Nordic Englishes on Twitter

Steven Coats
English Philology, University of Oulu, Finland
Digital Humaniora i Norden
16 March 2016

ULEÅBORGS UNIVERSITET
УЛЕАБОРГСКИЙ УНИВЕРСИТЕТ

Outline

1. Contexts of the present research
2. Data collection and processing
   - Twitter API and geo-encoded tweets
   - PoS tagging
   - Gender disambiguation
3. Results
   - Language distribution
   - Gender differences in use
     - Articles and personal pronouns, by gender and country
     - All grammatical features, by gender
     - Results: principal components analysis and clustering of aggregate lexical and grammatical feature frequencies
4. Summary

Contexts of the present research

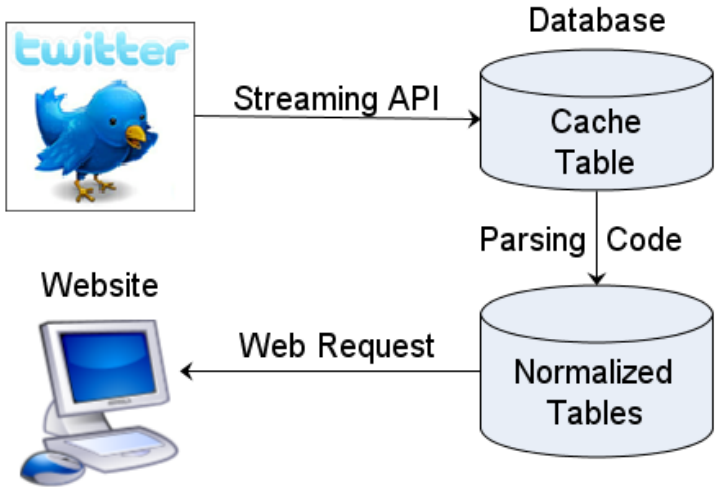
- English as it is used on Twitter in Northern Europe: Online Englishes and the status of English in (traditionally) non–Anglophone societies ("Expanding Circle", Kachru 1990, 1992)
- Categorization of language varieties based on aggregate frequencies of linguistic features (Biber 1988, 1995, 2006; Biber and Conrad 2009; Burrows 2002)
- Gendered differences in the language of computer-mediated communication (Baron 2004; Herring and Paolillo 2006; Herring 2013; Bamann, Eisenstein and Schnoebelen 2014)

Multi-feature/multidimensional analysis: Comparing Nordic Twitter Englishes

1. Create subcorpora of English-language Twitter messages for categories of interest (5 principal Nordic countries, US, males and females for each country)
2. Identify a large number of **lexical or grammatical features** in the corpora that can be **counted**
   - Lexical type frequencies or PoS frequencies
3. Test for differences between subcorpora for individual features
4. Calculate aggregate distances between the subcorpora
5. Use **principal components analysis, clustering** and similar techniques to explore underlying patterns of variation

Data collection and filtering

- Streaming API: open a connection and let the data pile up!
  - Unlimited Twitter stream ("firehose") is proprietary big data, only available to companies working in the "Twitter ecosystem"
  - 1% stream available to all
- Tweepy: python script access to Twitter (Roesslein 2015)
- Determining tweet country and region of origin
  - User-defined "location" entity vs. geo-coordinates (Pavalanathan and Eisenstein 2015)
  - Access levels, extent of geo-encoded user messages (1.6% of tweets according to Leetaru et al. 2013)
  - Filtering using data from GADM and packages in *R* (maptools, mapdata)
- Removing automated tweets from bots and unrenderable Unicode hexidecimal sequences
- Selecting only English-language messages



(https://fird0s.files.wordpress.com/2014/04/tweetcache.png)

Corpus summary statistics

	Tweets	Tokens
Iceland	8390	101,342
Denmark	23,571	274,726
Norway	35,298	321,670
Sweden	108,677	829,474
Finland	23,673	242,618
US	296,954	3,270,871

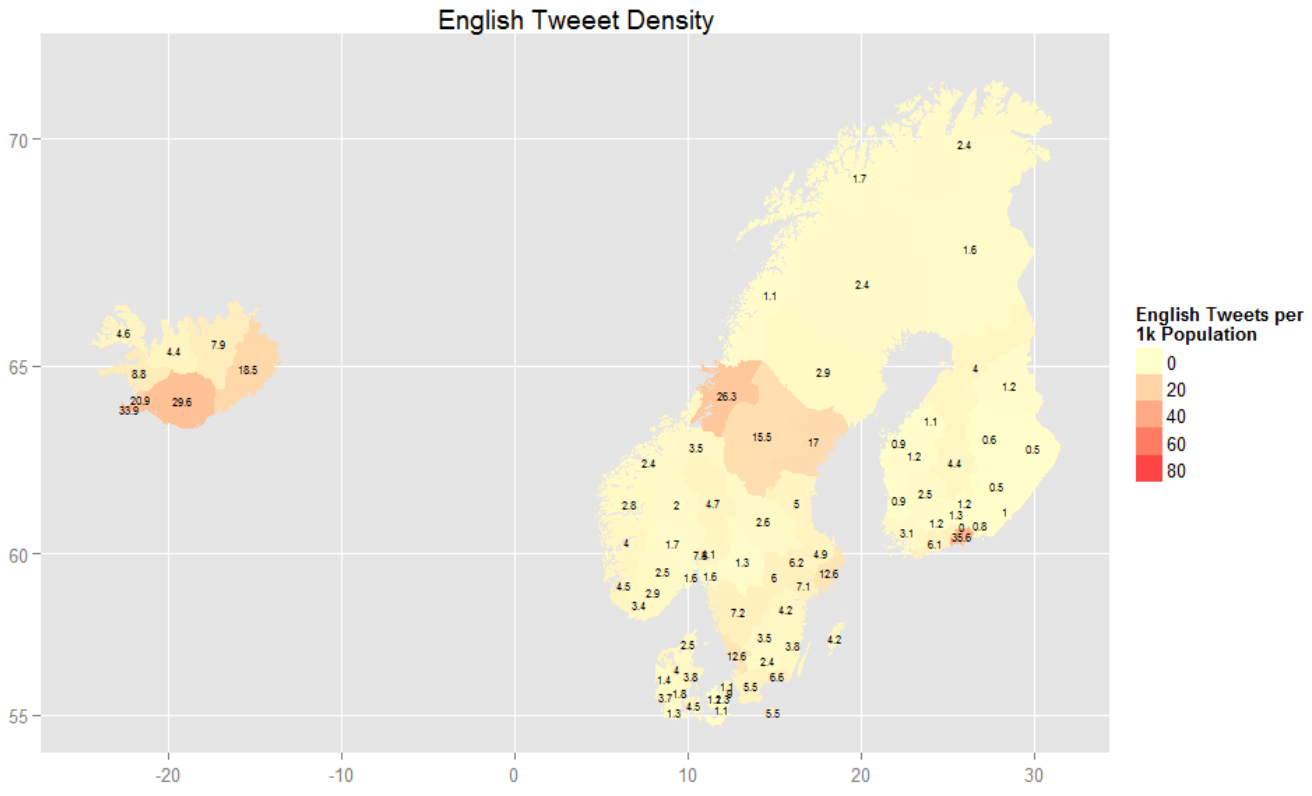
Percentage of tweets by language (comparison with Mocanu et al. 2013)

Iceland			Norway			Denmark			Sweden			Finland		
EN	45.0	34.1	NO	48.8	37.1	DK	45.0	33.5	SV	70.3	60.6	EN	56.4	18.3
IS	39.0	48.4	EN	24.6	35.6	EN	40.0	42.3	EN	18.1	23.2	FI	27.1	69.5
			DK	18.0	19.5	NO	6.2	2.2	NO	1.0	1.1	SV	4.1	1.0
			SV	2.9	3.4	SV	1.3	1.4				RU	3.9	0.9

- Norway, Denmark, and Sweden: Somewhat reduced use of principal national language, somewhat increased use of English
- Iceland: Increased use of Icelandic
- Finland: least use of English (2013 data seems off?)

