Gender and lexical type frequencies in Finland Twitter English

Steven Coats

English Philology, University of Oulu, Finland

22 October 2015
d2e: from Data to Evidence

ULEÅBORGS UNIVERSITET
УЛЕАБОРГСКИЙ УНИВЕРСИТЕТ

Table of Contents

1 Introduction: Twitter English and Finland Twitter English

2 Data Collection and Processing

- Collecting and Processing the Data
- Language Detection
- Geographical Distribution of Language
- Automatic PoS Tagging
- Gender Disambiguation

3 Selected Lexical Features

- Lexical Features

4 Conclusion and Summary

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 on Twitter in Finland: CMC “Global Englishes” and the status of English in (traditionally) non–Anglophone societies ("Expanding Circle", Kachru 1990)
- Gendered differences in CMC language (Wolf 2000, Baron 2004, Herring and Paolillo 2006, Herring 2013, Bamann et al. 2014)
- Selected non–standard lexical features: Expressive lengthening and emoticons

ULEÅBORGS UNIVERSITET
УЛЕАБОРГСКИЙ УНИВЕРСИТЕТ

Table of Contents

1 Introduction: Twitter English and Finland Twitter English

2 Data Collection and Processing

- Collecting and Processing the Data
- Language Detection
- Geographical Distribution of Language
- Automatic PoS Tagging
- Gender Disambiguation

3 Selected Lexical Features

- Lexical Features

4 Conclusion and Summary

Finland and Comparison Corpora Data Collection

- Finland data: Python script access to Twitter Streaming API
  - Access levels, extent of geo–encoded user messages (1.6% of tweets according to Leetaru et al. 2013)
  - Geo–coordinates bounding box
  - Filtering using data from GADM and packages in *R* (maptools, mapdata)
  - Two data sets collected: 2013 and 2015
- Comparison data: 2009 data from Texas A&M Univ. from Twitter Streaming API (no geo–coordinates), 2015 tweets geo-located to United States

