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

UNIVERSITY OF OULU

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

UNIVERSITY OF OULU

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

- 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?

UNIVERSITY OF OULU

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

UNIVERSITY OF OULU

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

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

```
83008   Slovenia        NNP 0,9735
83009   ,       ,       0,9905
83010   it's    PRP 0,8554
83011   very    RB  0,9529
83012   unimaginative   JJ  0,9728
83013   ,       ,       0,9903
83014   but CC  0,9990
83015   I   PRP 0,9940
83016   like    VBP 0,7748
83017   it  PRP 0,9969
83018   :D  UH  0,9909
83019   #BlueIsBlue HT  0,9774
83020   #RedIsRed   HT  0,9786
83021   #Eurovision HT  0,9851
83022   #ESC2016    HT  0,9375
83023   #ComeTogether   HT  0,9823
```

A tweet from Sweden about the Eurovision song contest

UNIVERSITY OF OULU

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

УЛЕАБОРГСКИЙ УНИВЕРСИТЕТ

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

English is used extensively in the Nordics on Twitter (Denmark > Norway > Finland > Sweden > Iceland

	National language(s)	English	Other languages
Iceland	7.76	3.13	0.10
Norway	3.43	7.32	9.23
Denmark	7.36	6.45	8.17
Sweden	6.62	2.24	2.13
Finland	3.56	3.27	4.16

Percent tweets by country and language

УЛЕАБОРГСКИЙ УНИВЕРСИТЕТ

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 θ): Iceland $= 1.96$, Norway $= 1.64$, Denmark $= 1.66$, Sweden $= 0.96$, Finland $= 0.82$

		female	male
Iceland	National language	5.71	8.80
	English	6.17	8.9
	Other language	9.10	4.9
Norway	National languages	3.37	6.46
	English	0.40	9.28
	Other language	7.22	5.24
Denmark	National language	7.25	4.45
	English	5.52	0.40
	Other language	8.21	6.14
Sweden	National language	8.63	9.61
	English	8.23	5.24
	Other language	4.12	6.13
Finland	National languages	5.58	2.57
	English	0.25	8.28
	Other language	5.16	0.14

Percent tweets by country, language, and gender

Geographical distribution of English tweets (gender-induced data)

1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection

Gender Disambiguation

PoS Tagging

3 Analysis

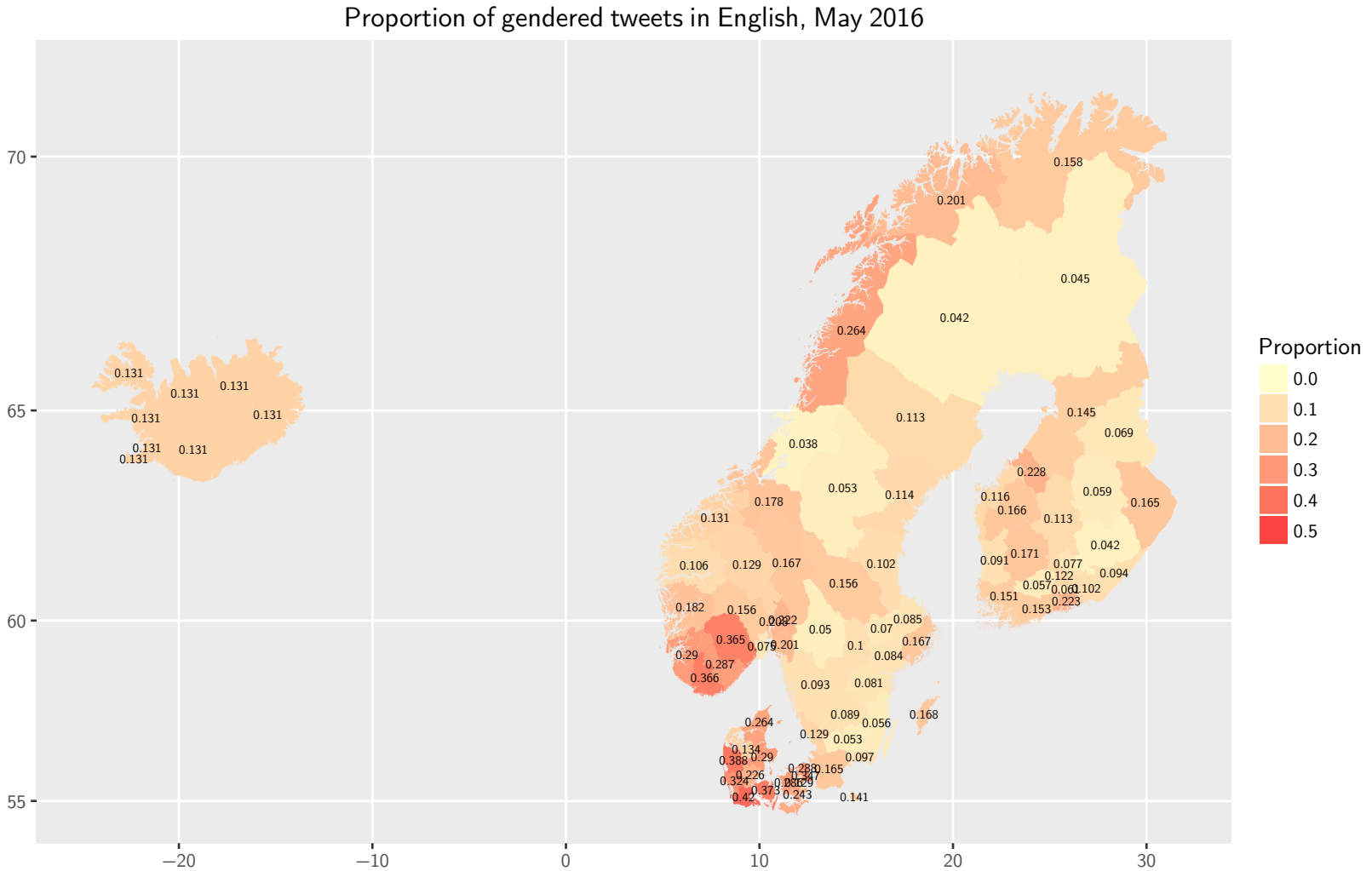
Language Profile

Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

4 Summary and Conclusion



- Many tweets from Denmark and Norway are in English - from rural Sweden, Finland or Iceland less so (Iceland values averaged across all provinces)