English Twitter activity

- English-language activity is somewhat higher in capital regions
- Clear geographical pattern not evident
- Corpus size considerations



Part-of-Speech tagging

- Carnegie-Mellon Twitter PoS Tagger (Gimpel et al. 2011; Gimpel et al. 2013, Owoputi et al. 2013)
- Tagger model applies Penn Treebank tags (Marcus et al. 1993) plus tags for CMC/Twitter-specific types (emoticons, usernames, retweets, URLs, hashtags)

```
217599  im  PRP 0.9811
217600  kinda   RB  0.9833
217601  worried VBN 0.7087
217602  because IN  0.9892
217603  my  PRP$    0.9966
217604  nose    NN  0.9941
217605  is  VBZ 0.9958
217606  huge    JJ  0.9778
```

```
230769  you PRP 0.9983
230770  did VBD 0.9267
230771  a   DT  0.9949
230772  damn    RB  0.5180
230773  good    JJ  0.9860
230774  job NN  0.9900
230775  as  IN  0.8833
230776  always  RB  0.9933
230777  :-) UH  0.9241
```


PoS tags applied by CMU Twitter Tagger

Number	Tag	Description	Number	Tag	Description
1.	-LRB-	Left-hand bracket	22.	NNS	Noun, plural
2.	-RRB-	Right-hand bracket	23.	NNP	Proper noun, singular
3.	''	Quotation mark (")	24.	NNPS	Proper noun, plural
4.	,	Comma	25.	PRP	Personal pronoun
5.	.	Period (. ? !)	26.	PRP$	Possessive pronoun
6.	:	Punctuation (: ; ... + - = < > / [ ] ~)	27.	RB	Adverb
7.	HT	Hashtag	28.	RBR	Adverb, comparative
8.	RT	Retweet	29.	RBS	Adverb, superlative
9.	URL	Universal Resource Locator	30.	RP	Particle
10.	USR	Username (preceded by @)	31.	SYM	Symbol
11.	CC	Coordinating conjunction	32.	TO	to
12.	CD	Cardinal number	33.	UH	Interjection/emoticon
13.	DT	Determiner	34.	VB	Verb, base form
14.	EX	Existential there	35.	VBD	Verb, past tense
15.	FW	Foreign word	36.	VBG	Verb, gerund or present participle
16.	IN	Preposition or subordinating conjunction	37.	VBN	Verb, past participle
17.	JJ	Adjective	38.	VBP	Verb, non-3rd person singular present
18.	JJR	Adjective, comparative	39.	VBZ	Verb, 3rd person singular present
19.	JJS	Adjective, superlative	40.	WDT	Wh-determiner
20.	MD	Modal verb	41.	WP	Wh-pronoun
21.	NN	Noun, singular or mass	42.	WRB	Wh-adverb

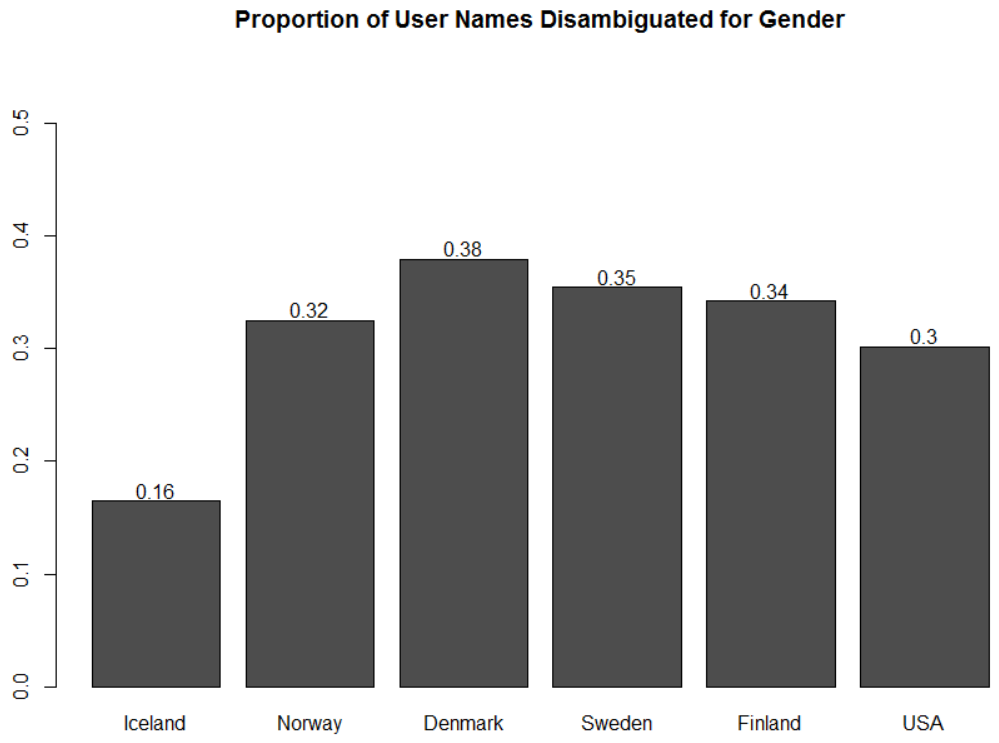
Gender disambiguation

- For each country, get a list of the 200 most common first names for each gender
- Filter tweet entity *tweet_author_screenname* for character sequences that correspond to the names
  - E.g. for tweets from Sweden:
  - Username *JohanLindman* → include in corpus of Swedish male tweets
  - Username *Twenty20Xxx* → Do not include in gendered subcorpora (etc.)
  - Exclude usernames that match male and female strings (e.g. *JohannaLindman*)

(cf. Mislove et al. 2011, Bamann et al. 2014)

Gender disambiguation

- Approximately 1/3 of user messages could be associated with gender in this manner
- Iceland: Fewer names in list; tendency not to use Icelandic names as author_screenname (?)
- This is an aggregate method but certainly not absolute

