1 Contexts of the Present Research

2 Data Collection and Processing
   Data Collection
   Gender Disambiguation
   PoS Tagging

3 Analysis
   Language Profile
   Correlation of Grammatical Features and Gender
   Principal Components Analysis
   Feature Dispersion

4 Summary and Conclusion

Grammatical Frequencies and Gender in Nordic Twitter Englishes

Steven Coats

English Philology, University of Oulu, Finland

27 September 2016
# Table of Contents

1. **Contexts of the Present Research**

2. Data Collection and Processing
   - Data Collection
   - Gender Disambiguation
   - PoS Tagging

3. Analysis
   - Language Profile
   - Correlation of Grammatical Features and Gender
   - Principal Components Analysis
   - Feature Dispersion

4. Summary and Conclusion
1 Contexts of the Present Research

- Categorization of discourse, language genres or varieties based on the principal communicative functions exemplified by configurations of linguistic features (Biber 1988, 1995, 2006; Biber and Conrad 2009)
- English as it is used in the Nordic countries: CMC “Global Englishes” and the status of Nordic languages
- Gender and language: Do differences reported for L1 English in CMC (e.g. Wolf 2000, Baron 2004, Herring and Paolillo 2006, Herring 2013, Bamman et al. 2014) also hold true for English used in the Nordics?
# Table of Contents

1 Contexts of
the Present
Research

2 Data
Collection and
Processing
- Data Collection
- Gender Disambiguation
- PoS Tagging

3 Analysis
- Language Profile
- Correlation of Grammatical Features and Gender
- Principal Components Analysis
- Feature Dispersion

4 Summary
and Conclusion

## Contexts of the Present Research

## Data Collection and Processing

### Data Collection

### Gender Disambiguation

### PoS Tagging

## Analysis

### Language Profile

### Correlation of Grammatical Features and Gender

### Principal Components Analysis

### Feature Dispersion

## Summary and Conclusion
Data Collection

- Nordic data: Python script to get tweets from Streaming API
- Geo-encoded tweets from a bounding box circumscribing the Nordic countries (longitude -26 to 32, latitude 53 to 72)
- Data collected May 2016: 16.2 m tweets
- Filtering by **country** field: 302,737 tweets from Nordic countries of Iceland, Norway, Denmark, Sweden, and Finland
- Further filtering by **language** field: 101,956 tweets in English (1,475,553 tokens)
Gender Disambiguation

- author_name field for each user filtered for strings that either begin with or include as a discrete element the most common male and female given names by country
- Data: Name frequency information from national statistical offices
- Tweets from Sweden by users with the (invented) author_name values:
  - Anna Andersson → Sweden female subcorpus
  - سُعاد الأطرش → ignored
  - zYlax85 → ignored
- Users matching both male and female names for a country → ignored
- Method assigned gender to 34% of Iceland, 49% of Norway, 62% of Denmark, 48% of Sweden, and 60% of Finland tweets
- 10 gendered subcorpora created (m and f for Iceland, Norway, Denmark, Sweden, and Finland)
## Tokenization and PoS Tagging

- **Carnegie-Mellon Twitter PoS Tagger** (Gimpel et al. 2011; Gimpel et al. 2013, Owoputi et al. 2013)

- **Penn Treebank tags** (Marcus et al. 1993) tags, additional tags for Twitter-specific types (retweet, username, hashtag), emoticons get “interjection” tag