Correlation of Grammatical Features and Gender

1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection

Gender Disambiguation

PoS Tagging

3 Analysis

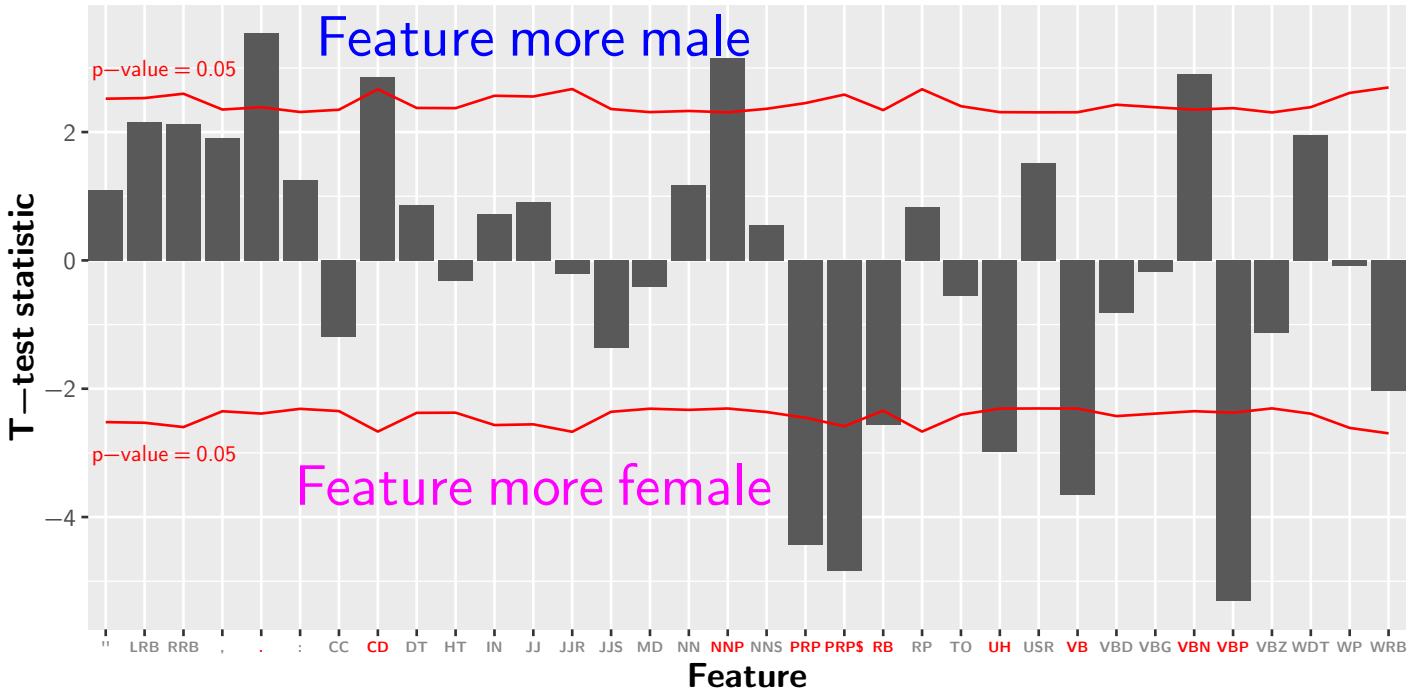
Language Profile

Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

4 Summary and Conclusion



- 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 (PRP$), 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

1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection

Gender Disambiguation

PoS Tagging

3 Analysis

Language Profile

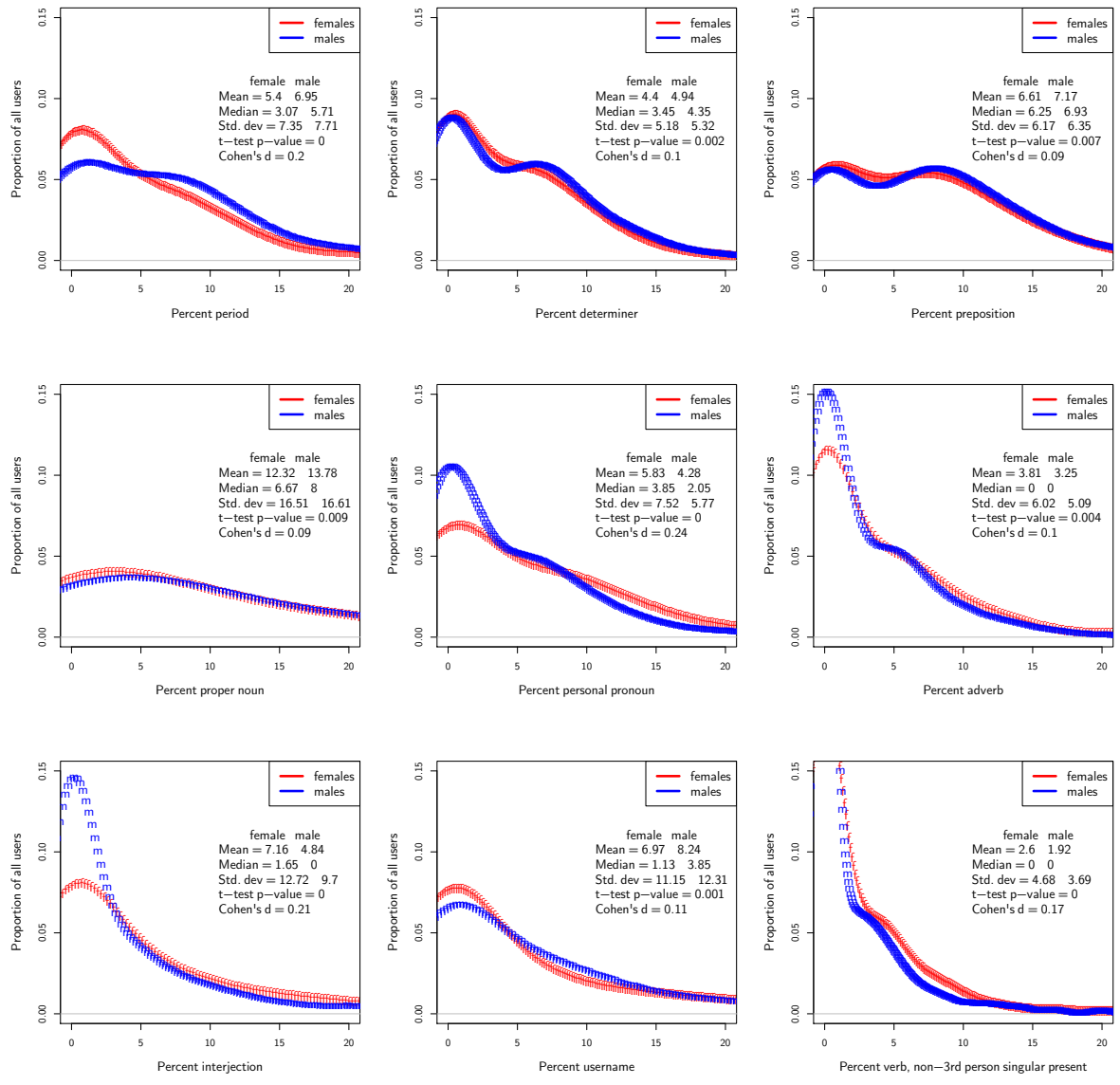
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