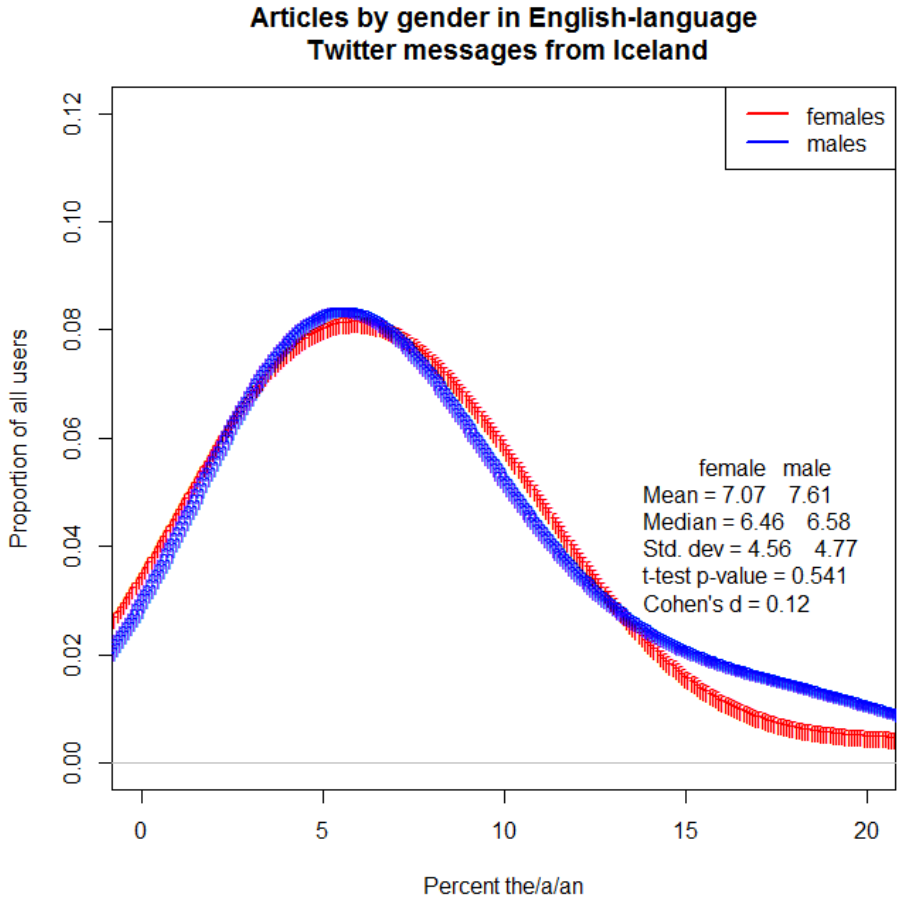


Gender differences in PoS frequencies

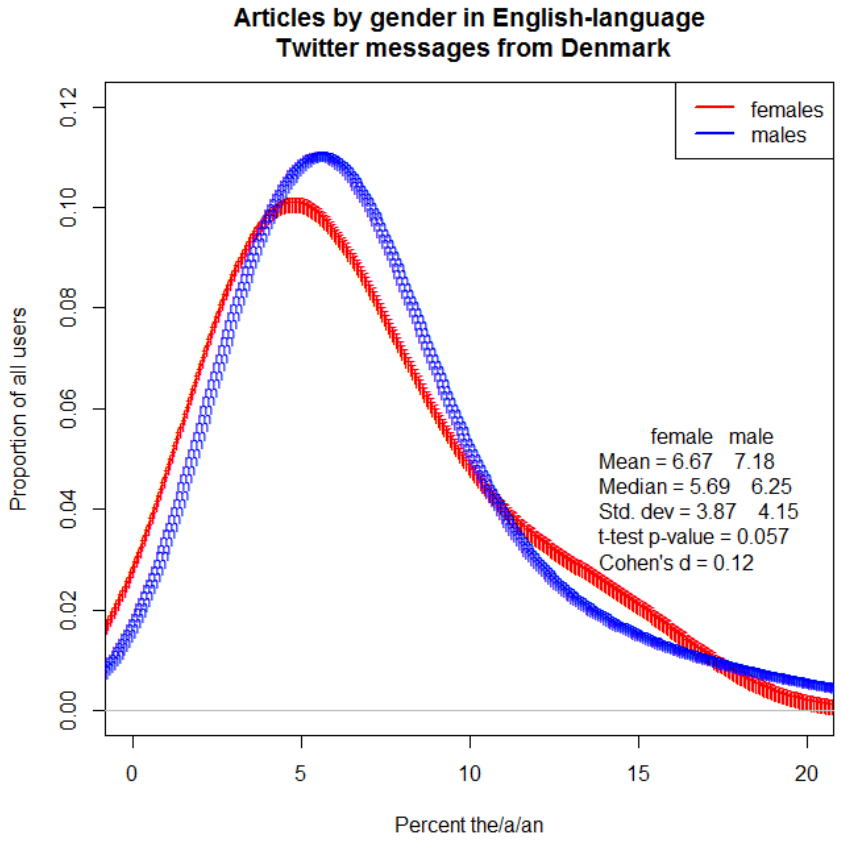
- Previous studies of CMC corpora (Baron 2004; Herring and Paolillo 2006; Argamon et al. 2007; Bamann, Eisenstein and Schnoebelen 2014) have shown different rates of use of particular word classes by males and females
  - Females use e.g. more personal pronouns, more modal verbs, and more emoticons
  - Males use e.g. more determiners such as articles or demonstrative pronouns and more numbers or numerals
- We can compare the total data set for male vs. female usage
- Some of these differences in use are significant for individual countries but not for others

Articles: Iceland

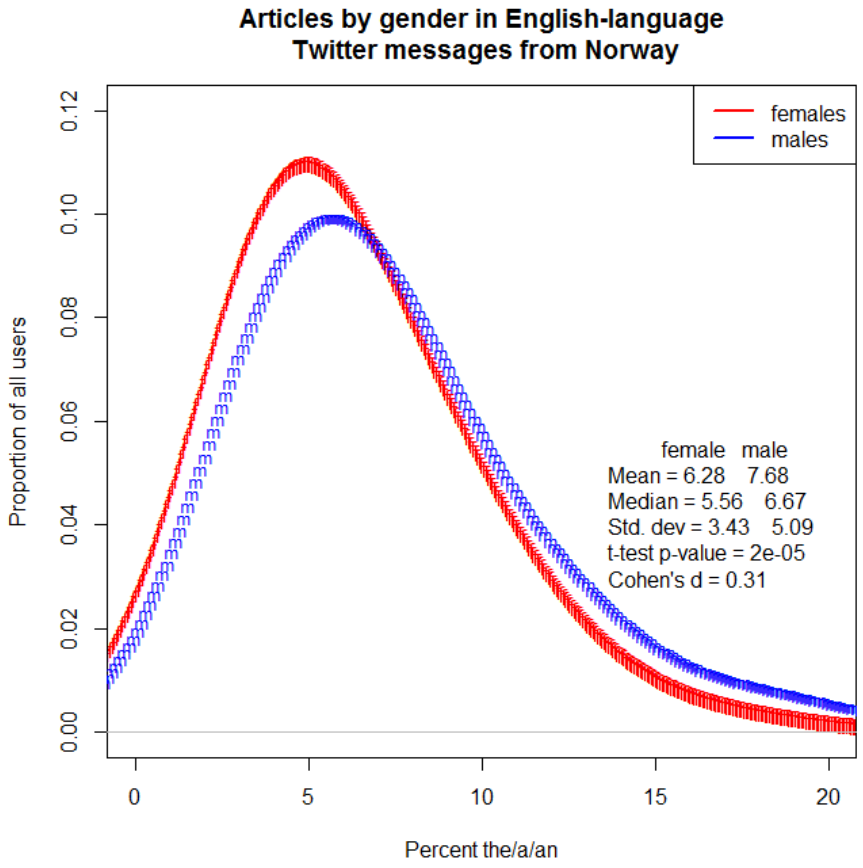
- Plotting the percent of tokens belonging to a particular word class vs. proportion of male and female users for individual countries reveals some differences