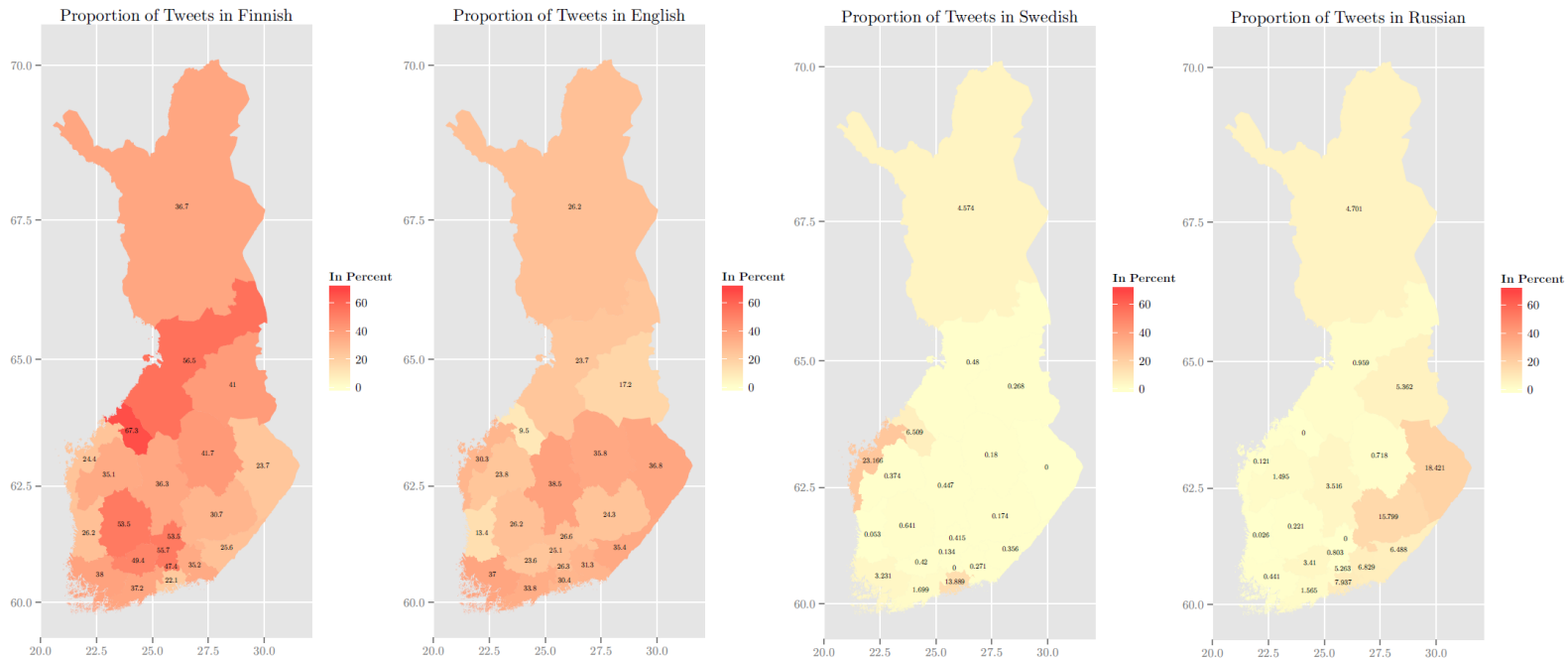
Language Detection

- Twitter provides a field in Tweet entity indicating language since mid–2013; prior to this have to detect language using own tools
- Automatic language disambiguation using langid.py (Lui and Baldwin 2012)

	Text	Lang	Prob
1	Yo are the stores open bc i was gonna go to herushinki tomorrow	en	0.999
2	@KristiinaKomula Have fun....	en	0.593
3	@rrebeckayes haha kul att jag skrattade i ca 10 min åt den själv :)	sv	0.999
4	Кто нибудь, дайте пищу для сквернословия.. а то в голову ничего не лезет.	ru	0.999
5	@rollersitar En oikein viä tiä viikonlopun suunnitelmista. :-/	fi	1.000
6	@wesa66 Tottakai.	fi	0.443

- Tweets with probabilistic language ID values > 0.6 retained for analysis.

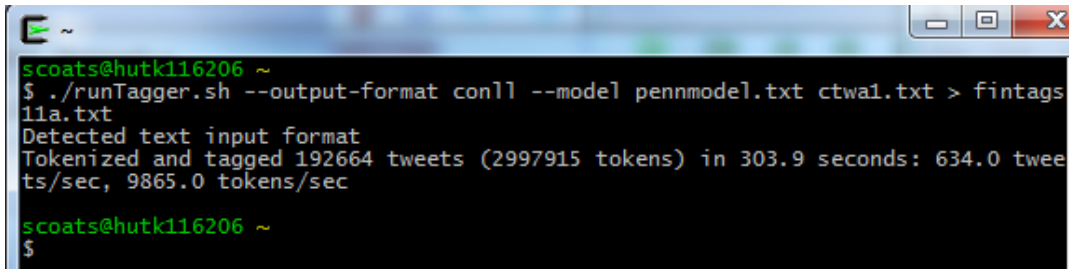
Percent of Tweets by Language and by Finnish Province



Languages of Finland Tweets (2013): 44.8% Finnish, 35.7% English, 2.2% Swedish 2.1% Russian, 15.2% Other

Tokenization and PoS Tagging

- Carnegie-Mellon Twitter PoS Tagger (Gimpel et al. 2011; Gimpel et al. 2013, Owoputi et al. 2013); Java code run in Unix shell (Cygwin)



- PoS tags, additional tags for Twitter–specific types (retweet, username, hashtag), can use Penn Treebank tags (Marcus et al. 1993)

```
317   @knislemc                    @   0,9968
318   My                           D   0,9986
319   toes                         N   0,9966
320   are                          V   0,9828
321   cold                         A   0,9985
322   .                            ,   0,9979
323   ;-)                          E   0,9889
324   #life                        #   0,9713
325   https://t.co/HwHcwlCzmZ      U   0,9860
```

