

# Multilingual Clusters and Gender in Nordic Twitter

Steven Coats

English Philology, Faculty of Humanities

[steven.coats@oulu.fi](mailto:steven.coats@oulu.fi) (<mailto:steven.coats@oulu.fi>)

CLARIN-PLUS Workshop “Creation and Use of Social Media Resources”

May 19th, 2017

# Goals of research

- Language use, bi- and multilingualism and gender: Small but significant differences in grammatical feature use according to gender have been found in large-scale corpus studies, including in online and social media (e.g. Argamon et al. 2007, Bamann et al. 2014, Coats 2016)
- Are analogous gender-based differences evident at a higher level of discourse organization (language choice, bi- and multilingualism)?
  1. Explore extent of language use by gender on Twitter for the Nordic countries
  2. Investigate similarities or differences by gender among bi- and multilingual Nordic Twitter users

# Data collection | Streaming API and gender disambiguation

- Tweets with populated place attributes collected using *Tweepy* (Roesslein 2015) from the Twitter Streaming API 9 November 2016 – 19 April 2017 (~530m tweets)
- Usernames of 162,035 unique authors of tweets with Nordic country\_code fields (Greenland, Iceland, Faroe Islands, Norway, Denmark, Sweden, Åland, Finland) identified
- Name frequency information for 15,284 names given to males and 17,763 names given to females obtained from Nordic statistical offices and used to induce a probability distribution for name gender

# Data collection | REST API and source filtering

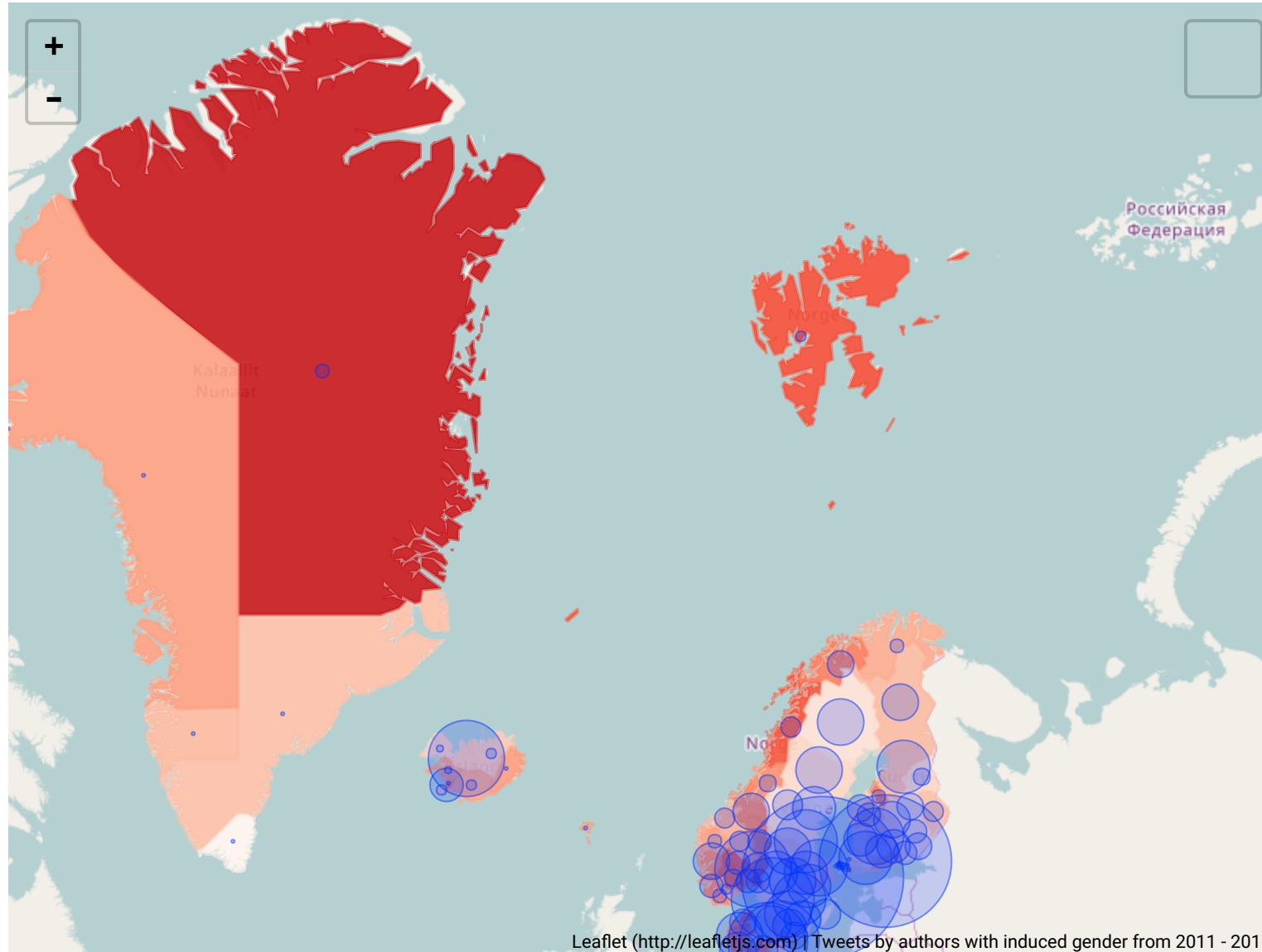
- Timelines (up to 3250 tweets) downloaded from the Twitter REST API for each gendered user
- Users with  $> 50\%$  Nordic place values and at least 20 tweets retained in data
- `tweet_source` filtered for sources more likely to be used by human agents to prevent (some) automated tweets sent by apps such that send English-language messages about music playlists, user location, workout details, weather, etc.

# Data collection | Language detection

- Languages of individual tweets were identified with the following procedure:
- Remove sequences that can cause inaccurate language detection (usernames, URLs, hashtags, emojis)
- Require three-way agreement: Retain only tweets whose language as reported by the native Twitter algorithm agrees with the language as reported by both the *compact language detector 2* (Sites 2014) and *langid* (Lui and Baldwin 2012)
- Working dataset: 25,509,920 tweets by 44,906 unique authors

# Tweet density by language

Map polygons from Natural Earth (<http://www.naturalearthdata.com>), maps from Open Street Maps ([www.openstreetmaps.org](http://www.openstreetmaps.org))