- Output consists of tab-separated token/tag/probability lines

### Example Tweet

<table>
<thead>
<tr>
<th>Token</th>
<th>Tag</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovenia</td>
<td>NNP</td>
<td>0.9735</td>
</tr>
<tr>
<td>,</td>
<td>,</td>
<td>0.9905</td>
</tr>
<tr>
<td>it's</td>
<td>PRP</td>
<td>0.8554</td>
</tr>
<tr>
<td>very</td>
<td>RB</td>
<td>0.9529</td>
</tr>
<tr>
<td>unimaginative</td>
<td>JJ</td>
<td>0.9728</td>
</tr>
<tr>
<td>but</td>
<td>CC</td>
<td>0.9903</td>
</tr>
<tr>
<td>I</td>
<td>PRP</td>
<td>0.9940</td>
</tr>
<tr>
<td>like</td>
<td>VBP</td>
<td>0.7748</td>
</tr>
<tr>
<td>it</td>
<td>PRP</td>
<td>0.9969</td>
</tr>
<tr>
<td>:D</td>
<td>UH</td>
<td>0.9909</td>
</tr>
<tr>
<td>#BlueIsBlue</td>
<td>HT</td>
<td>0.9774</td>
</tr>
<tr>
<td>#RedIsRed</td>
<td>HT</td>
<td>0.9786</td>
</tr>
<tr>
<td>#Eurovision</td>
<td>HT</td>
<td>0.9851</td>
</tr>
<tr>
<td>#ESC2016</td>
<td>HT</td>
<td>0.9375</td>
</tr>
<tr>
<td>#ComeTogether</td>
<td>HT</td>
<td>0.9823</td>
</tr>
</tbody>
</table>

A tweet from Sweden about the Eurovision song contest.
Table of Contents

1 Contexts of the Present Research

2 Data Collection and Processing
   Data Collection
   Gender Disambiguation
   PoS Tagging

3 Analysis
   Language Profile
   Correlation of Grammatical Features and Gender
   Principal Components Analysis
   Feature Dispersion

4 Summary and Conclusion
English is used extensively in the Nordics on Twitter (Denmark > Norway > Finland > Sweden > Iceland)

<table>
<thead>
<tr>
<th>Country</th>
<th>National language(s)</th>
<th>English</th>
<th>Other languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iceland</td>
<td>7.76</td>
<td>3.13</td>
<td>0.10</td>
</tr>
<tr>
<td>Norway</td>
<td>3.43</td>
<td>7.32</td>
<td>9.23</td>
</tr>
<tr>
<td>Denmark</td>
<td>7.36</td>
<td>6.45</td>
<td>8.17</td>
</tr>
<tr>
<td>Sweden</td>
<td>6.62</td>
<td>2.24</td>
<td>2.13</td>
</tr>
<tr>
<td>Finland</td>
<td>3.56</td>
<td>3.27</td>
<td>4.16</td>
</tr>
</tbody>
</table>

Percent tweets by country and language
1 Contexts of the Present Research

2 Data Collection and Processing
   Data Collection
   Gender Disambiguation
   PoS Tagging

3 Analysis
   Language Profile
   Correlation of Grammatical Features and Gender
   Principal Components Analysis
   Feature Dispersion

4 Summary and Conclusion

Language Profile

- Females use English more than males on Twitter in Iceland, Norway and Denmark, while males use the national languages slightly more in Sweden and Finland.

- The differences in English use by gender were statistically significant at $p < 0.05$ for all 5 countries (Fisher’s exact test).

- Effect sizes (odds ratio $\theta$):
  Iceland $= 1.96$, Norway $= 1.64$, Denmark $= 1.66$, Sweden $= 0.96$, Finland $= 0.82$.

<table>
<thead>
<tr>
<th>Country</th>
<th>Language Type</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iceland</td>
<td>National language</td>
<td>5.71</td>
<td>8.80</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>6.17</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>Other language</td>
<td>9.10</td>
<td>4.9</td>
</tr>
<tr>
<td>Norway</td>
<td>National languages</td>
<td>3.37</td>
<td>6.46</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>0.40</td>
<td>9.28</td>
</tr>
<tr>
<td></td>
<td>Other language</td>
<td>7.22</td>
<td>5.24</td>
</tr>
<tr>
<td>Denmark</td>
<td>National language</td>
<td>7.25</td>
<td>4.45</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>5.52</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Other language</td>
<td>8.21</td>
<td>6.14</td>
</tr>
<tr>
<td>Sweden</td>
<td>National language</td>
<td>8.63</td>
<td>9.61</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>8.23</td>
<td>5.24</td>
</tr>
<tr>
<td></td>
<td>Other language</td>
<td>4.12</td>
<td>6.13</td>
</tr>
<tr>
<td>Finland</td>
<td>National languages</td>
<td>5.58</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>0.25</td>
<td>8.28</td>
</tr>
<tr>
<td></td>
<td>Other language</td>
<td>5.16</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Percent tweets by country, language, and gender.
Many tweets from Denmark and Norway are in English - from rural Sweden, Finland or Iceland less so (Iceland values averaged across all provinces)
For the entire English corpus, ten features exhibited significant differences in frequency of use by gender according to the results of Welch's two-sample t-test.