- Output consists of tab–separated token/tag/prob tokens, including e.g. emoticons

Gender Disambiguation

- Usernames filtered for strings that include 200 most common male and female given names in Finland and the United States
- Tweets from Finland by a user with the (invented) username
  "PäiviMäkinen" → female
  "twenty10intl" → ignored
- Tweets from Comparison data by a user with the username
  "trevorOZ" → male
  "zYlax85" → ignored
- Names in both groups removed (e.g. "AnneliPuoskari" contains female string "anne" and male string "oskari")
- Data: names of all registered persons in Finland as of 1.1.2014 (from the Population Information Center, *Väestötietojärjestelmä*), 218 million persons born in the United States from 1950-2000 (Census Bureau)

Corpora Summary Statistics

Corpus	Messages	Tokens	Types
Comparison English 2009	192,664	3,067,142	236,082
Comparison English Male 2009	32,799	477,463	34,300
Comparison English Female 2009	16,514	269,873	20,066
US English 2015	250,756	2,907,048	242,320
US English Male 2015	32,799	477,574	55,420
US English Female 2015	16,514	270,024	32,533
Finland English 2013	32,956	388,511	70,949
Finland English Male 2013	2,999	41,150	11,607
Finland English Female 2013	4,648	57,098	11,423
Finland English 2015	20,882	271,188	52,145
Finland English Male 2015	2,056	25,849	7,985
Finland English Female 2015	2,447	30,539	6,907

ULEÅBORGS UNIVERSITET
УЛЕАБОРГСКИЙ УНИВЕРСИТЕТ

Table of Contents

1 Introduction: Twitter English and Finland Twitter English

2 Data Collection and Processing

- Collecting and Processing the Data
- Language Detection
- Geographical Distribution of Language
- Automatic PoS Tagging
- Gender Disambiguation

3 Selected Lexical Features

- Lexical Features

4 Conclusion and Summary

Quantifying Similarity

- A relative frequency statistic (log odds ratio) was calculated for all lexical types in the corpora.

	$corpus_1$	$corpus_2$	
$word/PoS$	O_{11}	O_{12}	$= R_1$
$\sim word/PoS$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

$$log\ odds\ ratio\ \theta = \log \frac{O_{11}O_{22}}{O_{12}O_{21}}$$

- Types with most extreme values are keywords, i.e. the lexemes most distinctive for the discourse of each corpus.

Lexical Features: Finland 2013 vs. Global 2009

- Lexical type frequencies reflect discourse–related topicality specific to the varieties
- Some highly frequent types reflect the temporal and geographical parameters of the data collection

	type	θ	type	θ
1	finland	7.94	last.fm	-5.16
2	❤	7.33	steelers	-4.97
3	niall	6.19	#listening	-4.32
4	helsinki	6.05	xbox	-4.20
5	finnish	5.65	2008	-4.11
6	♥	5.53	>>	-4.03
7	xx	5.44	scout	-3.89
8	sweden	5.27	flickr	-3.78
9	hel	5.15	herbal	-3.71
10	#party	4.74	dashboard	-3.67
11	#food	4.74	obama	-3.64
12	✌	4.71	palin	-3.62
13	:dd	4.57	nfl	-3.35
14	ikr	4.56	reader	-3.31
15	apparatus	4.56	firefox	-3.29
16	rn	4.53	ebay	-3.25
17	;))	4.51	blog	-3.19
18	gaga	4.46	twittering	-3.14
19	casually	4.46	bob	-3.11
20	youu	4.46	entry	-3.01