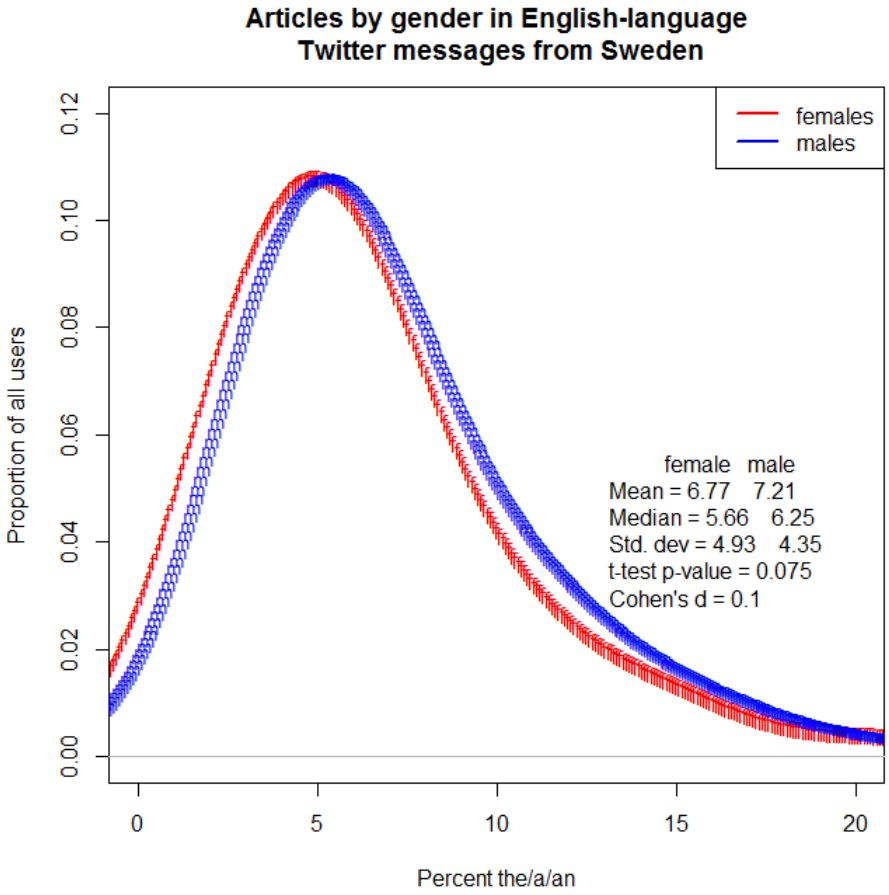
Articles: Denmark



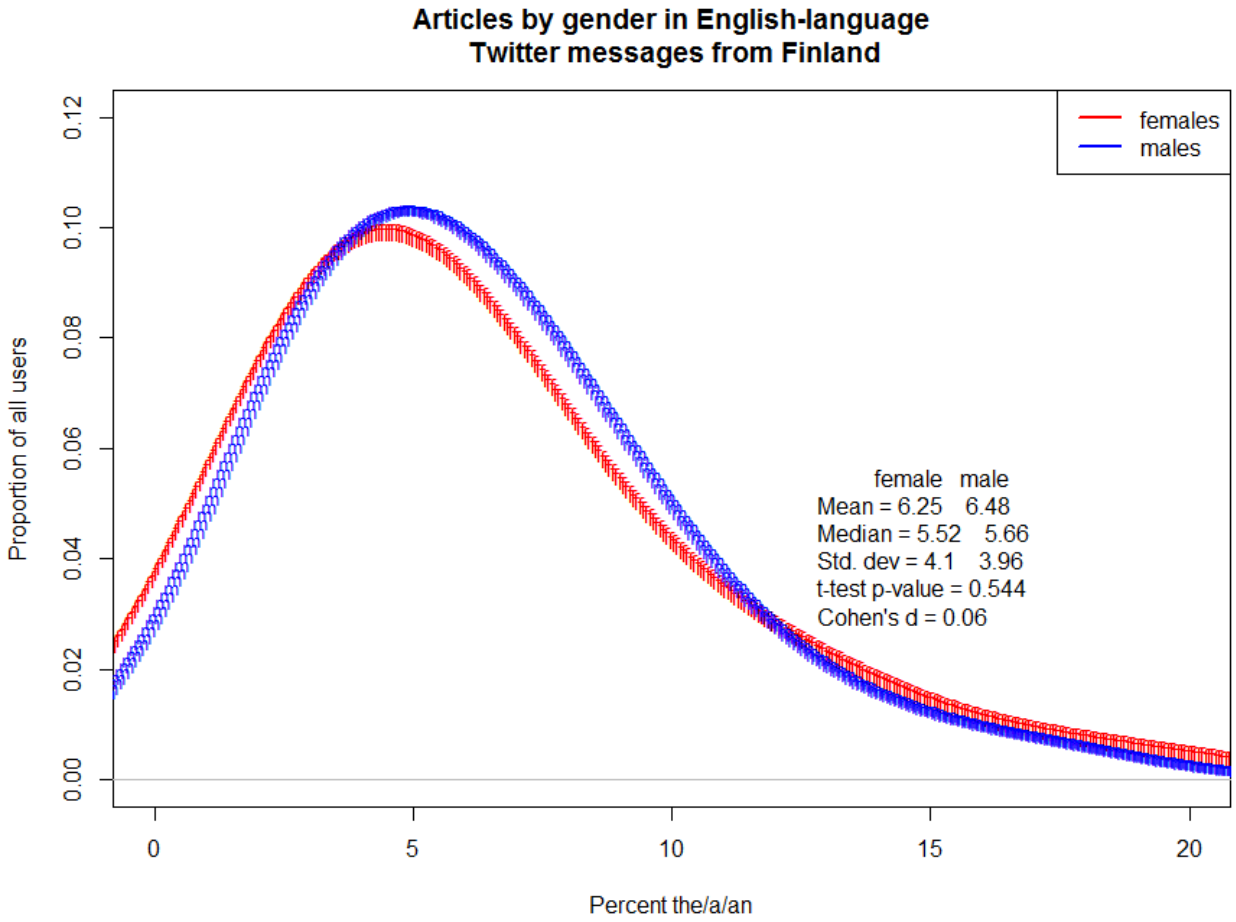
Articles: Norway



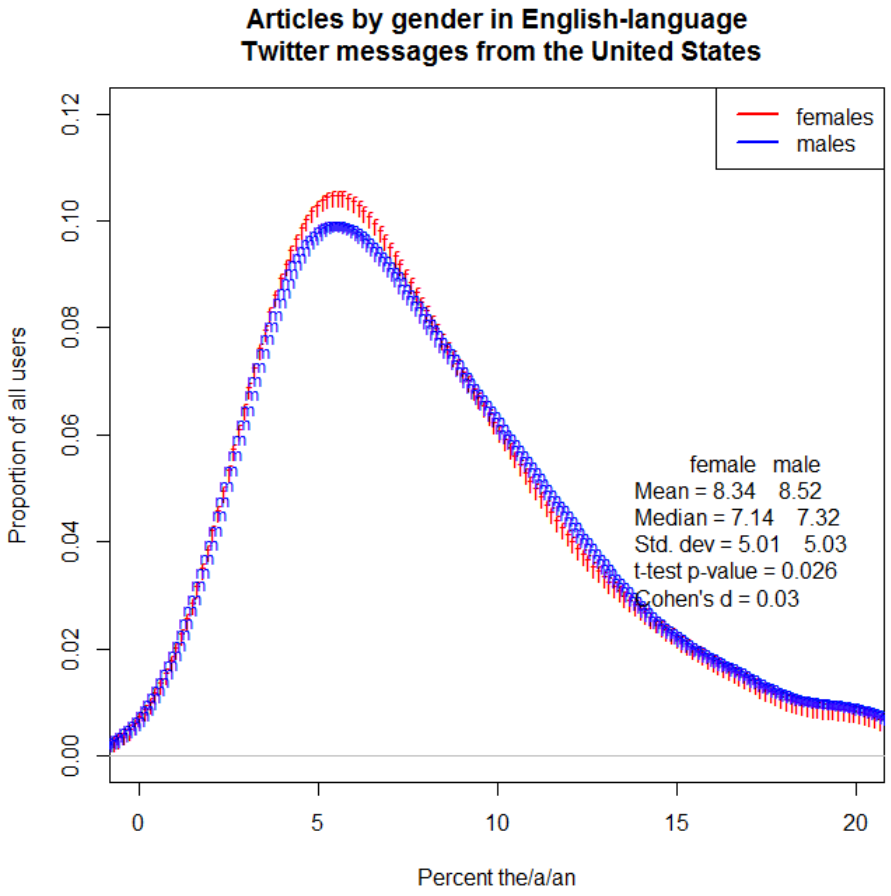
Articles: Sweden



Articles: Finland



Articles: US

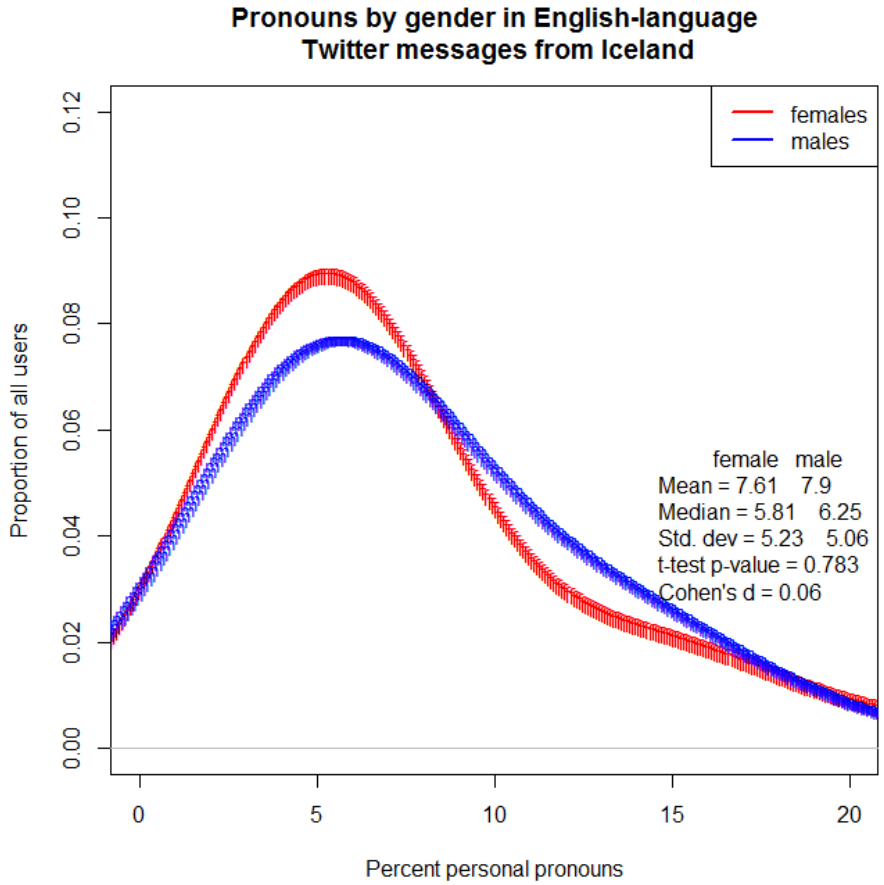

Articles by gender in English-language
Twitter messages from the United States
females
males
Proportion of all users
0.00
0.02
0.04
0.06
0.08
0.10
0.12
female male
Mean = 8.34 8.52
Median = 7.14 7.32
Std. dev = 5.01 5.03