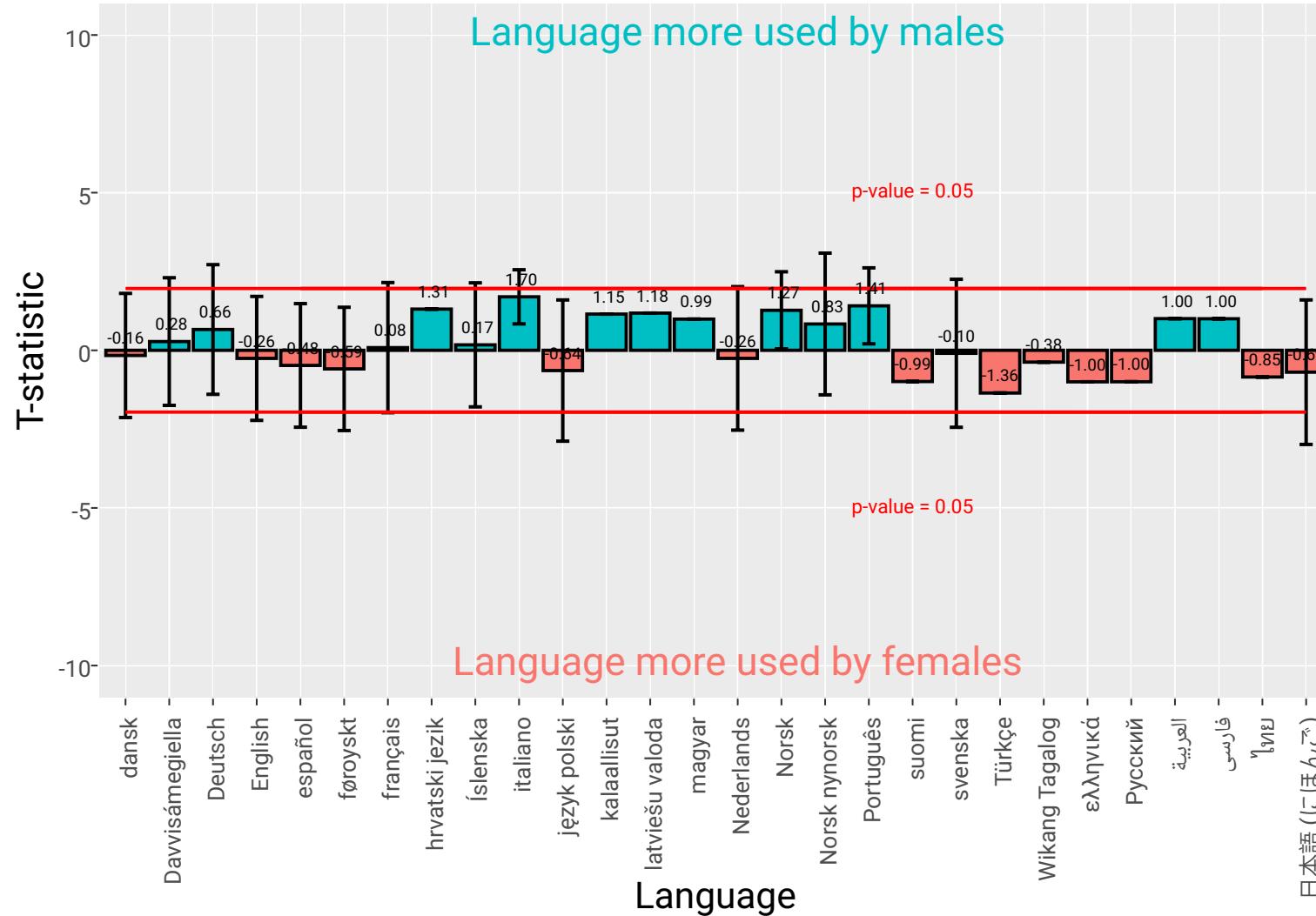


## Use of language by country and gender

- For each user, the proportion of all of his/her tweets in each of the 87 languages in the data was calculated
- For each of the five countries under consideration, the male-female balance was determined with a t-test of population means
- To account for unequal variance in the population distributions for some languages, 95% confidence intervals were constructed by using stratified non-parametric bootstrapping based on 10,000 resamples
- The 28 most frequently occurring languages overall are shown here

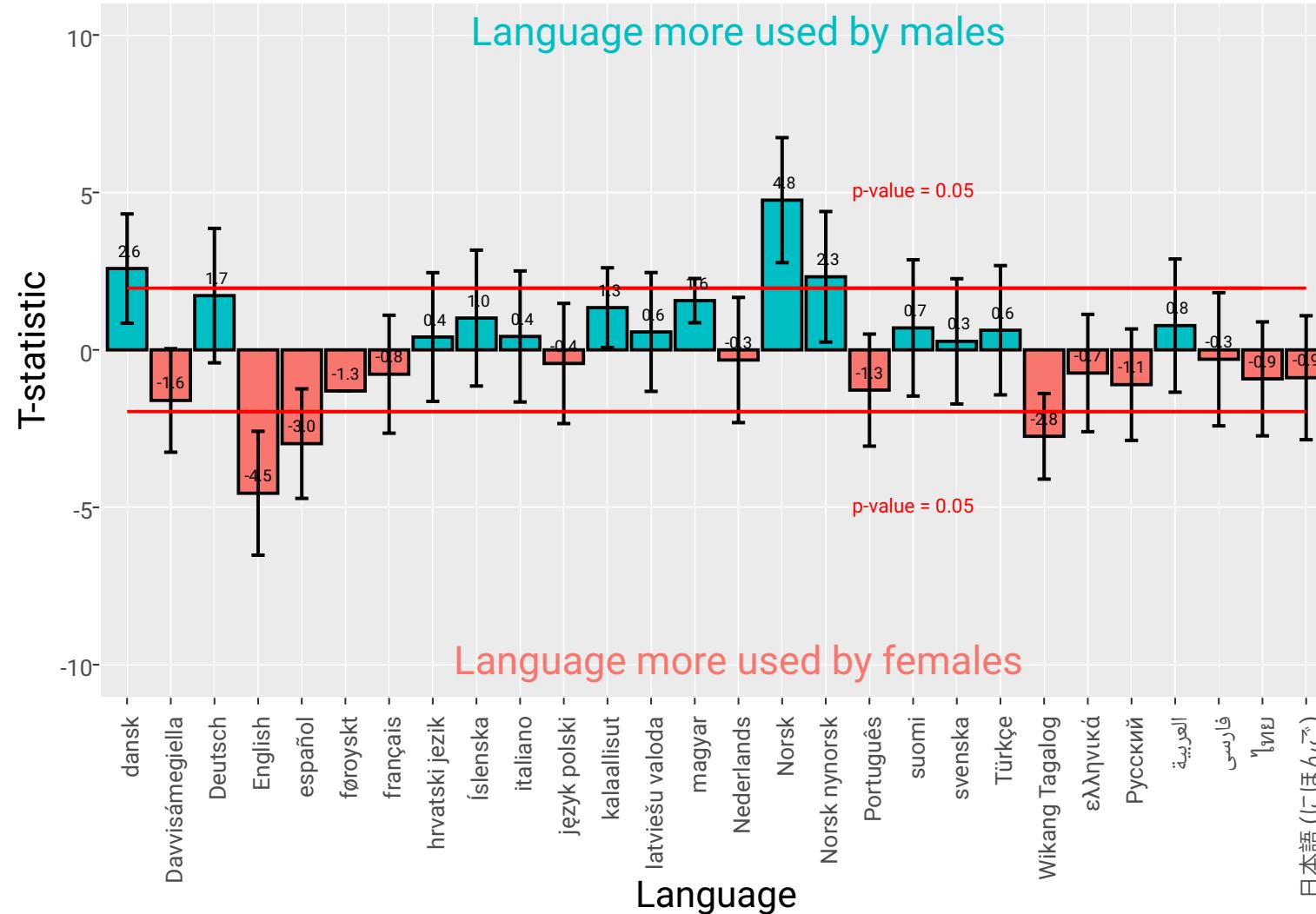
# Use of language by country and gender: Iceland