Lexical Features: Finland 2015 vs. US 2015

- Finland: local entities, anglophone celebrity culture
- Global/US: local entities, marketing entities, non-standard dialectal forms

	type	θ	type	θ
1	helsinki	8.87	tx	-3.85
2	finland	7.10	nigga	-3.68
3	109	7.06	fl	-3.63
4	hel	6.86	ca	-3.57
5	#cycling	6.79	#healthcare	-3.51
6	bicycle	6.47	#hospitality	-3.42
7	finnish	6.30	#memorialday	-3.42
8	oy	5.82	tf	-3.34
9	@real_liam_payne	5.75	#tweetmyjobs	-3.32
10	#risingstar	5.66	#hiring	-3.24
11	@chadwhitexxx	5.59	prom	-3.24
12	@planetjedward	5.47	humidity	-3.13
13	#survivor	5.42	va	-3.08
14	@brytonejames	5.42	#veteranjob	-3.06
15	#pens	5.33	#job	-3.05
16	hemmings	5.33	y'all	-3.05
17	#survivorsecondchance	5.29	sonic	-2.99
18	@bethreamer	5.24	yall	-2.98
19	counter	5.23	#retail	-2.89
20	@beamiller	5.15	asf	-2.89

Lexical Features: Finland 2013 Male vs. Female

- Males: nouns, initialisms, punctuation (automated tweets?)
- Females: usernames, personal names, hashtags, interactive types

	type	θ	type	θ
1	using	3.14	@harry_styles	-4.02
2	~	2.98	@austinmahone	-3.91
3	@alexstubb	2.84	@real_liam_payne	-2.66
4	w	2.49	#happybirthdayaustin	-2.58
5	vr	2.27	followers	-2.51
6	))	2.27	aww	-2.38
7	areena	2.27	@niallofficial	-2.36
8	project	2.22	#wtfjb	-2.33
9	north	2.17	ily	-2.33
10	☺	2.14	cry	-2.29
11	workout	2.13	x	-2.26
12	lunch	2.09	xxx	-2.24
13	design	2.09	hi	-2.14
14	running	2.06	boring	-2.13
15	style	2.06	@louis_tomlinson	-2.07
16	currently	2.02	ppl	-2.07
17	quite	1.98	:/	-1.99
18	event	1.94	harry	-1.99
19	star	1.94	justin	-1.97
20	center	1.94	mum	-1.94

Lexical Features: Finland 2015 Male vs. Female

- Males: named entities, places, hashtags (automated tweets?)
- Females: interactive types, emotional affect types

	type	θ	type	θ
1	#ff	3.59	sleep	-2.78
2	#wine	3.24	shit	-2.35
3	tampere	2.83	pic	-2.30
4	rt	2.54	justin	-2.29
5	friday	2.33	honour	-2.23
6	southern	2.20	@lauraschwartz	-2.10
7	(@	2.11	birthday	-2.07
8	#bubbles	2.05	haha	-2.03
9	espoo	2.05	actually	-2.03
10	#vscocam	2.01	head	-2.03
11	@helsinkiairport	1.99	:p	-1.99
12	hel	1.99	omg	-1.98
13	vantaa	1.94	dont	-1.84
14	vote	1.90	please	-1.82
15	xd	1.90	song	-1.82
16	learning	1.81	talk	-1.78
17	final	1.78	everyone	-1.78
18	design	1.78	direct	-1.78
19	system	1.78	you've	-1.78
20	set	1.78	won't	-1.78

Lexical Features: US 2015 Male vs. Female

- Males: professions (healthcare)
- Females: celebrity culture, usernames, interactive types

	type	θ	type	θ
1	#healthcare	4.99	citizen	-3.01
2	#rn	3.68	@nashgrier	-2.98
3	breakers	3.61	cam	-2.66
4	makers	3.60	#romance	-2.62
5	anti-war	3.60	nash	-2.57
6	60s	3.60	classroom	-2.55
7	ofthe	3.57	#gop	-2.45
8	pathologist	3.44	ancestors	-2.45
9	#education	2.92	suite	-2.40
10	nm	2.87	kayla	-2.39
11	physician	2.70	lo	-2.35
12	health	2.55	freakin	-2.29
13	therapist	2.55	@kdvr	-2.29
14	james	2.46	#sorrynotsorry	-2.19
15	speech	2.43	boyfriends	-2.19
16	lebron	2.41	extended	-2.19
17	#tcot	2.36	cuban	-2.13
18	law	2.30	#raw	-2.08
19	logic	2.27	ytd	-2.08
20	dakota	2.27	ships	-2.08