4 Summary and Conclusion

- Feature 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

1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection

Gender Disambiguation

PoS Tagging

3 Analysis

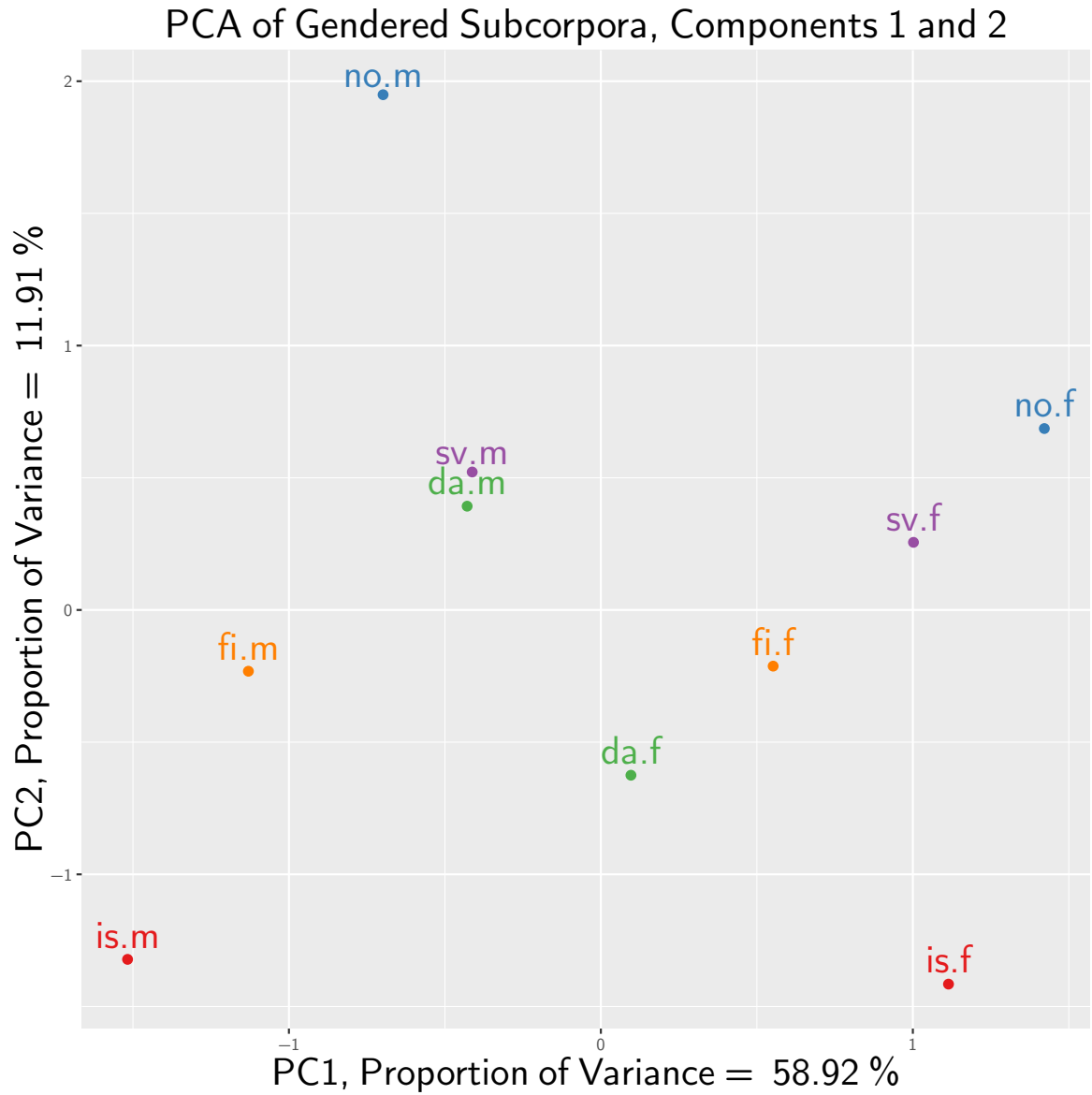
Language Profile

Correlation of Grammatical Features and Gender

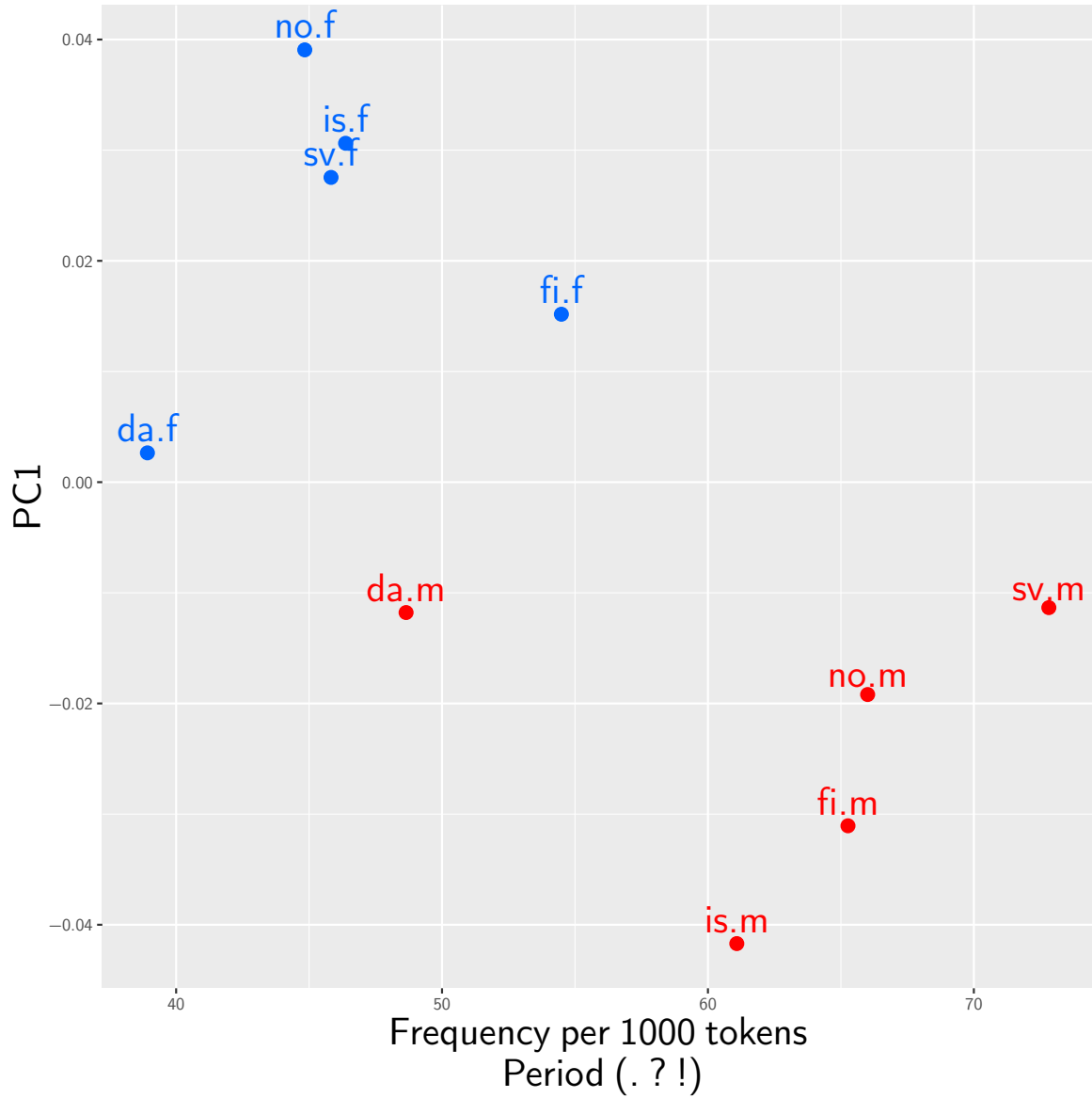