t-test p-value = 0.026
Cohen's d = 0.03
0
5
10
15
20
Percent the/a/an

Articles

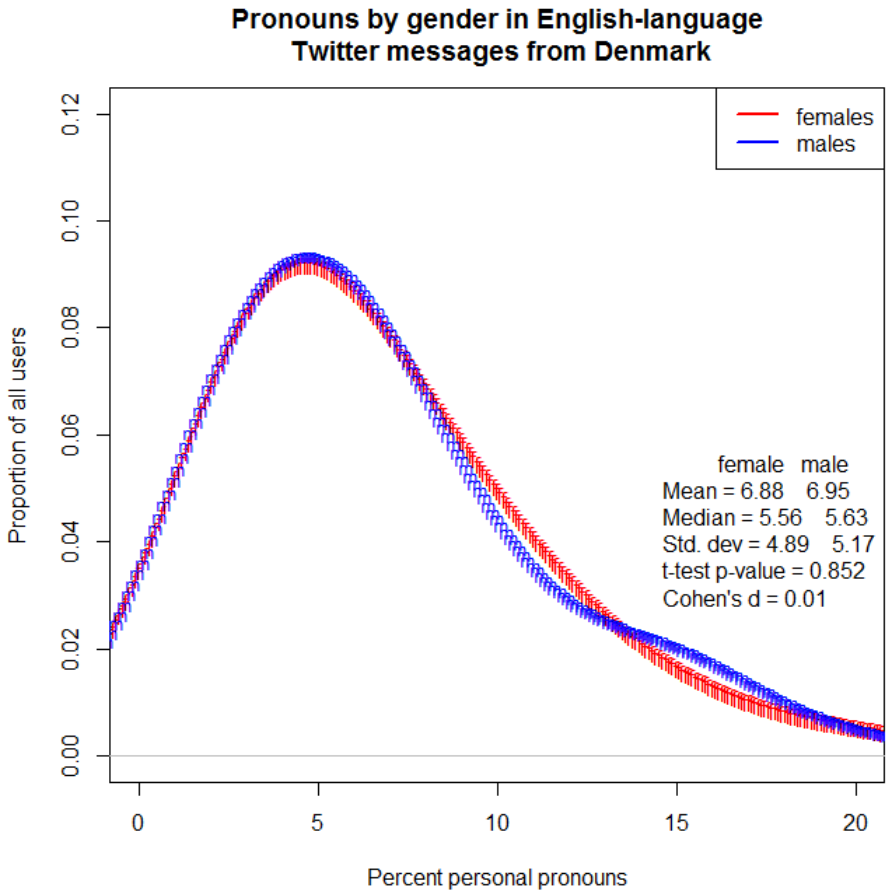
- Males in our subcorpora use more articles than do females
- The effect is significant at $p < 0.05$ for Norway and the US
- Except for Norway, the difference is not large (Cohen's *d* values)

Personal pronouns: Iceland

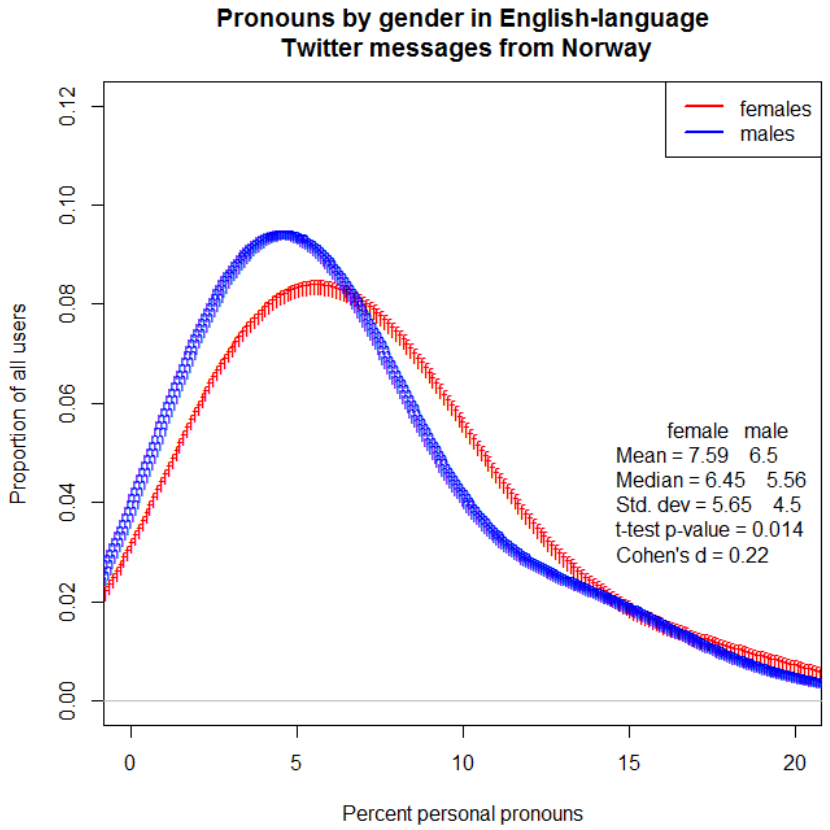
- Another feature that has been shown to differ according to author gender is use of personal pronouns in discourse