Non-standard Lexical Feature: Expressive Lengthening

- Repetition of individual characters in a word string (e.g. *cooooooool, yesssssss, dumbbbb*)
- Has been interpreted as discourse marker of emotional affect in CMC (Rao et al. 2010, Schnoebelen 2012, Bamann, Eisenstein and Schnoebelen 2014)
- Identification: All tokens that contain three or more characters in sequence considered
- Procedure: Loop regex over all letters for sequences of length 3-10 characters; exclude urls, usernames, hashtags

```
for(i in 1:26){assign(paste("C.letters.male.3.",letters[i],+
sep=""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{3}+
(?!",letters[i],")(.*)",sep=""),us.male.words,perl=T))}+
# assigns every instance of 3 letters in sequence to a variable
for(t in 1:8){for(b in 1:26){eval(parse(text=paste(+
"length.C.",t+2,"[[",b,"]]","<-","length(C.letters.male",t+2,+
".",letters[b],")",sep="")))}}
```

Expressive Lengthening by Geography

- Finland English tweets exhibit more expressive lengthening in 2013 than 2015 ; Comparison/US tweets less expressive lengthening in 2009 than 2015.

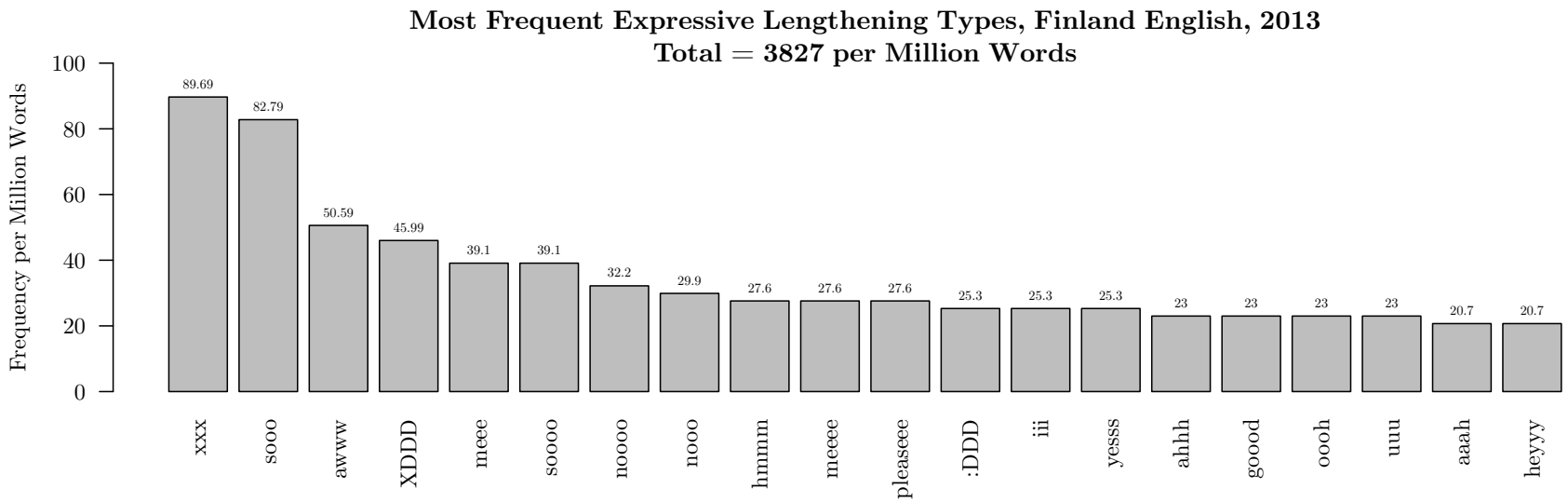
Corpus	Lengthenings/1m Tok
Comparison English 2009	1590
Finland English 2013	3827
US English 2015	2623
Finland English 2015	2125