Principal Components Analysis

Feature Dispersion

4 Summary and Conclusion



Feature Dispersion



1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection

Gender Disambiguation

PoS Tagging

3 Analysis

Language Profile

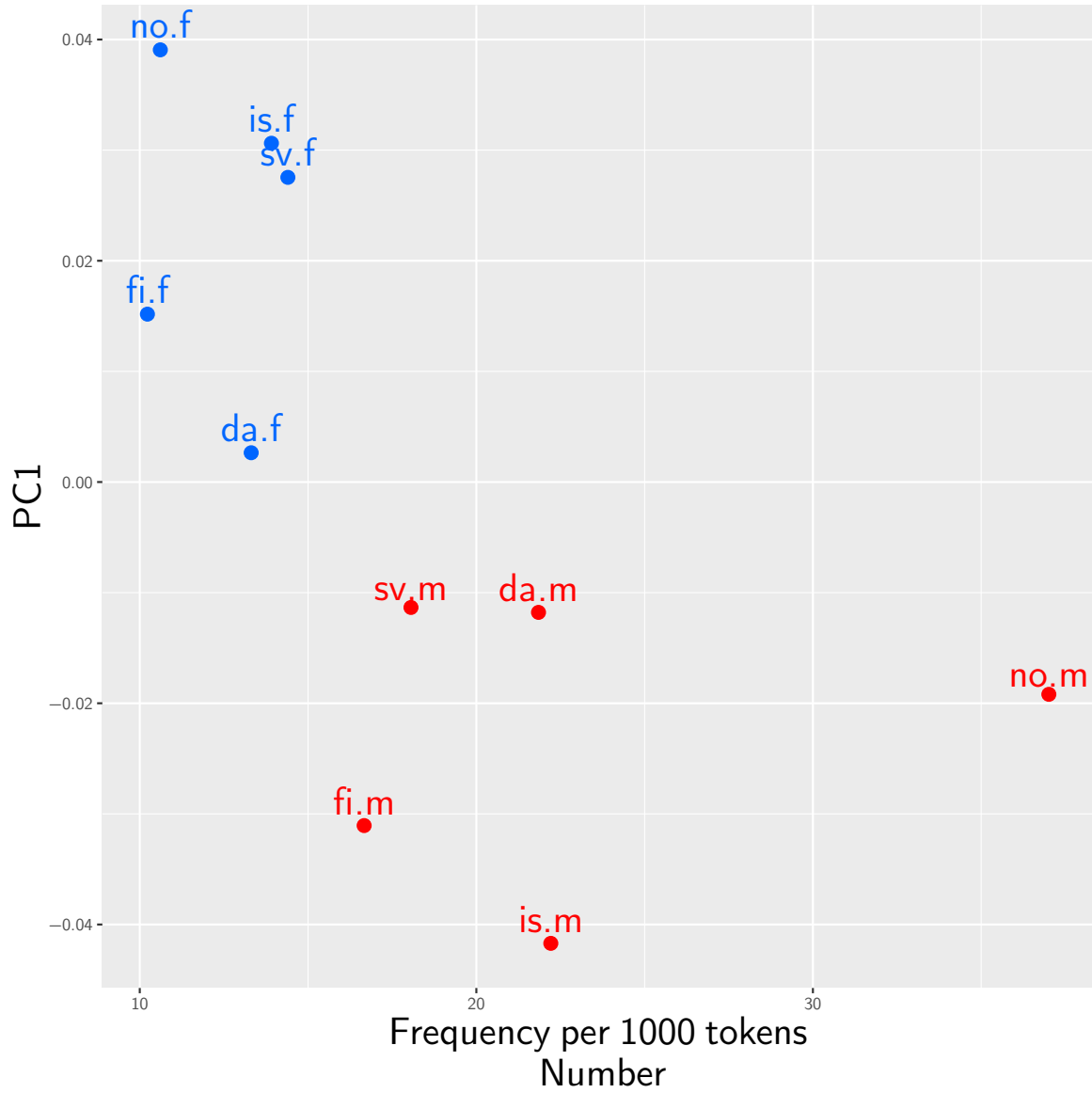
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