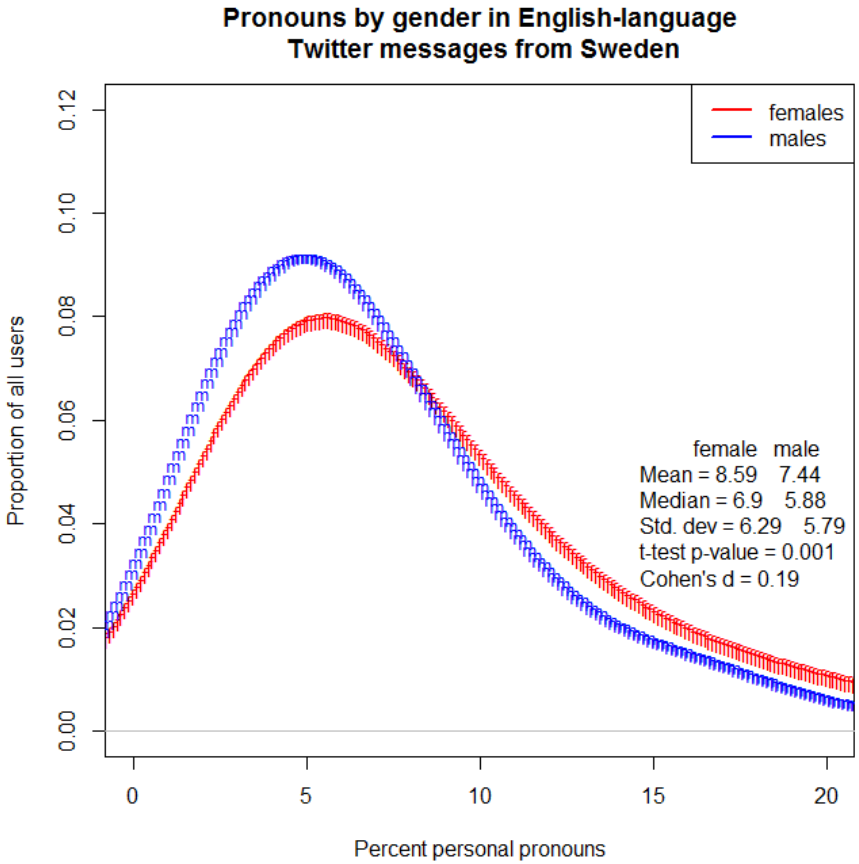
Personal pronouns: Denmark



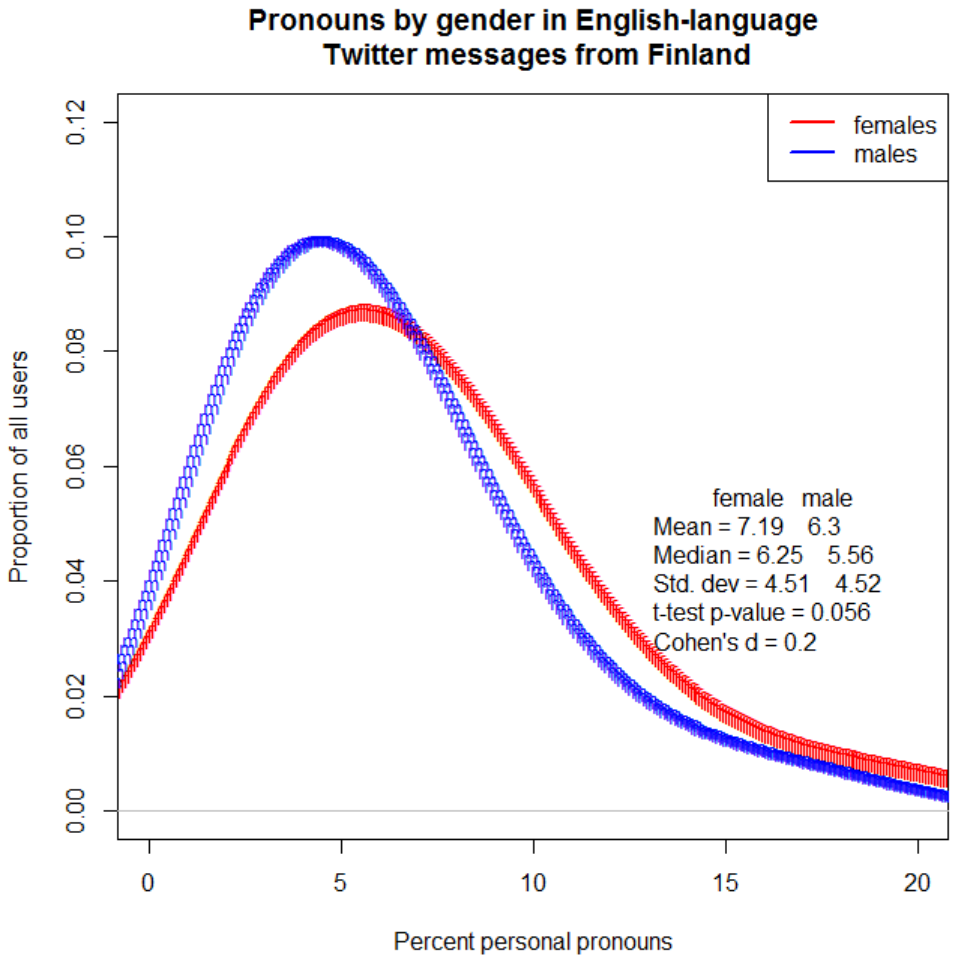
Personal pronouns: Norway



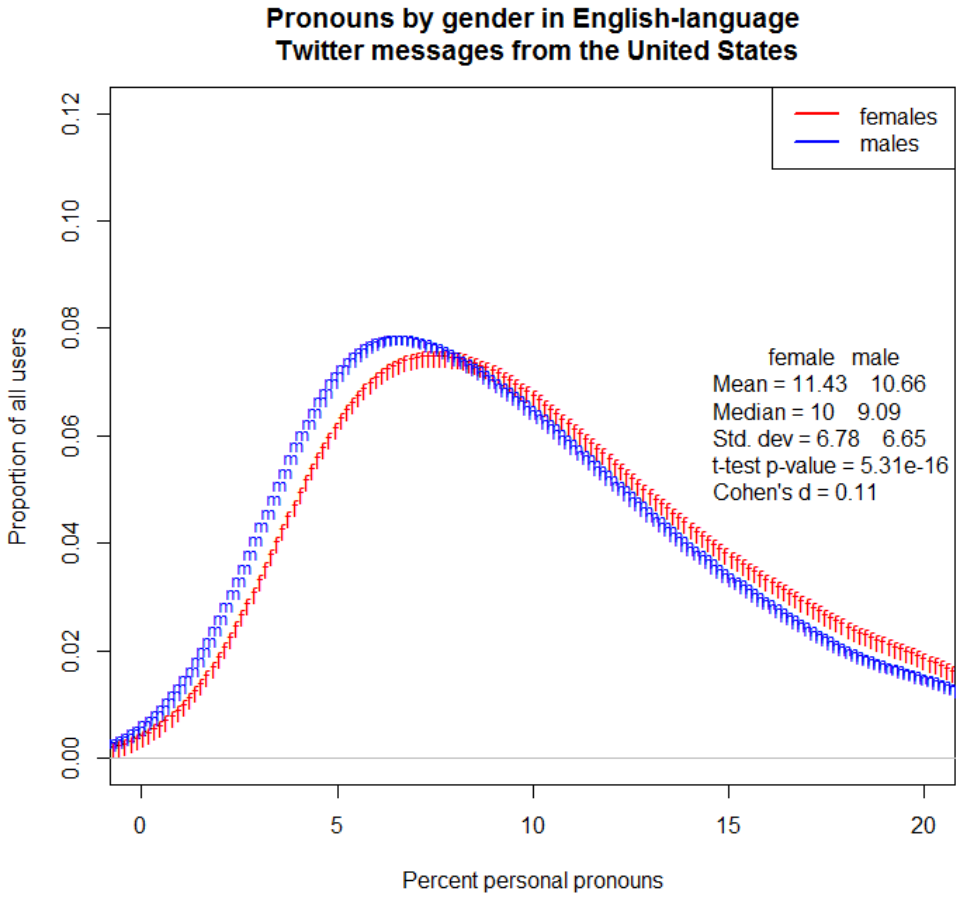
Personal pronouns: Sweden



Personal pronouns: Finland



Personal pronouns: US



Personal pronouns

- Females in our subcorpora use more personal pronouns than do males
- The effect is significant at $p < 0.05$ for Norway, Sweden, and the US
- The effect is not large (but larger than the difference in article use)