Iceland: 1362 users, 911297 tweets



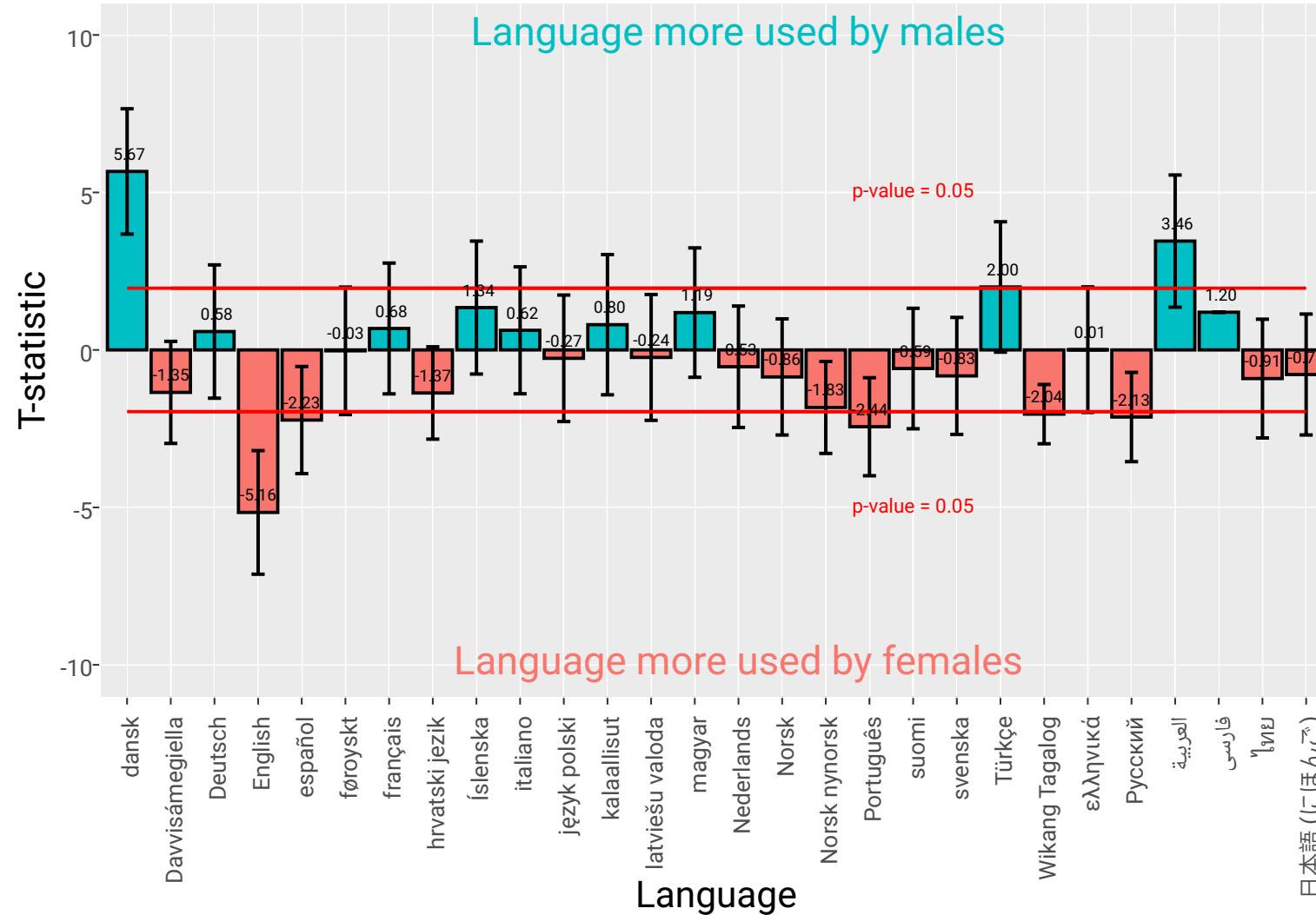
# Use of language by country and gender: Norway

Norway: 7162 users, 3229143 tweets



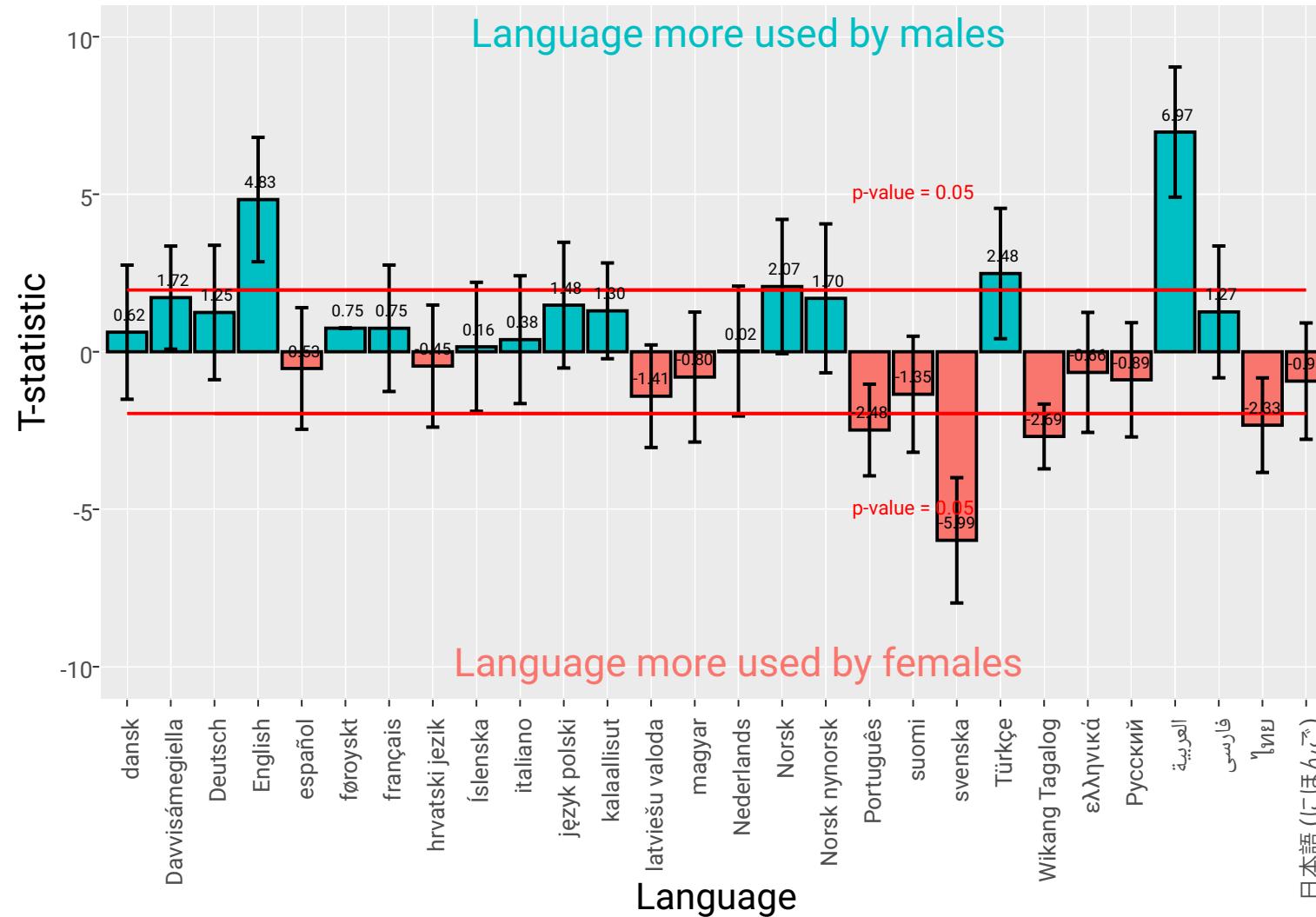
# Use of language by country and gender: Denmark