4 Summary and Conclusion

Feature Dispersion



Feature Dispersion

1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection

Gender Disambiguation

PoS Tagging

3 Analysis

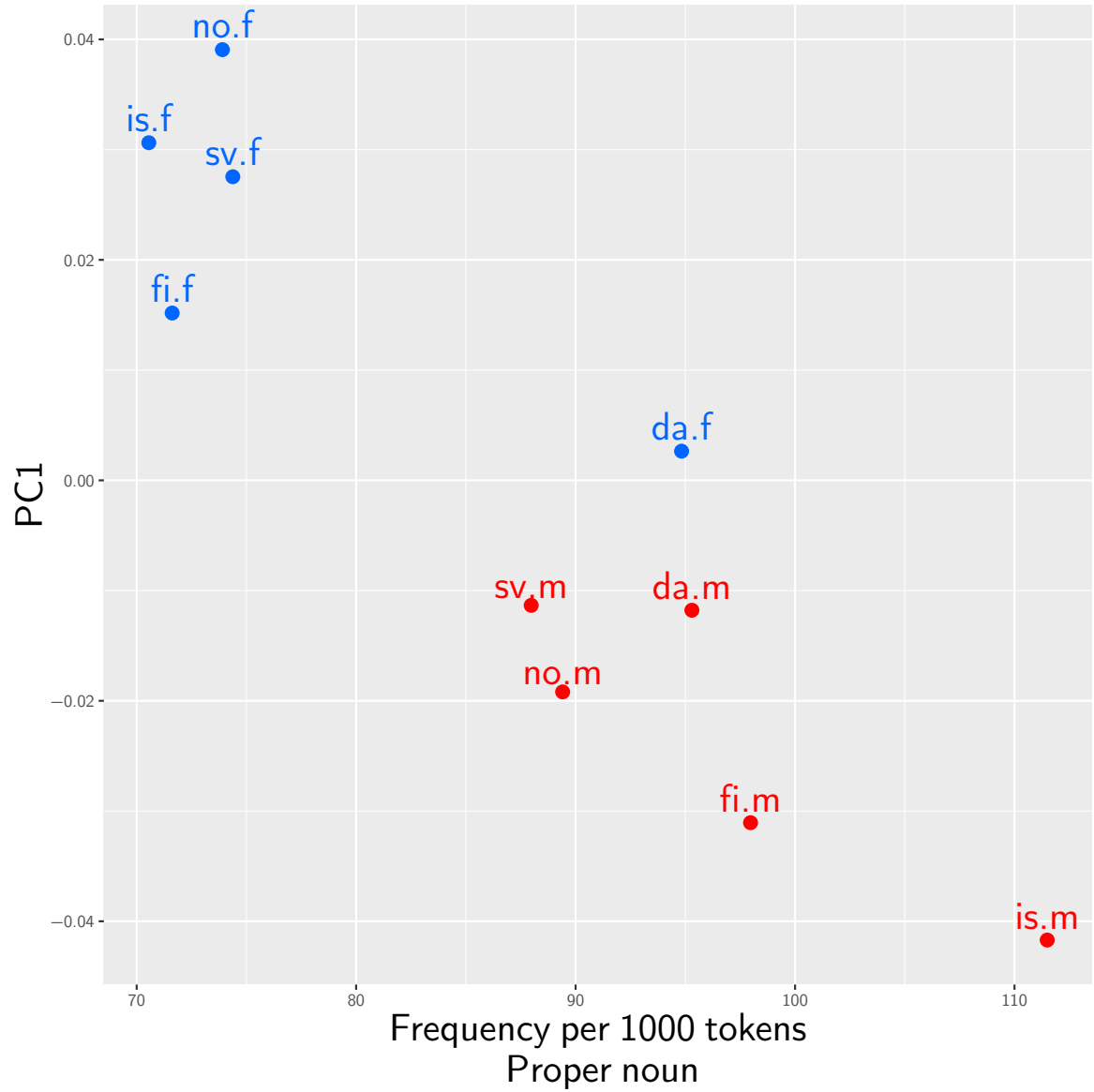
Language Profile

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

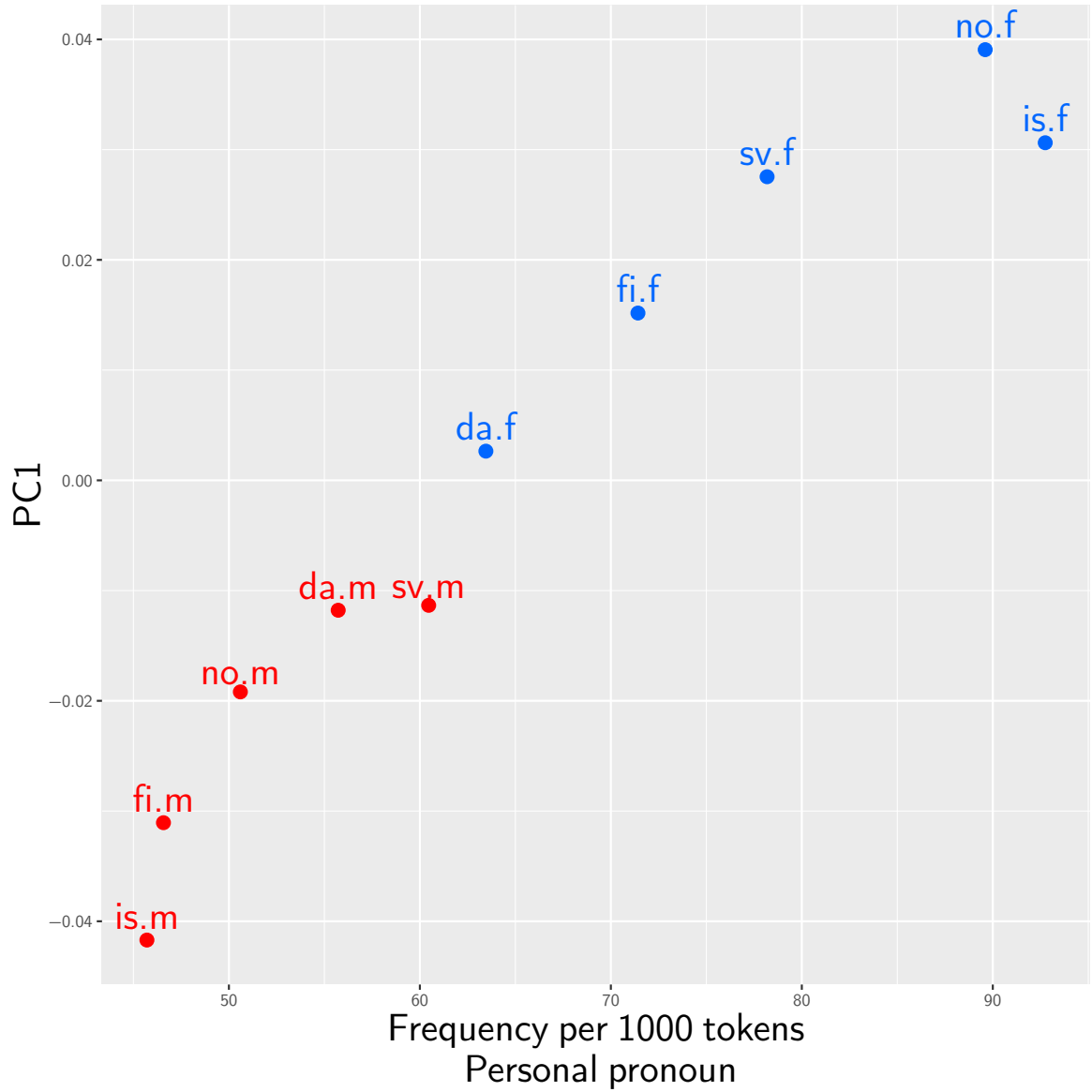
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

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

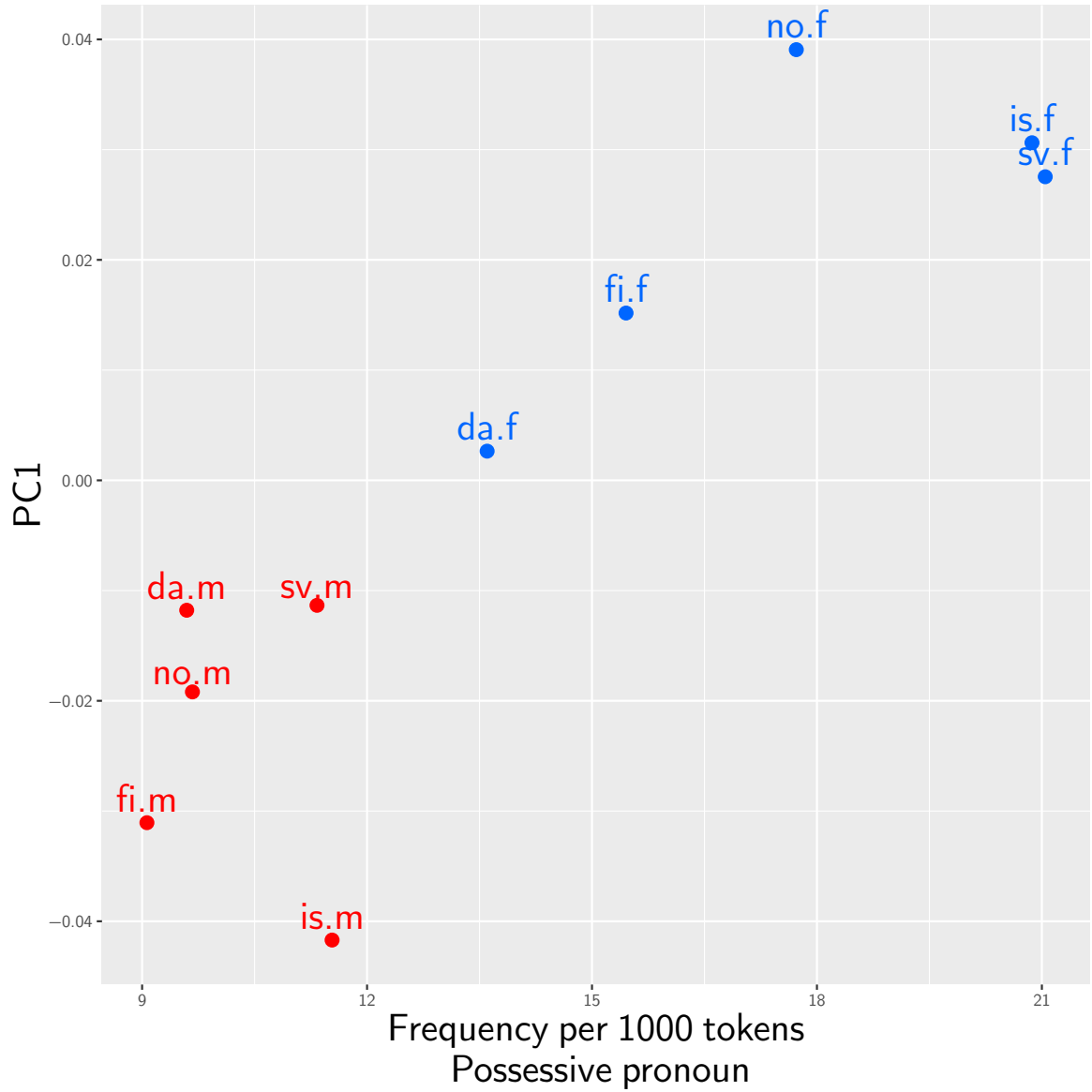
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