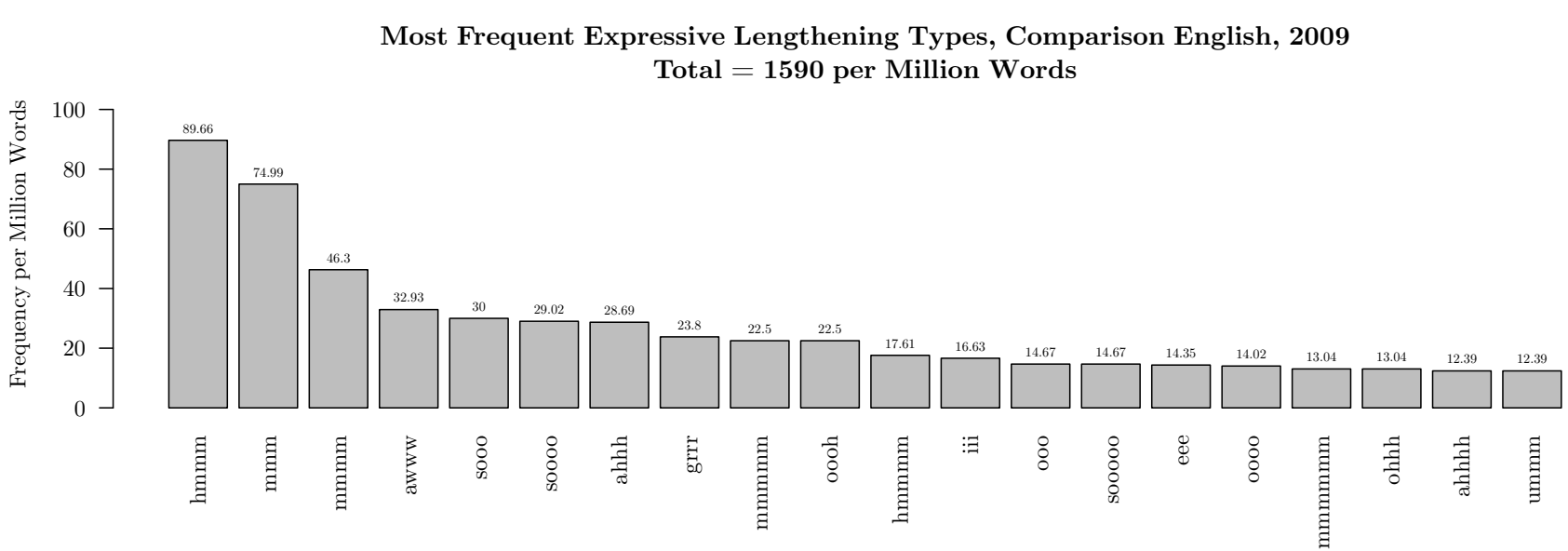
Expressive Lengthening by Gender

- Tweets by females exhibit more expressive lengthening; Male/female discrepancy more pronounced for Finland English.