T-tests of significance on differences in use of 34 grammatical features

- When comparing the langauge of all female and all male users from the Nordic subcorpora, a few differences in feature use attain statistical significance (t-test of population means, $p < 0.05$)
- Males: sentence-ending punctuation, proper nouns
- Females: adjectives, modal verbs, personal pronouns, verbal infinitives, non 3rd-person present singular verb forms

Feature	Gender (or not significnat at p < 0.05)	Feature	Gender
''	ns	PRP	F
-RRB-	ns	PRP$	ns
,	ns	RB	ns
.	M	RBR	ns
:	ns	RP	ns
CC	ns	TO	ns
CD	ns	UH	ns
DT	ns	URL	ns
HT	ns	USR	ns
IN	ns	VB	F
JJ	F	VBD	ns
JJR	ns	VBG	ns
JJS	ns	VBN	ns
MD	F	VBP	F
NN	ns	VBZ	ns
NNP	M	WP	ns
NNS	ns	WRB	ns

Aggregate comparisons: distance measure

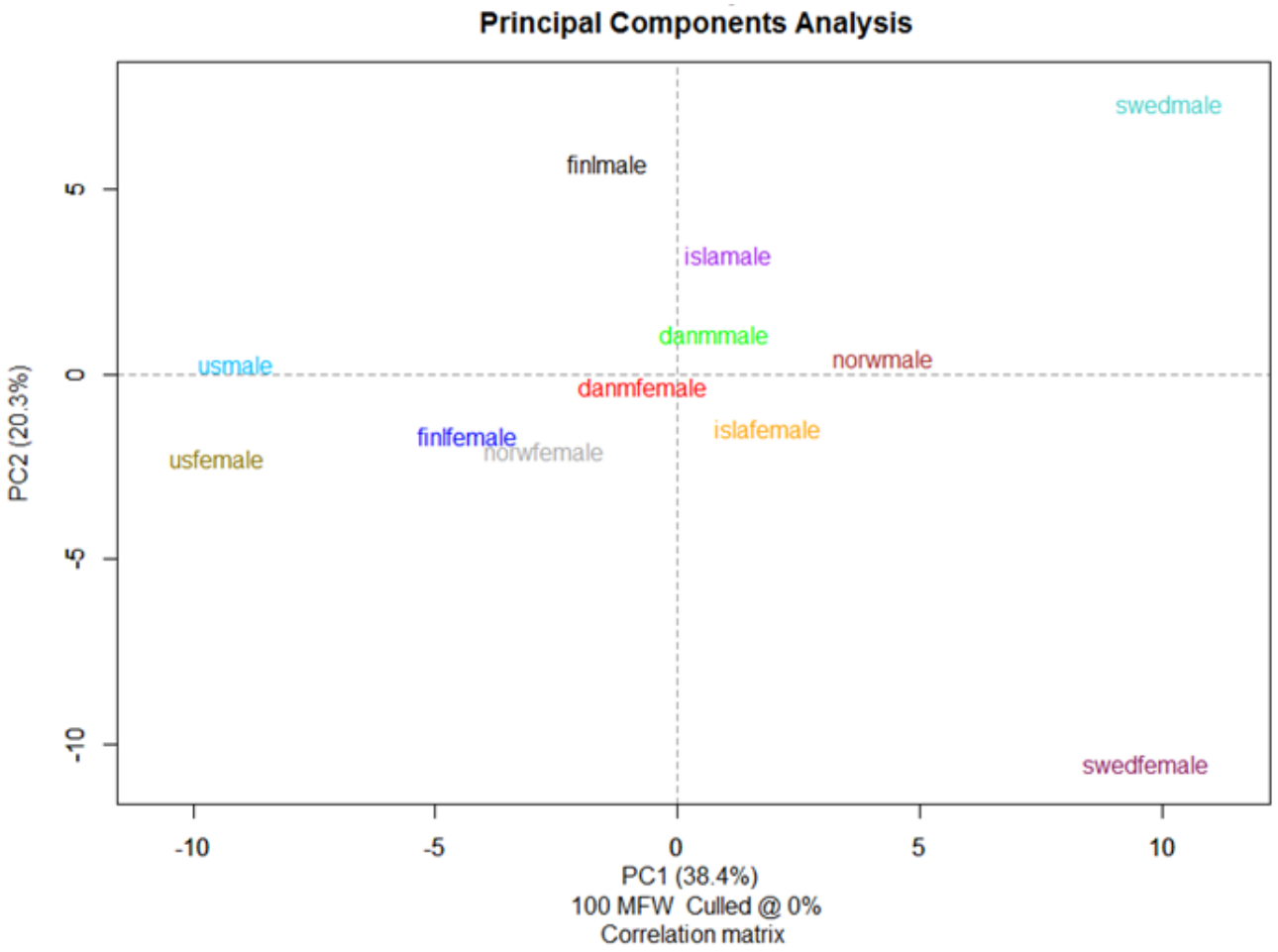
- The relationship between any two subcorpora is calculated with an aggregate distance measure
- A measure from computational stylometry is used: **Burrows' Delta**

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - B_i}{\sigma_i}\right|$$

- The mean of the absolute differences between the z-scores for a set of word variables or PoS variables in a given text A and the z-scores for the same set of word variables or PoS variables in a target set B (Burrows 2002: 271).
- Lexical comparison: calculate Burrows' Delta for the 100 most frequent word types in all of the subcorpora
- Grammatical comparison: calculate Burrows' Delta for the 34 PoS tags in all of the subcorpora
- The R package "stylo" is used (Eder, Kestemont and Rybicki 2013)