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

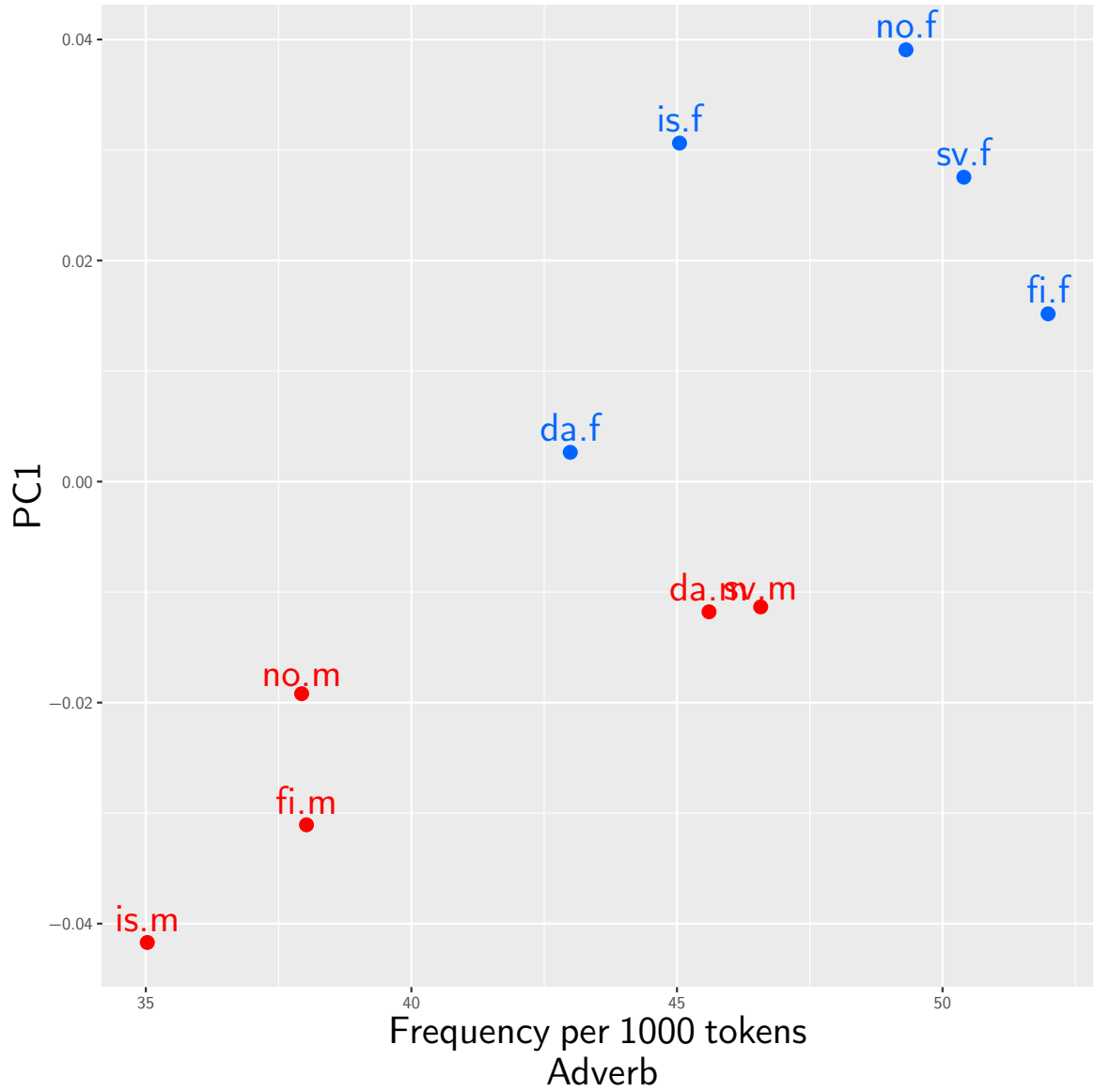
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

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

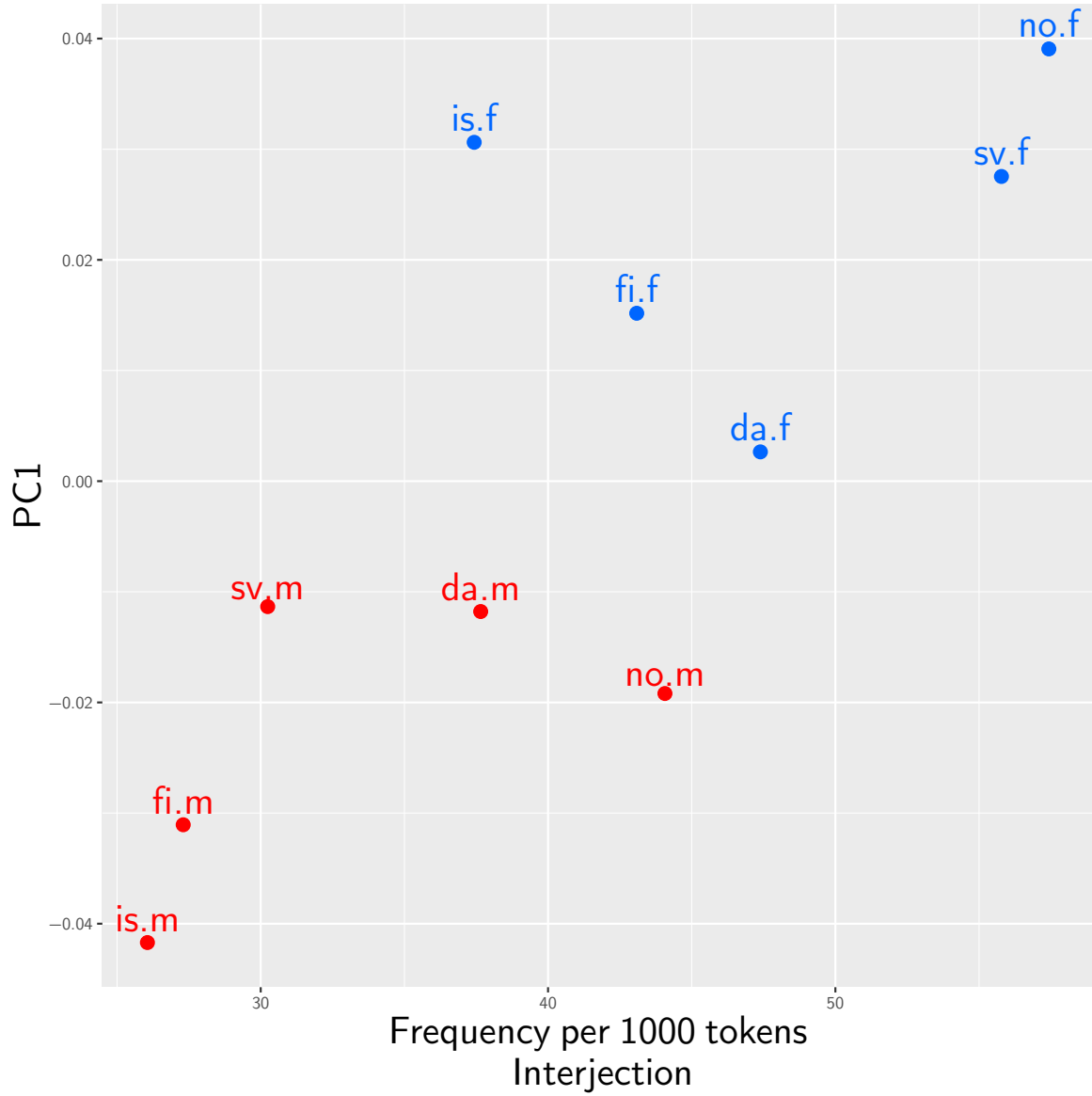
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

