., Koppel, M., Pennebaker, J., and Schler, J. (2007). Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday*, 12(9).

Bamann, D., Eisenstein, J., and Schnoebelen, T. (2014).

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

Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics*, 18(2):135–160.

Baron, N. 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, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, UK.

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

Blommaert, J. (2012). Supervernaculars and their dialects. *Dutch Journal of Applied Linguistics*, 1(1):1–14.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Stroudsburg, PA. Association for Computational Linguistics.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PLoS ONE*, 9(1).

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Stroudsburg, PA. Association for Computational Linguistics.

Görlach, M. (1995). *Still More Englishes*. John Benjamins, Amsterdam.

Gustafson-Capková, S. and Hartmann, B. (2008). *Manual of the Stockholm Umeå Corpus version 2.0*. Stockholm University.

Herring, S. and Paolillo, J. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

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

Leetaru, K. H., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5/6).

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Mislove, A., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. In *Proceedings of ICWSM*, pages 554–557, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., and Vespignani, A. (2013). The Twitter of babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4).

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390. NAACL-HLT.

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

Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to Twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 281–288, Menlo Park, CA. Association for the Advancement of Artificial Intelligence.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, pages 37 – 44. ACM.

Roesslein, J. (2015). Tweepy. Python programming language module.

Schenk-van Witsen, R. (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.

Squires, L. (2015). Twitter: Design, discourse, and implications of public text. In Alexandra Georgakopoulou et al., editors, *The Routledge Handbook of Language and Digital Communication*, pages 239–256. Routledge, London and New York.

Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics*, 51:1–12.

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

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