Denmark: 6602 users, 2858740 tweets



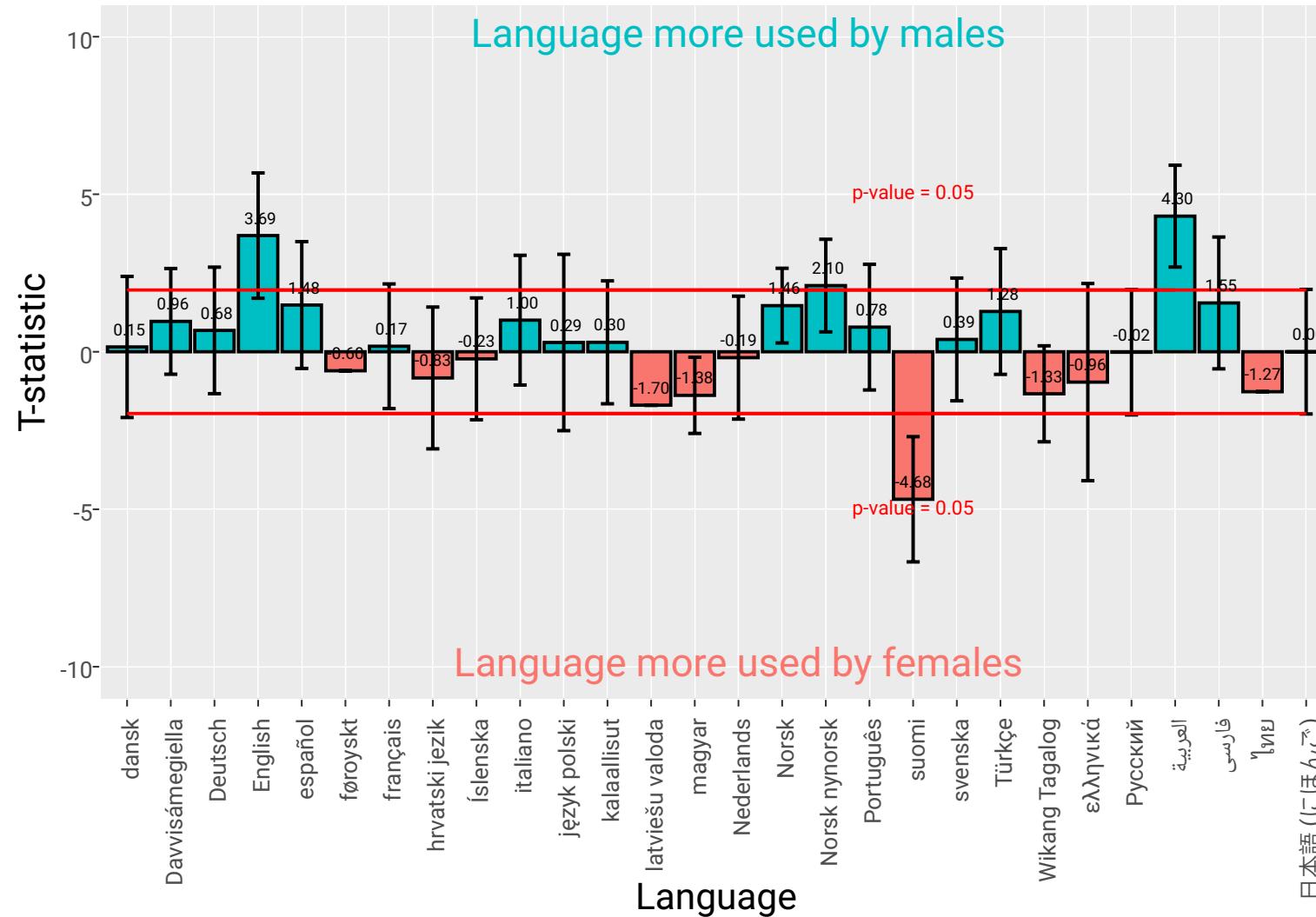
# Use of language by country and gender: Sweden

Sweden: 19933 users, 13249744 tweets



# Use of language by country and gender: Finland

Finland: 9617 users, 5064426 tweets

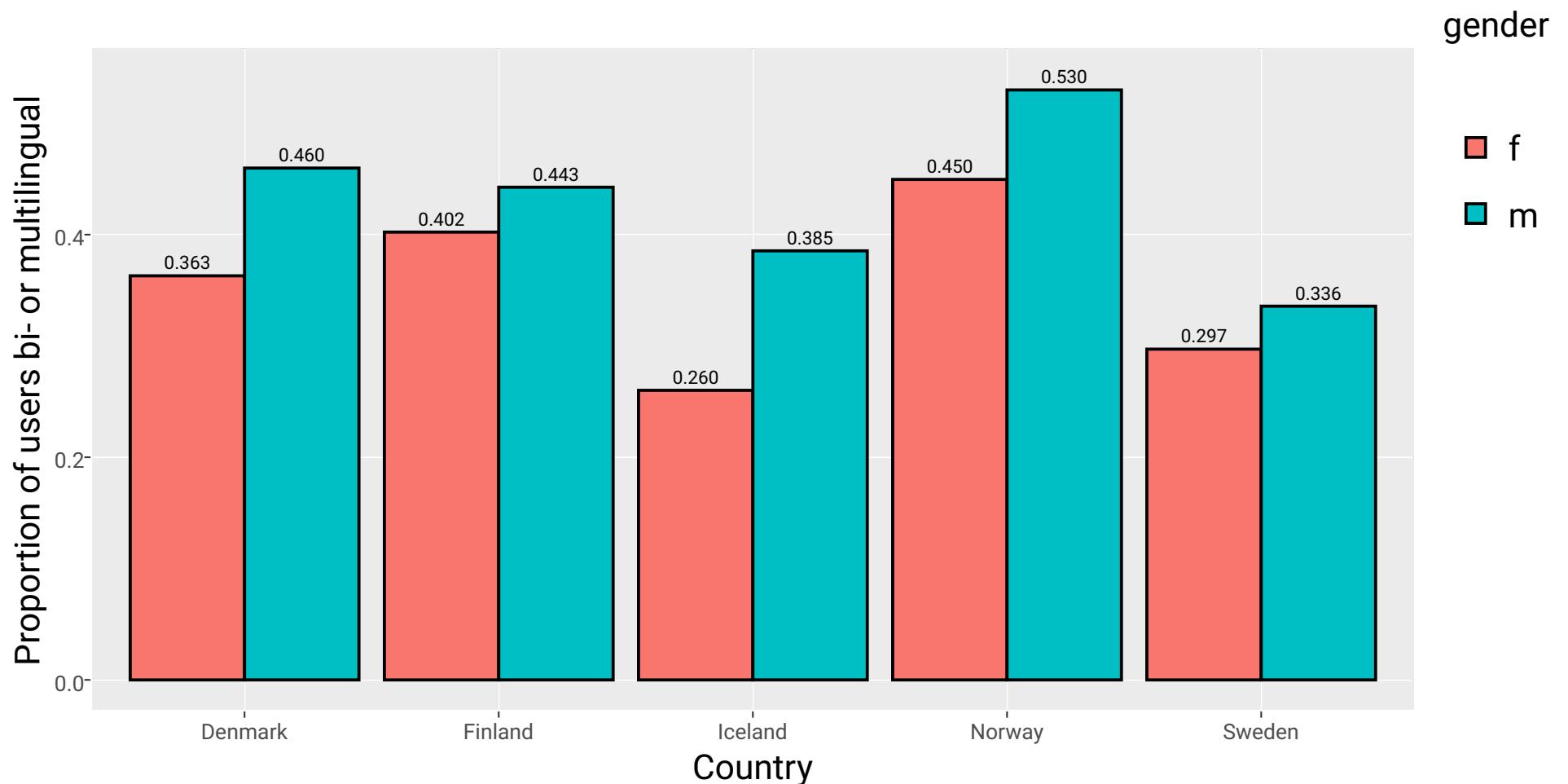


# Summary language use by gender

- Males use more Arabic, Farsi, Turkish; females use more Thai, Tagalog, Japanese, Russian
- These facts may represent gender differences in patterns of migration to the Nordics
- Iceland, Norway and Denmark: Males use more national languages, females use more English
- Sweden and Finland: Females use more national languages, males use more English
- These facts may represent language shift
- Labovian sociolinguistics: In changing linguistic situations, females use prestige markers more than males (1990)

# Quantifying bi- and multilingualism

- Users were defined as bi- or multilingual if for at least 2 languages each comprised  $\geq$  10% of the total number of his/her tweets
- In this data, males are more bi- or multilingual than females



# Creating networks of bi- and multilinguals

- To quantify the strength of connections between languages on Twitter in the Nordics, networks of bi- and multilingual users were created
- For each language pair  $(i, j)$  a contingency table of the number of bilinguals was set up:

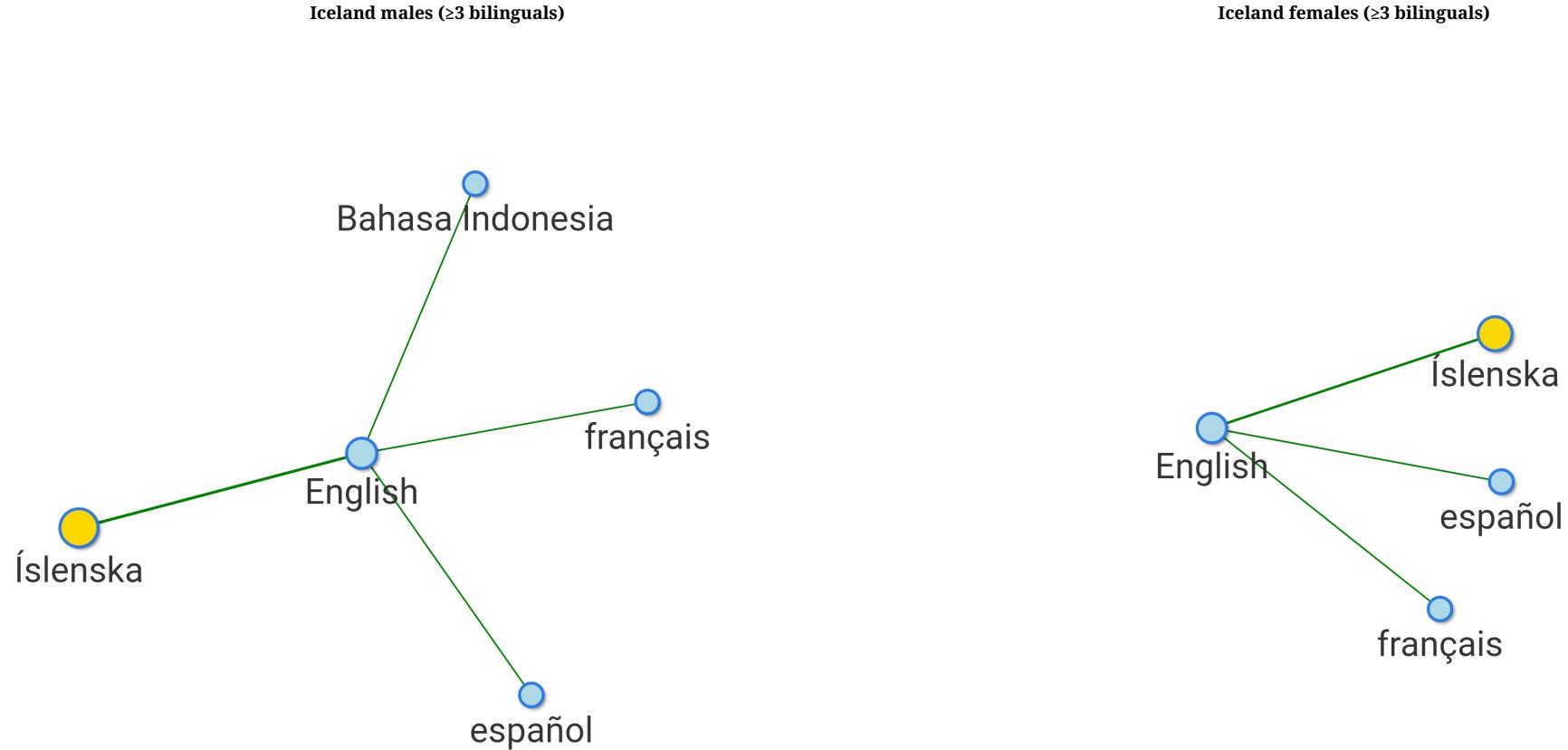
$$\begin{array}{ccccc} & \textit{language}_i & \sim & \textit{language}_i \\ \textit{language}_j & O_{11} & & O_{12} & = R_1 \\ \sim \textit{language}_j & O_{21} & & O_{22} & = R_2 \\ & = C_1 & & = C_2 & = N \end{array}$$

- From this table a correlation coefficient  $\phi$  can be calculated:

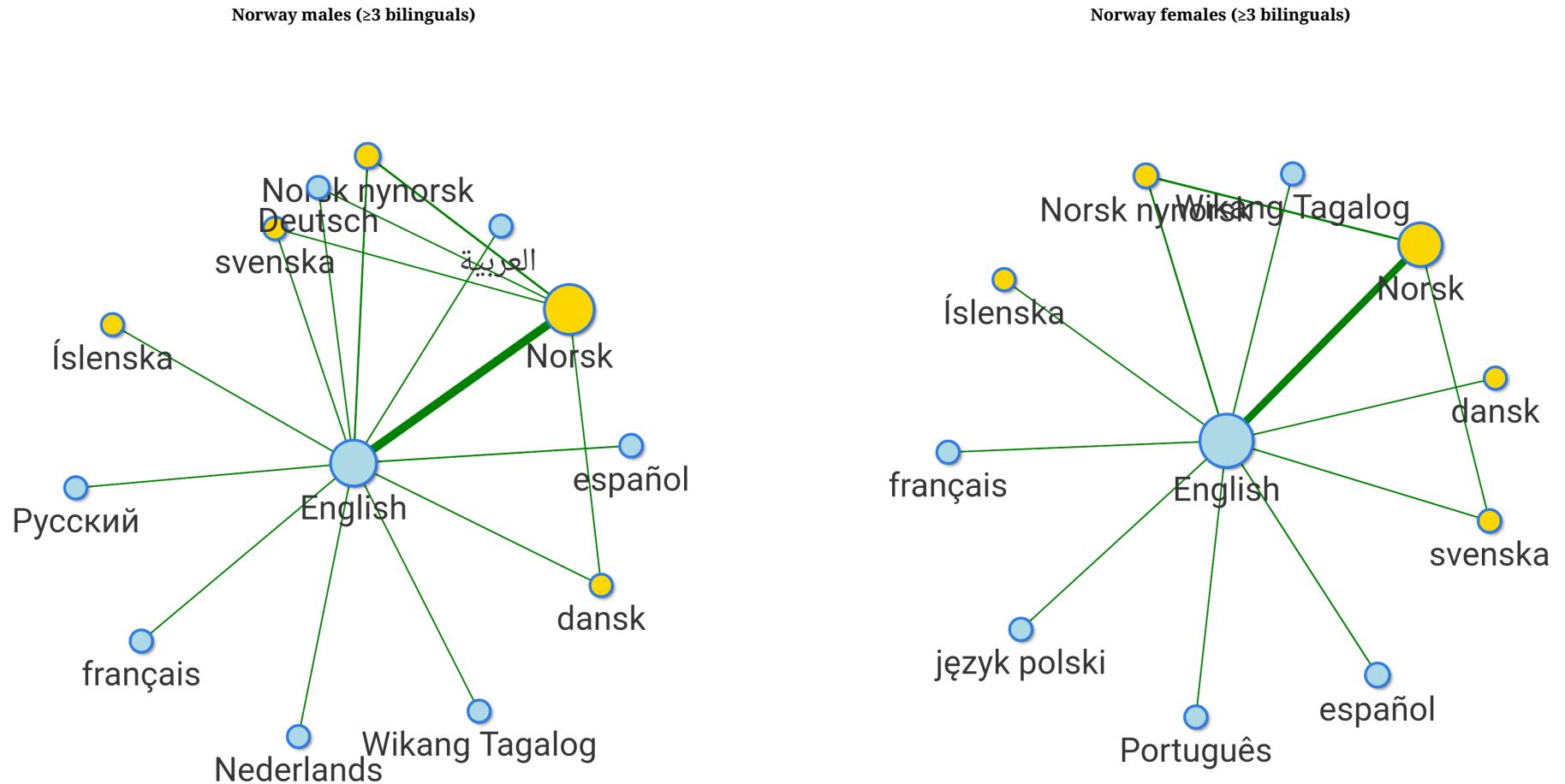
$$\phi_{ij} = \frac{(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{R_1R_2C_1C_2}}$$