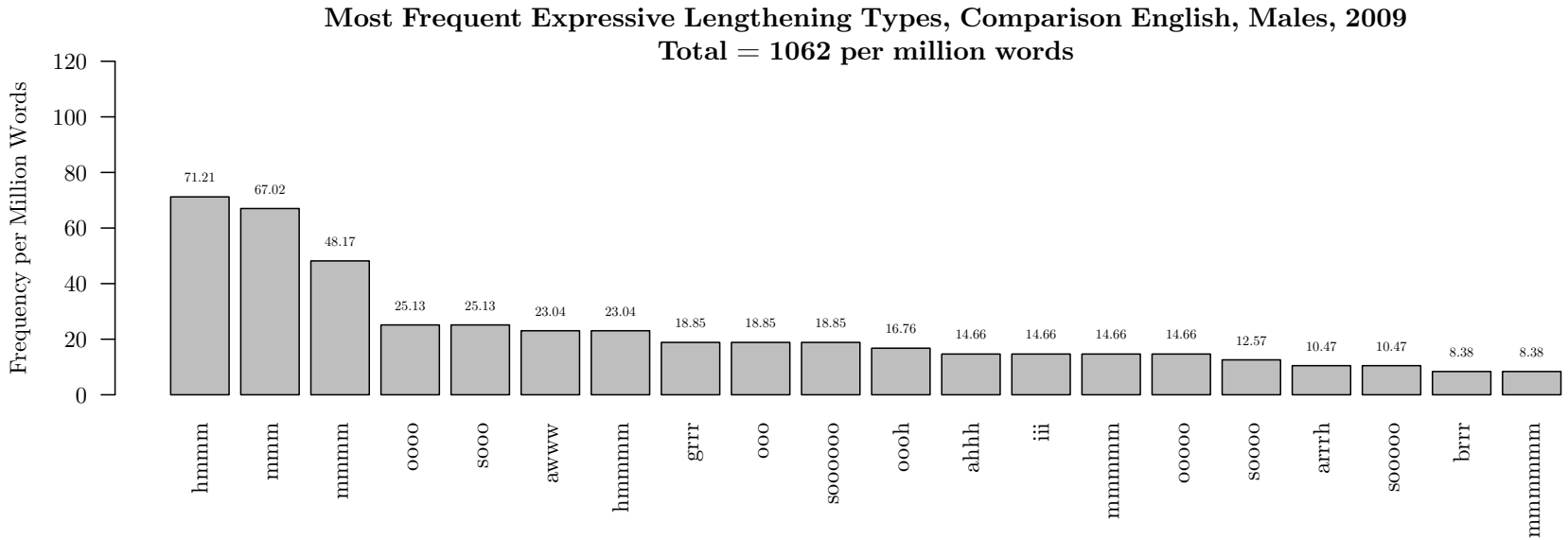
Corpus	Lengthenings/1m Tok
Comparison English 2009 Male	1062
Comparison English 2009 Female	1434
Finland English 2013 Male	1884
Finland English 2013 Female	4645
US English 2015 Male	2219
US English 2015 Female	3210
Finland English 2015 Male	1199
Finland English 2015 Female	3831

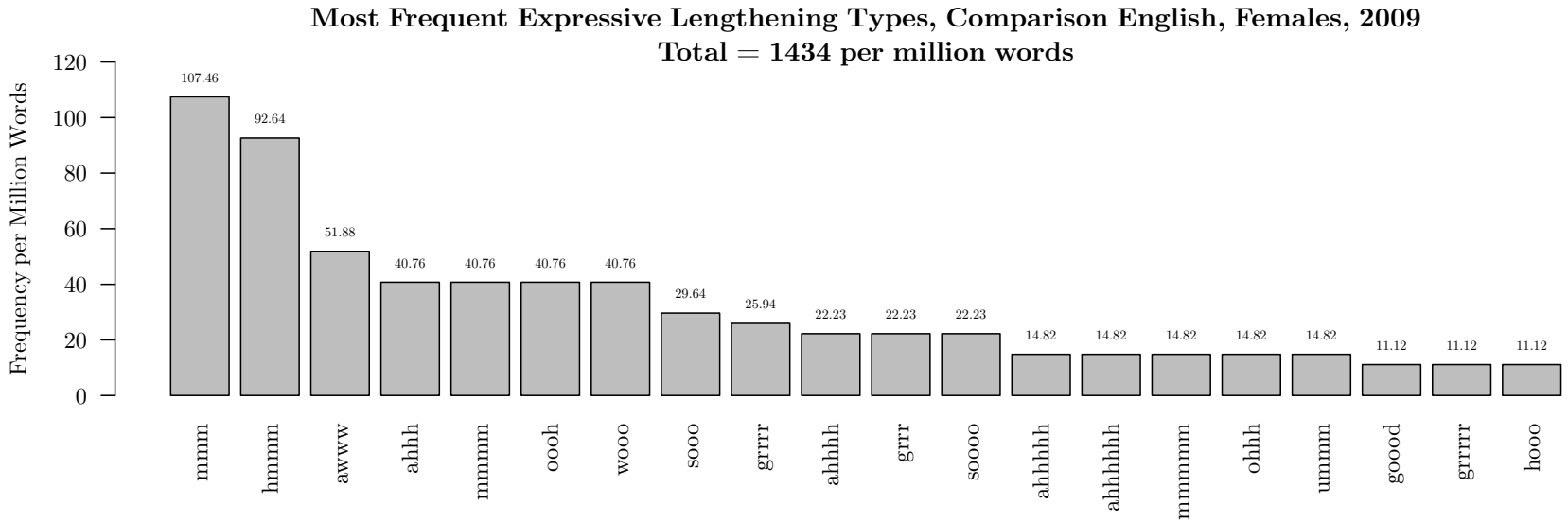
Expressive Lengthening by Geography, 2013/2009



