Nordic Englishes on Twitter

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

- **Selecting only English-language messages**

---

Nordic Englishes on Twitter | Coats | 16.3.2016
## Corpus summary statistics

<table>
<thead>
<tr>
<th>Country</th>
<th>Tweets</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iceland</td>
<td>8390</td>
<td>101,342</td>
</tr>
<tr>
<td>Denmark</td>
<td>23,571</td>
<td>274,726</td>
</tr>
<tr>
<td>Norway</td>
<td>35,298</td>
<td>321,670</td>
</tr>
<tr>
<td>Sweden</td>
<td>108,677</td>
<td>829,474</td>
</tr>
<tr>
<td>Finland</td>
<td>23,673</td>
<td>242,618</td>
</tr>
<tr>
<td>US</td>
<td>296,954</td>
<td>3,270,871</td>
</tr>
</tbody>
</table>
### Percentage of tweets by language (comparison with Mocanu et al. 2013)

<table>
<thead>
<tr>
<th>Language</th>
<th>Iceland</th>
<th>Norway</th>
<th>Denmark</th>
<th>Sweden</th>
<th>Finland</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>45.0</td>
<td>34.1</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
</tr>
<tr>
<td>NO</td>
<td>48.8</td>
<td>37.1</td>
<td>45.0</td>
<td>33.5</td>
<td>No data</td>
</tr>
<tr>
<td>DK</td>
<td>45.0</td>
<td>33.5</td>
<td>No data</td>
<td>70.3</td>
<td>60.6</td>
</tr>
<tr>
<td>SV</td>
<td>70.3</td>
<td>60.6</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
</tr>
<tr>
<td>FI</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>56.4</td>
</tr>
<tr>
<td>RU</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>18.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Iceland</th>
<th>Norway</th>
<th>Denmark</th>
<th>Sweden</th>
<th>Finland</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>39.0</td>
<td>48.4</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
</tr>
<tr>
<td>EN</td>
<td>24.6</td>
<td>35.6</td>
<td>40.0</td>
<td>42.3</td>
<td>18.1</td>
</tr>
<tr>
<td>NO</td>
<td>6.2</td>
<td>2.2</td>
<td>1.0</td>
<td>1.1</td>
<td>23.2</td>
</tr>
<tr>
<td>FI</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>27.1</td>
</tr>
<tr>
<td>SV</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>69.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Iceland</th>
<th>Norway</th>
<th>Denmark</th>
<th>Sweden</th>
<th>Finland</th>
</tr>
</thead>
<tbody>
<tr>
<td>DK</td>
<td>18.0</td>
<td>19.5</td>
<td>6.2</td>
<td>2.2</td>
<td>1.0</td>
</tr>
<tr>
<td>NO</td>
<td>1.0</td>
<td>1.1</td>
<td>4.1</td>
<td>1.0</td>
<td>No data</td>
</tr>
<tr>
<td>SV</td>
<td>2.9</td>
<td>3.4</td>
<td>1.3</td>
<td>1.4</td>
<td>No data</td>
</tr>
<tr>
<td>RU</td>
<td>3.9</td>
<td>0.9</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
</tr>
</tbody>
</table>

- 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


- Tagger model applies Penn Treebank tags (Marcus et al. 1993) plus tags for CMC/Twitter-specific types (emoticons, usernames, retweets, URLs, hashtags)
### PoS tags applied by CMU Twitter Tagger

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

- **Mean** for females: 6.67, males: 7.18
- **Median** for females: 5.69, males: 6.25
- **Std. dev** for females: 3.87, males: 4.15
- **t-test p-value**: 0.057
- **Cohen's d**: 0.12
Articles: Norway
Articles: Sweden

Articles by gender in English-language
Twitter messages from Sweden

- Females
- Males

Mean: Females 6.77, Males 7.21
Median: Females 6.66, Males 6.25
Std. dev: Females 4.93, Males 4.35
T-test p-value: 0.075
Cohen's d: 0.1
Articles: Finland

Articles by gender in English-language Twitter messages from Finland

- Females
  - Mean: 6.25
  - Median: 5.52
  - Std. dev: 4.1
  - t-test p-value: 0.544
  - Cohen's d: 0.06

- Males
  - Mean: 6.48
  - Median: 5.66
  - Std. dev: 3.96
  - t-test p-value: 0.544
  - Cohen's d: 0.06
Articles: US

![Graph showing articles by gender in US English-language Twitter messages](image)

- **female**
  - Mean = 8.34
  - Median = 7.14
  - Std. dev = 5.01
  - t-test p-value = 0.026
  - Cohen's d = 0.03

- **male**
  - Mean = 8.52
  - Median = 7.32
  - Std. dev = 5.03
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

Pronouns by gender in English-language Twitter messages from Denmark

<table>
<thead>
<tr>
<th></th>
<th>females</th>
<th>males</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>6.88</td>
<td>6.95</td>
</tr>
<tr>
<td>median</td>
<td>5.56</td>
<td>5.63</td>
</tr>
<tr>
<td>std. dev</td>
<td>4.89</td>
<td>5.17</td>
</tr>
<tr>
<td>t-test p-value</td>
<td>0.852</td>
<td></td>
</tr>
<tr>
<td>Cohen's d</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>

Proportion of all users

Percent personal pronouns
Personal pronouns: Norway

Pronouns by gender in English-language Twitter messages from Norway

- Females
- Males

- Mean: 7.59
- Median: 6.45
- Std. dev: 5.65
- t-test p-value: 0.014
- Cohen's d: 0.22
Personal pronouns: Sweden

Pronouns by gender in English-language
Twitter messages from Sweden

- Females
  - Mean = 8.59
  - Median = 6.9
  - Std. dev = 6.29
  - t-test p-value = 0.001
  - Cohen's d = 0.19

- Males
  - Mean = 7.44
  - Median = 5.88
  - Std. dev = 5.79

Proportion of all users vs. Percent personal pronouns
Personal pronouns: Finland

Pronouns by gender in English-language Twitter messages from Finland

- **Mean**: Females: 7.19, Males: 6.3
- **Median**: Females: 6.25, Males: 5.56
- **Std. dev**: Females: 4.51, Males: 4.52
- **t-test p-value**: 0.056
- **Cohen’s d**: 0.2
Personal pronouns: US

Pronouns by gender in English-language Twitter messages from the United States

- Females: Mean = 11.43, Median = 10, Std. dev = 6.78, t-test p-value = 5.31e-16, Cohen's d = 0.11
- Males: Mean = 10.66, Median = 9.09, Std. dev = 6.85

Proportion of all users vs. Percent personal pronouns
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 language 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
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 correlation 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


References