- $\phi$  ranges in value from  $-1$  to  $1$ . A t-statistic and p-values were calculated for the  $\phi$  scores

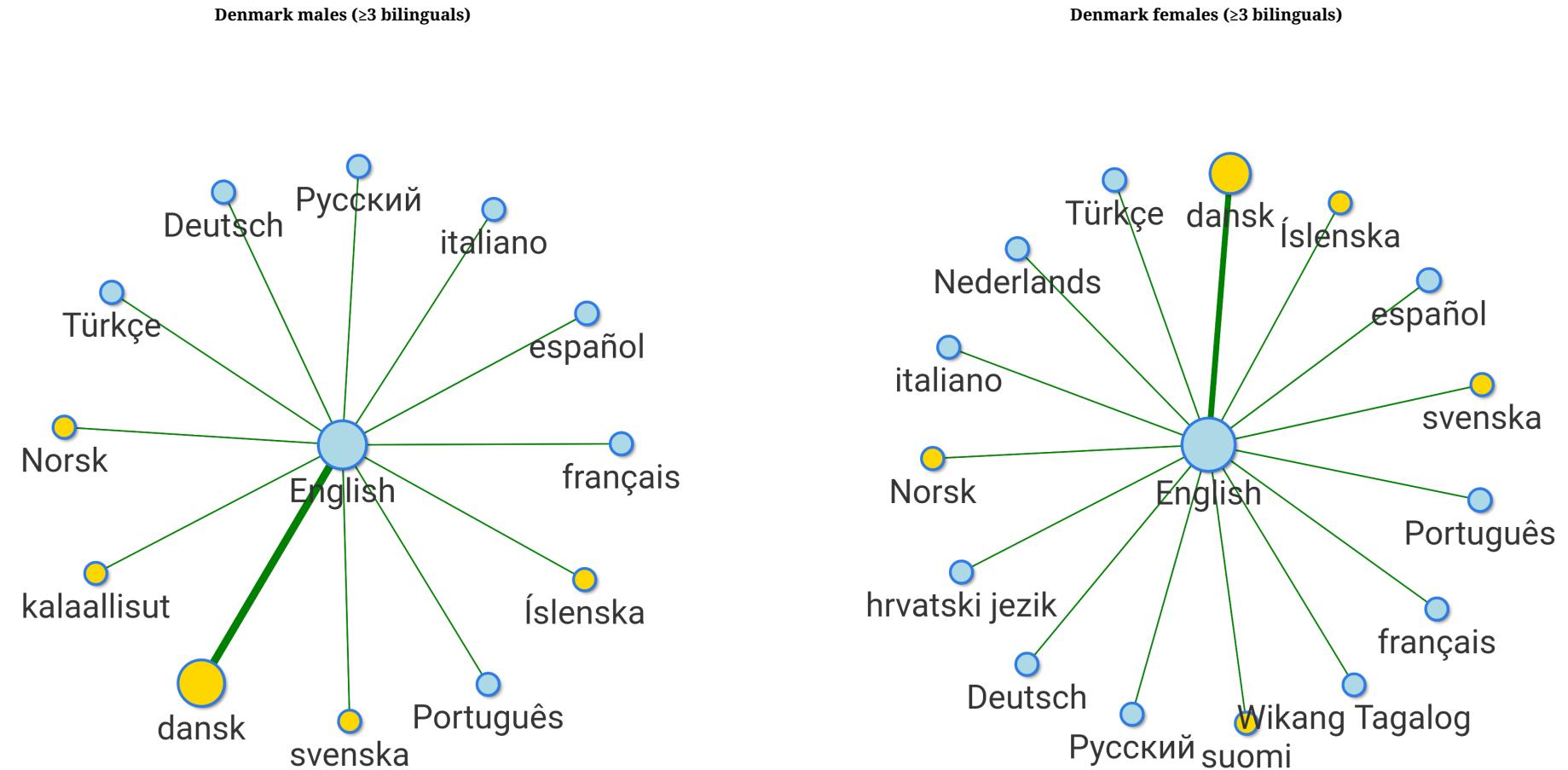
# Bilinguals' use of language by country and gender: Iceland



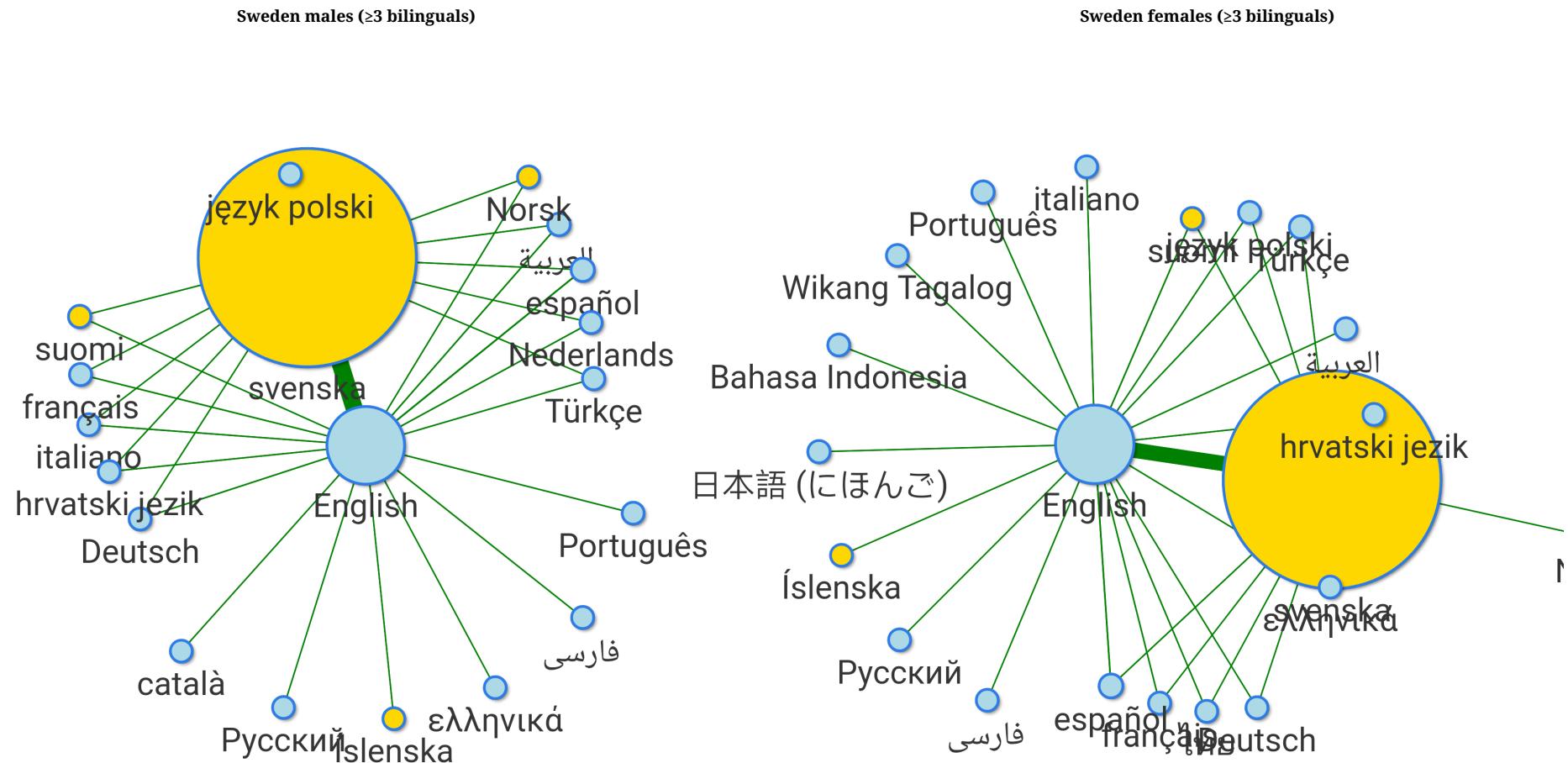
# Bilinguals' use of language by country and gender: Norway