- Males used more sentence-ending punctuation (.), numbers (CD), proper nouns (NNP), and past participles (VBN).
- Females used more personal pronouns (PRP), possessive pronouns (PRPS), adverbs (RB), interjections (UH), verbal base forms (VB), and non-3rd-person-present singular verb forms (VBP).
Correlation of Grammatical Features and Gender by Country

- Feature frequencies by country and gender
- For example: Sweden; significant difference in use by gender for nine features
- Modest effect sizes (Cohen’s D) – largest here is personal pronoun use
Principal Components Analysis

PCA of Gendered Subcorpora, Components 1 and 2

PC1, Proportion of Variance = 58.92%
PC2, Proportion of Variance = 11.91%

PCA of Gendered Subcorpora, Components 1 and 2

- is.m
- is.f
- no.m
- no.f
- da.m
- da.f
- sv.m
- sv.f
- fi.m
- fi.f

1 Contexts of the Present Research
2 Data Collection and Processing
- Data Collection
- Gender Disambiguation
- PoS Tagging
3 Analysis
- Language Profile
- Correlation of Grammatical Features and Gender
4 Summary and Conclusion
Feature Dispersion

1 Contexts of the Present Research

2 Data Collection and Processing
- Data Collection
- Gender Disambiguation
- PoS Tagging

3 Analysis
- Language Profile
- Correlation of Grammatical Features and Gender
- Principal Components Analysis

4 Summary and Conclusion
1 Contexts of the Present Research

2 Data Collection and Processing
  Data Collection
  Gender Disambiguation
  PoS Tagging

3 Analysis
  Language Profile
  Correlation of Grammatical Features and Gender
  Principal Components Analysis
  Feature Dispersion

4 Summary and Conclusion
1 Contexts of the Present Research

2 Data Collection and Processing
   - Data Collection
   - Gender Disambiguation
   - PoS Tagging

3 Analysis
   - Language Profile
   - Correlation of Grammatical Features and Gender
   - Principal Components Analysis

4 Summary and Conclusion

Feature Dispersion

![Feature Dispersion Diagram]

- Proper noun
- Frequency per 1000 tokens
- PC1

Languages:
- is.m
- is.f
- no.m
- no.f
- da.m
- da.f
- sv.m
- sv.f
- fi.m
- fi.f

Values:
-0.04
-0.02
0.00
0.02
0.04
70 80 90 100 110
1 Contexts of the Present Research

2 Data Collection and Processing
   - Data Collection
   - Gender Disambiguation
   - PoS Tagging

3 Analysis
   - Language Profile
   - Correlation of Grammatical Features and Gender
   - Principal Components Analysis

4 Summary and Conclusion

Feature Dispersion

- Frequency per 1000 tokens
- Personal pronoun

Graph showing the distribution of personal pronouns across different PC1 values.
1 Contexts of the Present Research

2 Data Collection and Processing
- Data Collection
- Gender Disambiguation
- PoS Tagging

3 Analysis
- Language Profile
- Correlation of Grammatical Features and Gender
- Principal Components Analysis
- Feature Dispersion

4 Summary and Conclusion

Feature Dispersion

Frequency per 1000 tokens
Possessive pronoun

PC1
1 Contexts of the Present Research

2 Data Collection and Processing
Data Collection
Gender Disambiguation
PoS Tagging

3 Analysis
Language Profile
Correlation of Grammatical Features and Gender
Principal Components Analysis
Feature Dispersion

4 Summary and Conclusion

Feature Dispersion

Adverb
PC1

Frequency per 1000 tokens

is.m
is.f
no.m
no.f
da.m
da.f
sv.m
sv.f
fi.m
fi.f
−0.04
−0.02
0.00
0.02
0.04
35 40 45 50
1 Contexts of the Present Research