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

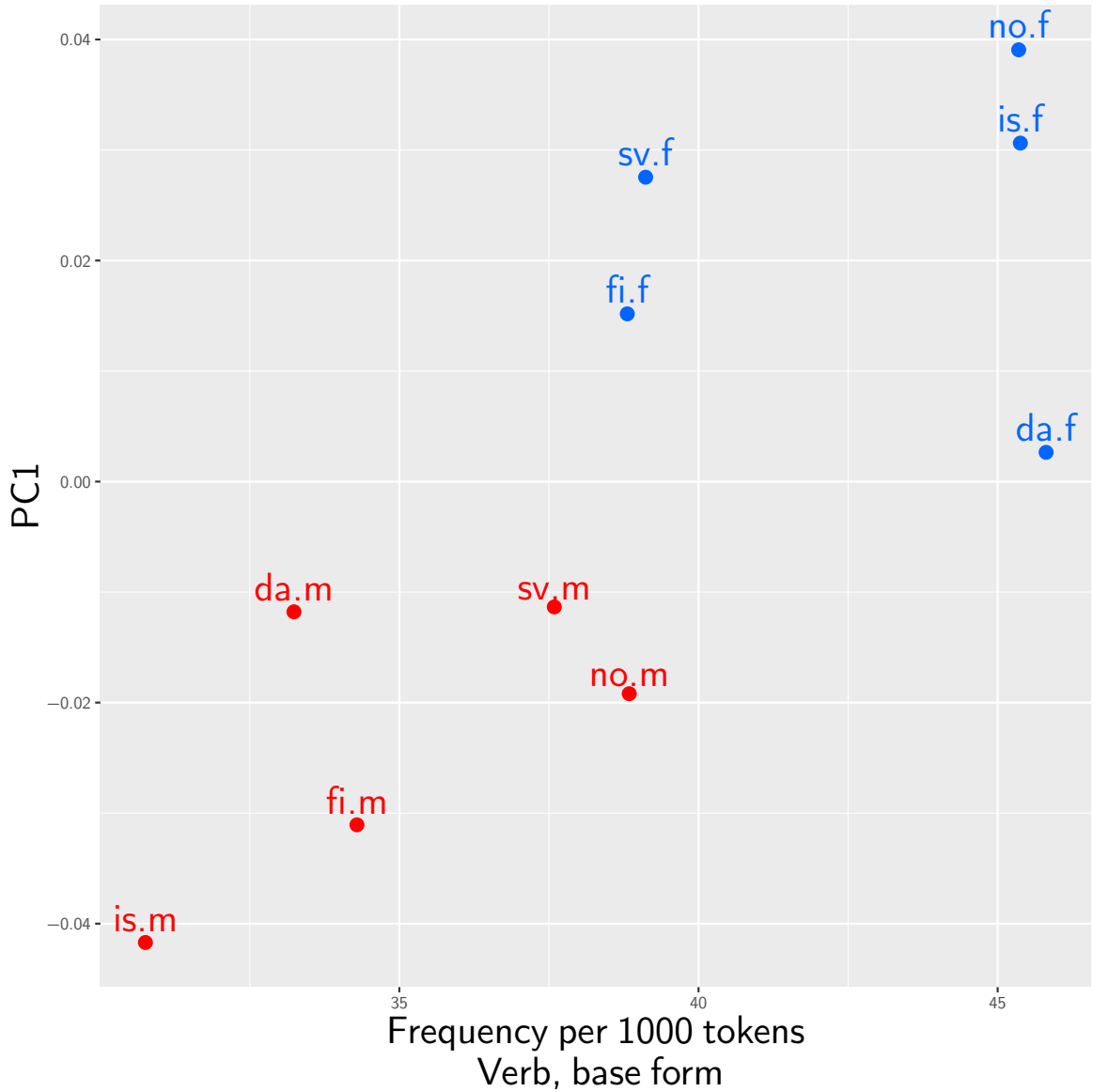
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