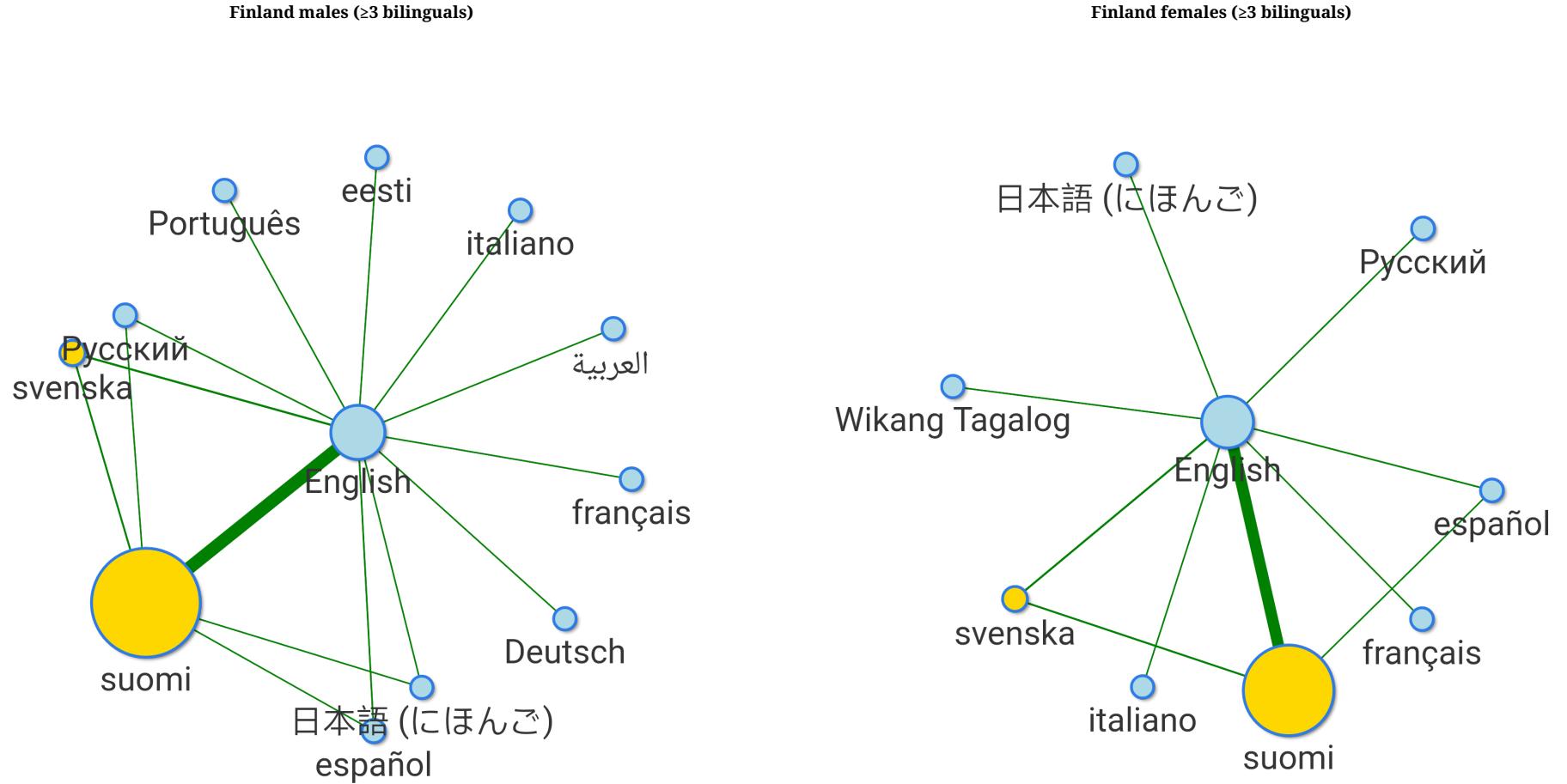
# Bilinguals' use of language by country and gender: Denmark



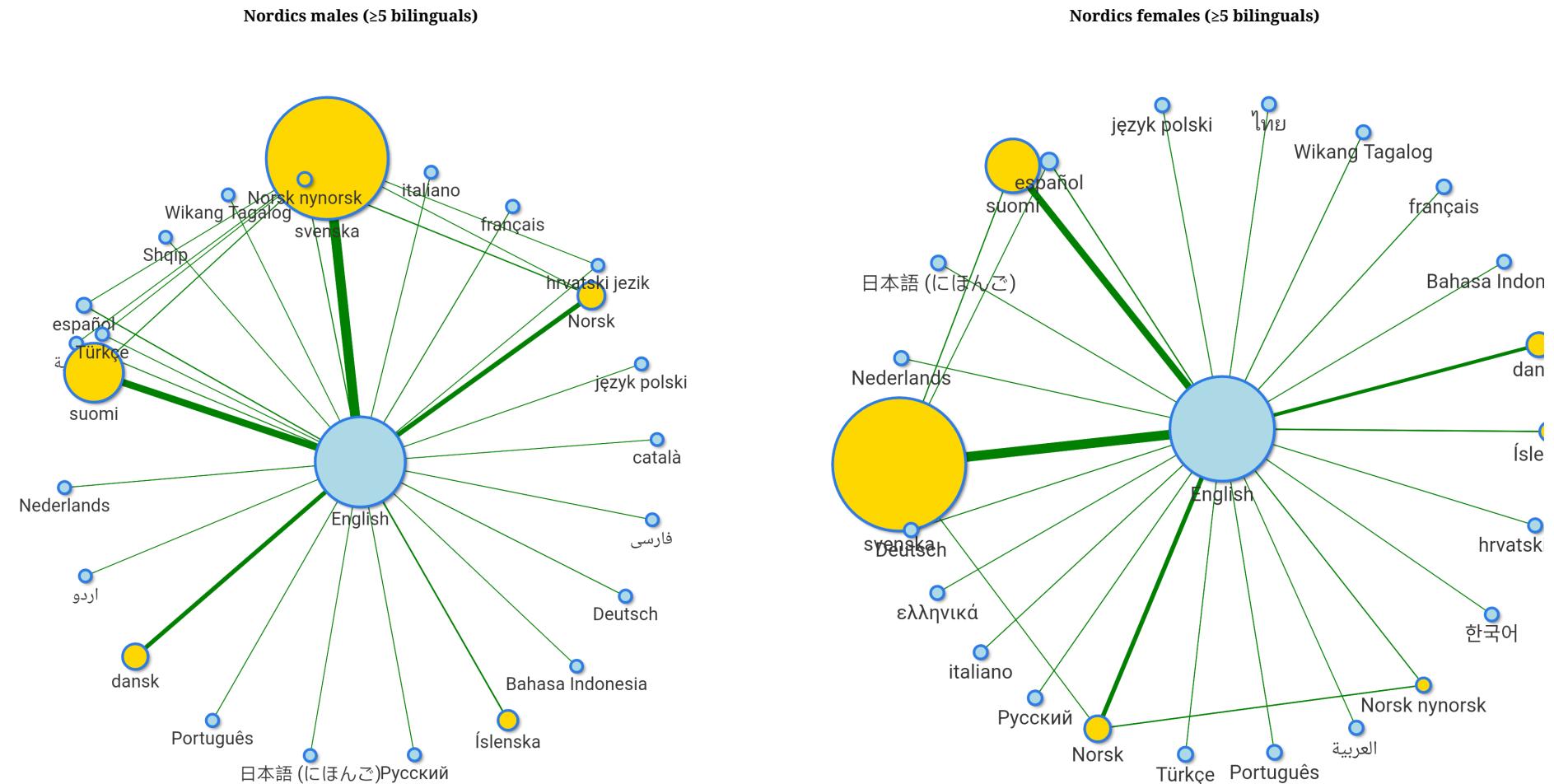
# Bilinguals' use of language by country and gender: Sweden



# Bilinguals' use of language by country and gender: Finland



# Clusters for all Nordic males and females



# Significance of gender differences

- A z-score was calculated via a Fisher transform of  $\phi$  values according to the formula:

$$z_{(i,j)} = \frac{z_{\phi_{(i,j)m}} - z_{\phi_{(i,j)f}}}{\sqrt{\frac{1}{n_{(i,j)m}-3} + \frac{1}{n_{(i,j)f}-3}}}$$

(Sheskin 2000: 792)

For languages  $(i, j)$ :

$z_{\phi_{(i,j)m}}$  is the Fisher transformed value of  $\phi$  for males

$z_{\phi_{(i,j)f}}$  is the Fisher transformed value of  $\phi$  for females

$n_{(i,j)m}$  and  $n_{(i,j)f}$  are the number bilingual speakers in the male and female networks.