2 Data Collection and Processing
- Data Collection
- Gender Disambiguation
- PoS Tagging

3 Analysis
- Language Profile
- Correlation of Grammatical Features and Gender
- Principal Components Analysis
- Feature Dispersion

4 Summary and Conclusion

Feature Dispersion

![Feature Dispersion Graph](image)

PC1

Frequency per 1000 tokens

Interjection

- is.m
- is.f
- no.m
- no.f
- da.m
- da.f
- sv.m
- sv.f
- fi.m
- fi.f

-0.04
-0.02
0.00
0.02
0.04

30 40 50
1 Contexts of the Present Research

2 Data Collection and Processing
   Data Collection
   Gender Disambiguation
   PoS Tagging

3 Analysis
   Language Profile
   Correlation of Grammatical Features and Gender
   Principal Components Analysis

Feature Dispersion

PC1

Frequency per 1000 tokens
Verb, base form

Verb, base form

is.m
is.f
no.m
no.f
da.m
da.f
sv.m
sv.f
fi.m
fi.f
−0.04
−0.02
0.00
0.02
0.04
35 40 45

Feature Dispersion
Feature Dispersion

Frequency per 1000 tokens
Verb, past participle

PC1

-0.04 -0.02 0.00 0.02 0.04
5.0 5.5 6.0 6.5 7.0 7.5

Contexts of the Present Research

Data Collection and Processing
Data Collection
Gender Disambiguation
PoS Tagging

3 Analysis
Language Profile
Correlation of Grammatical Features and Gender
Principal Components Analysis

Feature Dispersion

16.9.27 Coats
Grammar and Gender Nordic Twitter English
Feature Dispersion

- Frequency per 1000 tokens
- Verb, non—3rd person singular present

PC1

- 20 25 30 35
- -0.04 -0.02 0.00 0.02 0.04

Contexts of the Present Research

Data Collection and Processing
- Data Collection
- Gender Disambiguation
- PoS Tagging

Analysis
- Language Profile
- Correlation of Grammatical Features and Gender
- Principal Components Analysis

Summary and Conclusion
Table of Contents

1 Contexts of the Present Research

2 Data Collection and Processing
   Data Collection
   Gender Disambiguation
   PoS Tagging

3 Analysis
   Language Profile
   Correlation of Grammatical Features and Gender
   Principal Components Analysis
   Feature Dispersion

4 Summary and Conclusion

Summary and Conclusion
Summary – Language Preference

- English is extensively used on Twitter in the Nordic countries (Denmark > Norway > Finland > Sweden > Iceland)
- Overall, females make slightly more use of English than males, and males more use of the national languages. At the country level, this is true for Iceland, Norway, and Denmark.
- Sociolinguistic interpretation: Females are more ready to embrace new language practices when they are introduced into a community from social groups perceived to have high social standing? (parallel to observed change in feature use by gender, Labov 2001: 266)
- For Sweden and Finland, the pattern is reversed (but effect sizes are also much smaller) – this supports findings from Finnish survey data (Leppanen et al. 2011)
Summary – Feature Preference

- Females use features associated with interaction and the negotiation of **stance and affect concerns**
- Males use features associated with **information or discourse organization**
- Similar to findings from L1 English contexts for various genres (Argamon et al. 2007; Bamman et al. 2014)
- Different preferences in communicative style? **Involved versus informational** dimensions suggested by Biber (1988; 1995)
- Principal components analysis of 34 features in the ten gendered Nordic subcorpora: First principal component, which explains 9%.58 of total variation, **discriminates author gender**
Summary – Conclusion

- Confirmation of findings from L1 anglophone cultures in L2 contexts → communicative styles between males and females
- Future work: Larger corpora, feature frequency analysis in Nordic (or other L1) languages by gender
- Thanks for your attention!
References I


References II


References III

1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection
Gender Disambiguation
PoS Tagging

3 Analysis
Language Profile
Correlation of Grammatical Features and Gender
Principal Components Analysis
Feature Dispersion

4 Summary and Conclusion