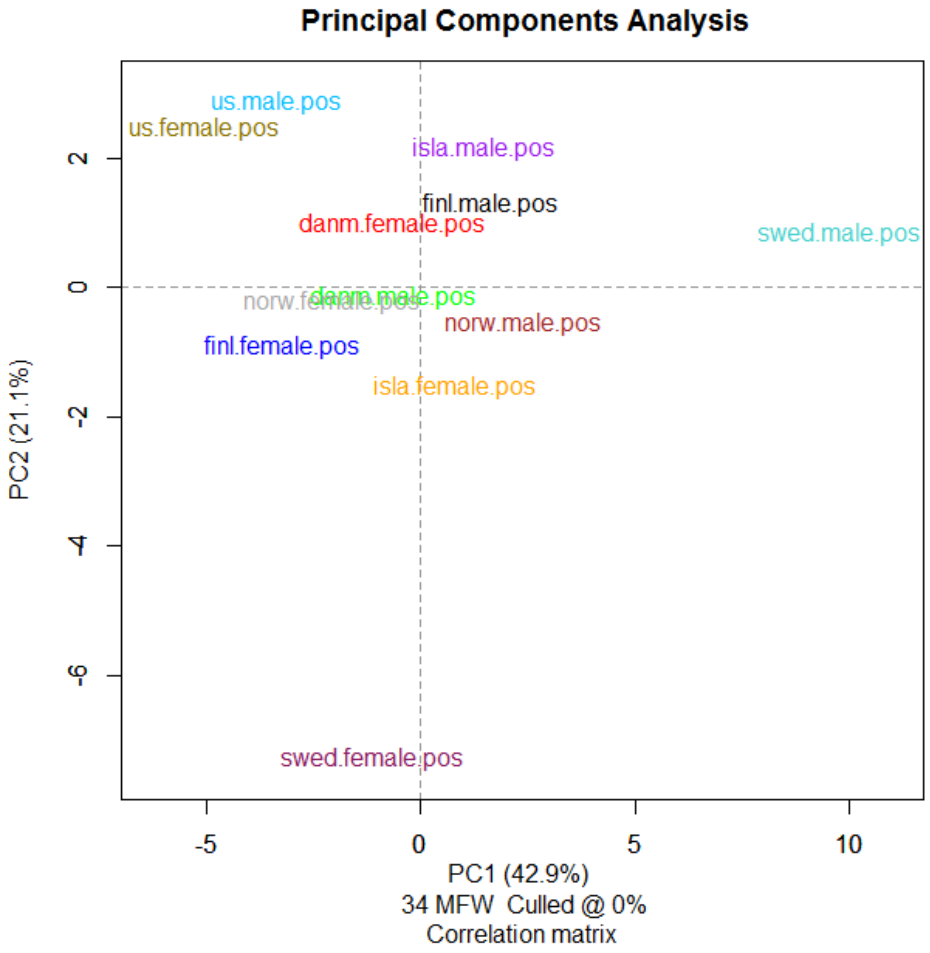
Lexical features: PCA

- Underlying patterns of variance in the data can be explored using techniques such as principal components analysis, agglomerative clustering, factor analysis, etc.
- Here Burrows' Delta values for each subcorpus have been converted to a correlaltion matrix and principal components identified; plotted are the loadings on PCs 1 and 2
- Analysis of the 100 most frequent word types suggests a slight functional separation between genders along a principal component
- "Male" corpora have positive values on PC2, "female" negative
- Nordic varieties are mostly closer to each other than to US English



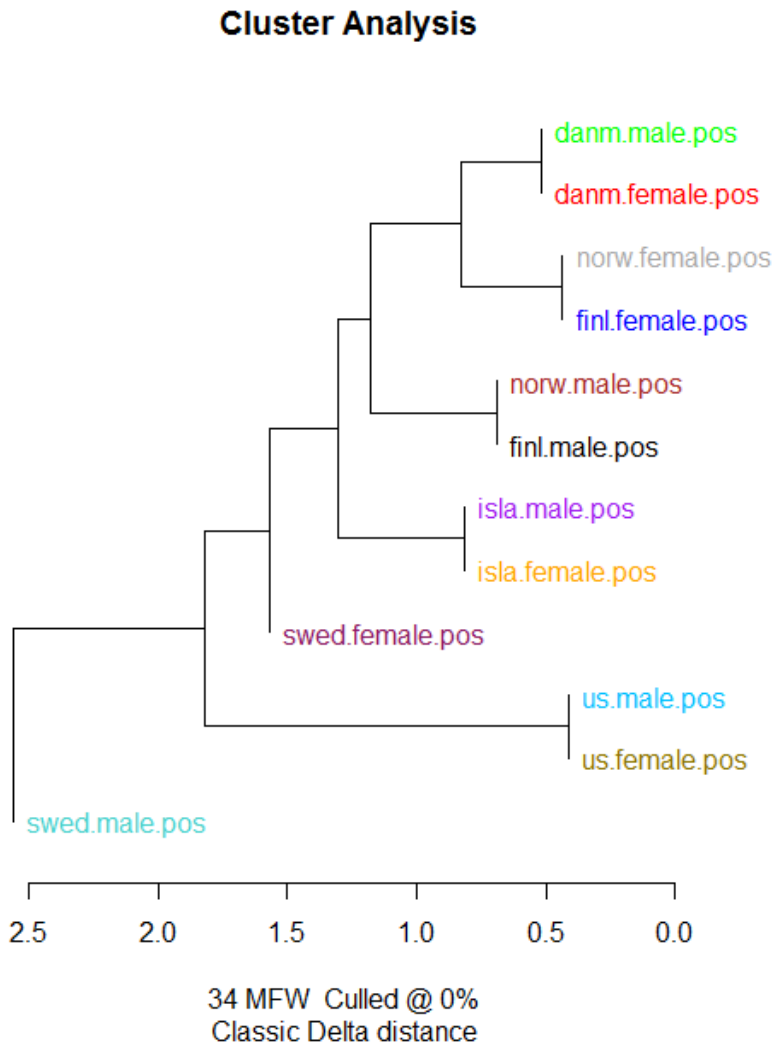
Grammatical features: PCA

- Analysis of grammatical features shows a similar configuration along the first two principal components
- No clear gender distinction



Grammatical features: clustering

- Cluster analysis of Burrows' Delta values for subcorpora based on aggregate frequencies of 34 grammatical features
- Males and females from US, Iceland, and Denmark cluster together
- Males and females for Sweden, Norway, and Finland do not



Summary and outlook

- Extensive use of English on Twitter in the Nordics (Denmark > Norway > Iceland > Sweden > Finland)
- Preliminary confirmation of functional gendered differences in the use of certain word classes for Nordic Twitter users writing in English
- Larger sample sizes are needed
- Gender differences are slight – an analysis from the perspective of topic-based social groups may be informative (Bamann, Eisenstein and Schnoebelen 2014)
- The process by which language features become associated with social categories in CMC (”enregisterment”, Squires 2010) deserves further scrutiny

References

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, UK: Cambridge University Press.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge, UK: Cambridge University Press.

Biber, D. 2006. *University Language. A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D. and Conrad, S. 2009. *Register, Genre and Style*. Cambridge, UK: Cambridge University Press.

Burrows, J. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17: 267–287.

Eder, M., Kestemont, M. and Rybicki, J. 2013. Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference* Abstracts, 487–89. Lincoln, NE: University of Nebraska-Lincoln. https://sites.google.com/site/computationalstylistics/stylo.

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

References

Herring, S. 2013. Discourse in Web 2.0: Familiar, reconfigured, and emergent" In Tannen, D. and Trester, A.M., (eds.), *Discourse 2.0: Language and New Media*, 1–25. Washington, DC: Georgetown University Press.

Hoover, D. 2004. Testing Burrows's Delta. *Literary and Linguistic Computing* 19, 453–475.

Kachru, B. 1990. *The Alchemy of English: The Spread, Functions, and Models of Nonnative Englishes*. Urbana, IL: University of Illinois Press.

Kachru, B. 1992. World Englishes: approaches, issues and resources. *Language Teaching* 25, 1–14. Cambridge, UK: Cambridge University Press.

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).

Mislove, A., Lehmann, A., Ahn, Y.-Y., Onnela, J.-P. and Rosenquist, J.N. 2011. Understanding the demographics of Twitter users. In *Proceedings of the International Conference on Web and Social Media (ICWSM),* 554–557. Menlo Park, CA: AAAI.

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).

Pavalanathan, U. and Eisenstein, J. 2015. Confounds and consequences in geotagged Twitter data. http://arxiv.org/abs/1506.02275

Pennebaker, J., Francis, M. and Booth, R. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahway, NJ: Lawrence Erlbaum Associates.

Roesslein, J. 2015. *Tweepy*. Python programming language module. URL: https://github.com/tweepy/tweepy

Squires, L. 2010. Enregistering internet language. *Language in Society* 39, 457–492.