# Summary of gender differences

Show 100 ▾ entries

Search:

	lang.i	lang.j	phi.m	phi.f	z.score	p.value
1	አማርኛ	English	0.002	0.002	0.023	0.491
2	العربية	dansk	-0.019	0	-0.965	0.167
3	العربية	English	-0.071	-0.061	-0.536	0.296
4	العربية	svenska	-0.028	-0.015	-0.701	0.242
5	العربية	Türkçe	0.075	0.074	0.029	0.489
6	català	English	-0.032	-0.046	0.709	0.239
7	català	español	0.114	0.107	0.342	0.366
8	čeština	English	0.004	0.002	0.076	0.47
10	dansk	Deutsch	-0.022	-0.012	-0.518	0.302
11	dansk	English	0.029	0.038	-0.473	0.318
12	dansk	français	-0.019	-0.012	-0.35	0.363
13	dansk	kalaallisut	0.027	0.056	-1.49	0.068

	<b>lang.i</b>	<b>lang.j</b>	<b>phi.m</b>	<b>phi.f</b>	<b>z.score</b>	<b>p.value</b>
14	dansk	Norsk	-0.176	-0.154	-1.162	0.123
15	dansk	Português	-0.005	-0.014	0.46	0.323
16	dansk	svenska	-0.308	-0.286	-1.243	0.107
17	dansk	Türkçe	-0.013	-0.002	-0.607	0.272
19	Deutsch	English	-0.008	-0.037	1.492	0.068
20	Deutsch	svenska	-0.04	-0.029	-0.53	0.298
21	ελληνικά	English	0.004	0.005	-0.054	0.479
29	English	español	-0.013	-0.015	0.065	0.474
30	English	eesti	0.003	0.004	-0.053	0.479
31	English	فارسی	-0.018	0.004	-1.161	0.123
32	English	suomi	0.003	0.004	-0.038	0.485
33	English	føroyskt	0.002	0.002	-0.014	0.494
34	English	français	-0.055	-0.043	-0.619	0.268
35	English	hrvatski jezik	-0.05	-0.054	0.173	0.431

	<b>lang.i</b>	<b>lang.j</b>	<b>phi.m</b>	<b>phi.f</b>	<b>z.score</b>	<b>p.value</b>
36	English	Bahasa Indonesia	0.005	0.005	-0.001	0.5
37	English	Íslenska	0.015	0.022	-0.357	0.36
39	<b>English</b>	日本語 (にほんご)	-0.06	0.006	-3.389	0
40	English	kalaallisut	0.004	0.003	0.033	0.487
44	<b>English</b>	<b>Nederlands</b>	0.008	-0.027	1.814	0.035
45	<b>English</b>	<b>Norsk nynorsk</b>	-0.505	-0.477	-1.905	0.028
46	English	Norsk	-0.143	-0.158	0.767	0.221
47	<b>English</b>	<b>język polski</b>	-0.024	-0.06	1.887	0.03
48	English	Português	-0.016	-0.013	-0.157	0.438
49	English	Română	0.003	0.003	0.008	0.497
50	English	Русский	-0.013	-0.032	0.969	0.166
51	English	Soomaaliga	0.003	0.002	0.052	0.479
52	English	Shqip	0.004	0.002	0.098	0.461
53	English	svenska	0.012	-0.003	0.777	0.219

	<b>lang.i</b>	<b>lang.j</b>	<b>phi.m</b>	<b>phi.f</b>	<b>z.score</b>	<b>p.value</b>
55	English	ไทย	0.003	0.007	-0.192	0.424
56	English	Wikang Tagalog	0.005	0.011	-0.328	0.372
57	<b>English</b>	<b>Türkçe</b>	-0.154	-0.106	-2.503	0.006
62	español	suomi	-0.064	-0.064	-0.018	0.493
63	español	svenska	-0.08	-0.075	-0.263	0.396
65	eesti	suomi	0.004	-0.001	0.269	0.394
70	suomi	日本語 (にほんご)	0.017	-0.013	1.548	0.061
71	suomi	Русский	-0.015	-0.034	0.991	0.161
72	suomi	svenska	-0.401	-0.422	1.319	0.094
76	français	svenska	-0.046	-0.048	0.069	0.473
78	hrvatski jezik	svenska	0.004	-0.02	1.267	0.103
90	Nederlands	Norsk	-0.018	-0.006	-0.628	0.265
91	Nederlands	svenska	-0.026	-0.024	-0.106	0.458
93	<b>Norsk nynorsk</b>	<b>Norsk</b>	0.262	0.301	-2.171	0.015

# Preliminary summary and conclusions

- Aggregate language usage statistics may reflect both patterns of international mobility as well as cultural associations of languages/language learner preferences
- English is the central language of bilingualism for both males and females
- Males are more likely to be bilingual overall
- Comparison of phi values somewhat supports sociolinguistic interpretation (more data needed)
- There is some evidence for higher idea density (i.e. longer tweets) for bilinguals

# Acknowledgements

Thanks to

- Dorthe Larsen (Danmarks Statistik) for Danish name lists
- Joar Skott (Statistika centralbyrån, Sweden) for Swedish name lists
- Finland's Center for Scientific Computing ([www.csc.fi](http://www.csc.fi)) for data storage and computing resources

# References

- Argamon, S., M. Koppel, J. W. Pennebaker and J. Schler. 2007. Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday* 12/9. <http://pear.accc.uic.edu/ojs/index.php/fm/article/view/2003/1878> (<http://pear.accc.uic.edu/ojs/index.php/fm/article/view/2003/1878>)
- Bamann, D., J. Eisenstein and T. Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2), 135–160. <http://onlinelibrary.wiley.com/doi/10.1111/josl.12080/full> (<http://onlinelibrary.wiley.com/doi/10.1111/josl.12080/full>)
- Coats, S. 2016. Grammatical feature frequencies of English on Twitter in Finland. In L. Squires (ed.), *English in computer-mediated communication: Variation, representation, and change*. Berlin: De Gruyter. 179–210.
- Labov, W. 1990. “The intersection of sex and social class in the course of linguistic change.” *Language Variation and Change* 2, 205–254.
- Lui, M. and T. Baldwin. 2012. “Langid.py: An off-the-shelf language identification tool”. 50th Proceedings of the Association for Computational Linguistics, 25–30. Stroudsburg, PA: ACL. <http://dl.acm.org/citation.cfm?id=2390475> (<http://dl.acm.org/citation.cfm?id=2390475>)
- Roesslein, J. 2015. Tweepy. Python package [Computer software]. <http://www.tweepy.org> (<http://www.tweepy.org>)
- Sheskin, D. 2000. *Handbook of parametric and non-parametric statistical procedures*, 2nd ed. Boca Raton: Chapman and Hall.
- Sites, D. 2013. Compact language detector 2. <https://github.com/CLD2Owners/cld2> (<https://github.com/CLD2Owners/cld2>)