4 Summary and Conclusion

Feature Dispersion



Feature Dispersion

1 Contexts of the Present Research

2 Data Collection and Processing

Data Collection

Gender Disambiguation

PoS Tagging

3 Analysis

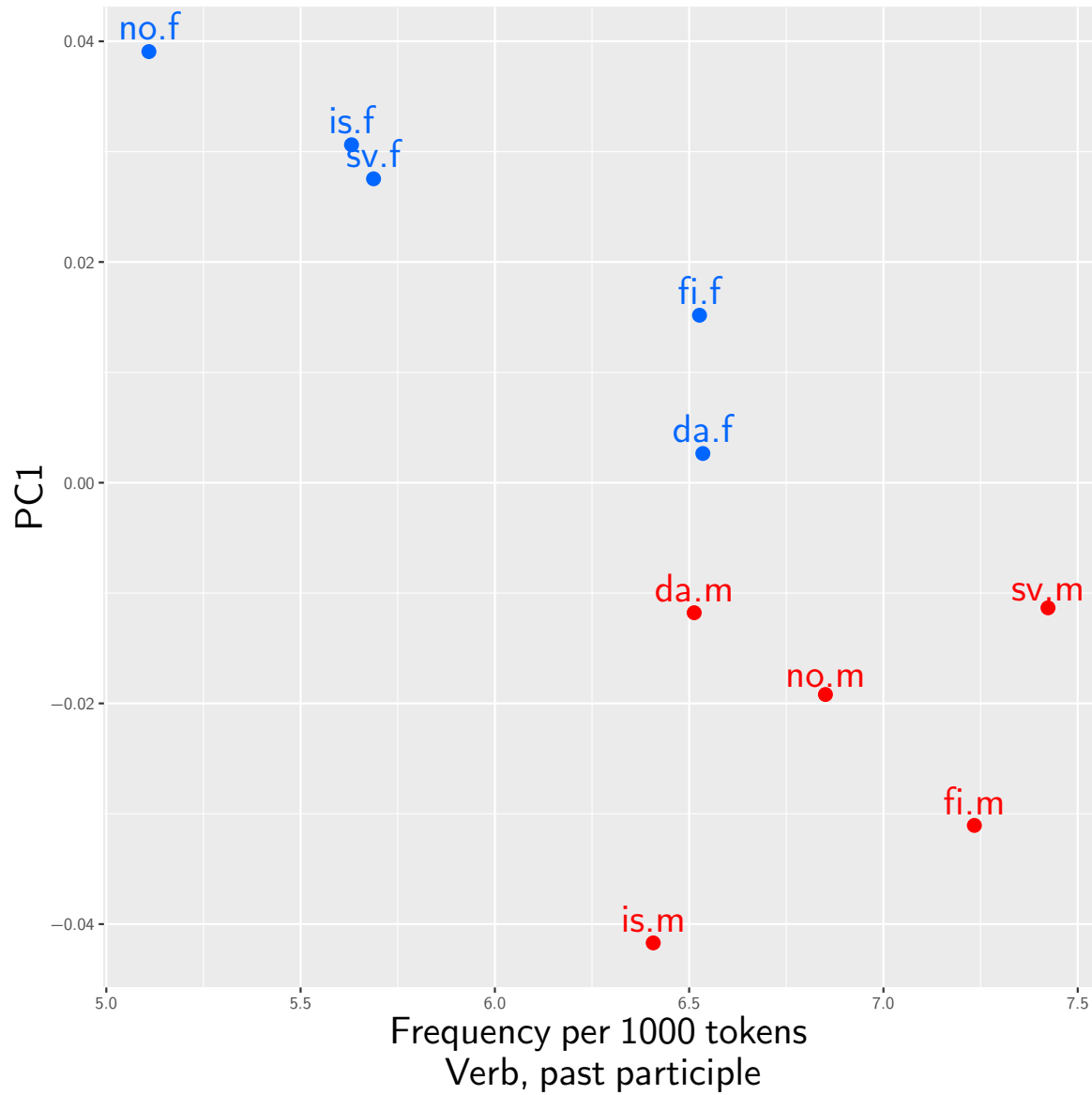
Language Profile

Correlation of Grammatical Features and Gender

Principal Components Analysis

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

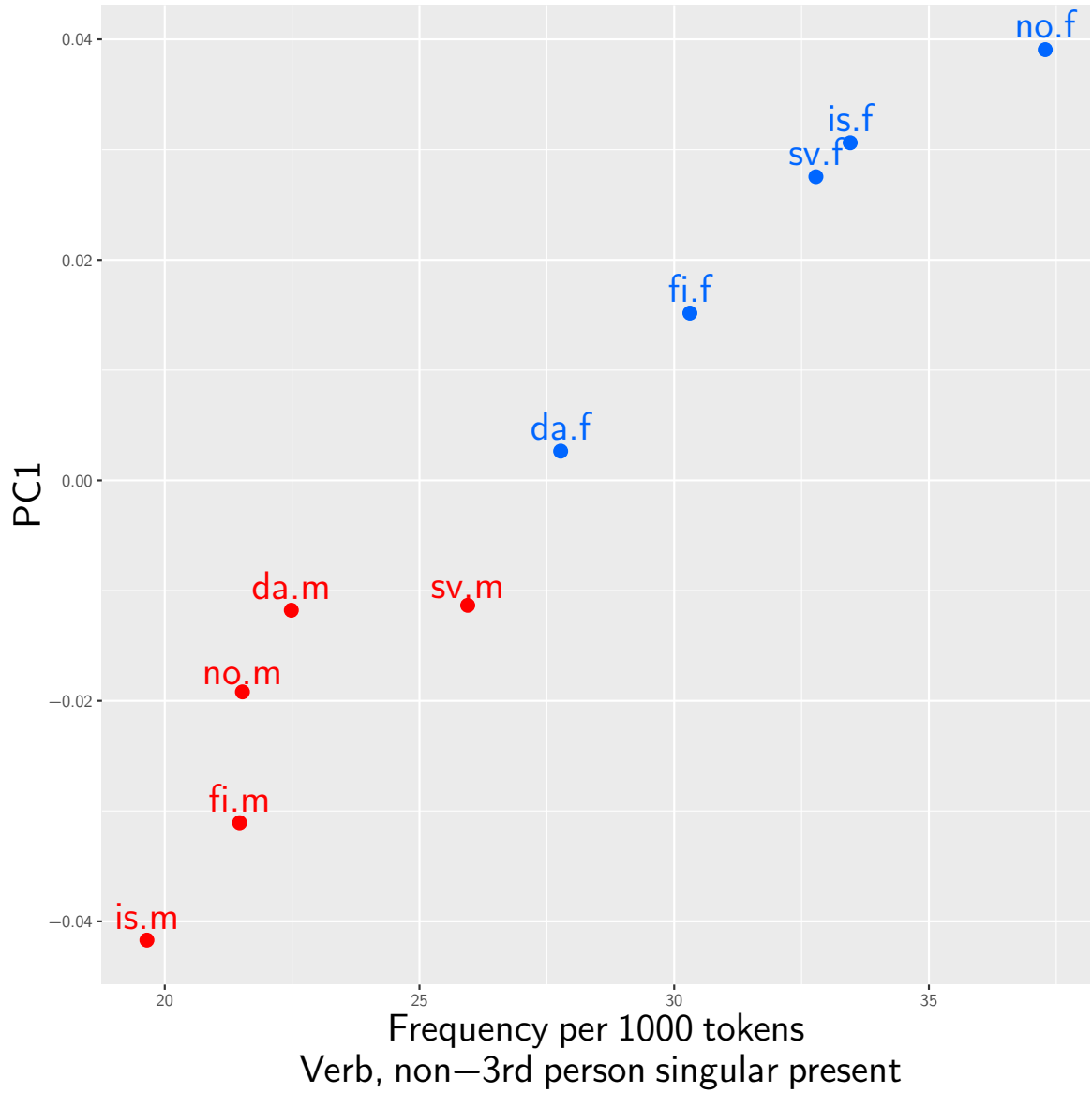
Correlation of Grammatical Features and Gender

Principal Components Analysis

Feature Dispersion

4 Summary and Conclusion

Feature Dispersion



UNIVERSITY OF OULU

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

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 – 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

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

- 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**

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 – Conclusion

- Confimation 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!

UNIVERSITY OF OULU

References I

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

Bamman, D., J. Eisenstein and T. Schnoebelen. (2014). "Gender Identity and Lexical Variation in Social Media". *Journal of Sociolinguistics* 18(2), .160–135

Baron, N. (2004). "See you online: Gender issues in college student Use of instant messaging". *Journal of Language and Social Psychology* 23(4), .423–397

Gimpel, K., N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan and N. A. Smith. (2011). "Part-of-speech tagging for Twitter: Annotation, features, and experiments". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, .47–42 Stroudsburg, PA: ACM. www.ark.cs.cmu.edu/TweetNLP/gimpel+etal.acl11.pdf. (accessed 16 September 2016)

Gimpel, K. N. Schneider and B. O'Connor. (2013). "Annotation Guidelines for Twitter Part-of-Speech Tagging Version "3.0 Computational Science Department, Carnegie Mellon University. http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf. (accessed 16 September 2016)

References II

Herring, S. (2013). “Discourse in Web :0.2 Familiar, reconfigured, and emergent”. In Deborah Tannen & Anna Marie Trester (eds.), *Discourse :0.2 Language and New Media*, .25–1 Washington, DC: Georgetown University Press.

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

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

Mocanu, D., A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani (2013). “The Twitter of Babel: Mapping world languages through microblogging platforms”. *PLoS ONE 4.8*.

References III


Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, N. Schneider and N. A. Smith. (2013). "Improved part-of-speech tagging for online conversational text with word clusters". *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, .390–380 Stroudsburg, PA: ACM. http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.naacl13.pdf. (accessed 16 September 2016)


Wolf, A. (2000). "Emotional expression online: Gender differences in emoticon use". *Cyber Psychology and Behavior* 3, .833–827


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 S. Conrad (2009). *Register, Genre and Style*. Cambridge: Cambridge University Press.

References IV


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


Labov, W. (2001). *Principles of Linguistic Change, Vol. 2: Social Factors*. Oxford: Blackwell.


Leppänen, S., A. Pitkänen-Huhta, T. Nikula, S. Kytölä, T. Törmäkangas, K. Nissinen, L. Kääntä, T. Räisänen, M. Laitinen, H. Koskela, S. Lähdesmäki and H. Jousmäki. (2011). *National Survey on the English Language in Finland: Uses, meanings and attitudes* (= Studies in Variation, Contacts and Change in English, Volume 5). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/05/evarieng-vol5.pdf