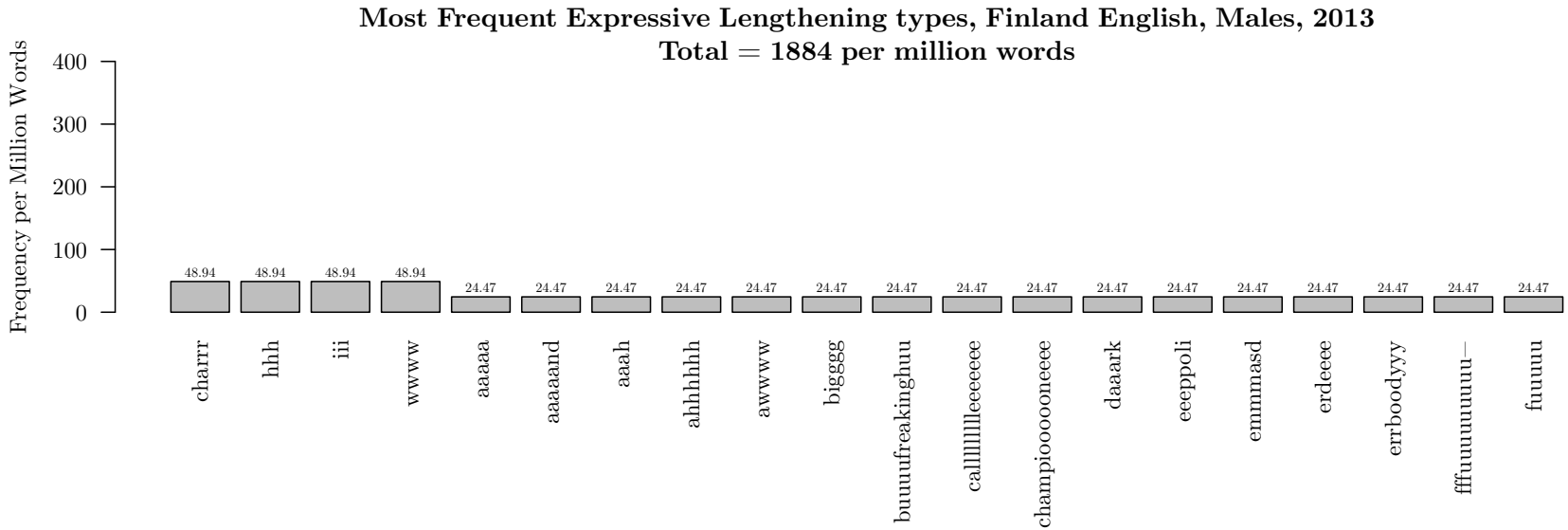


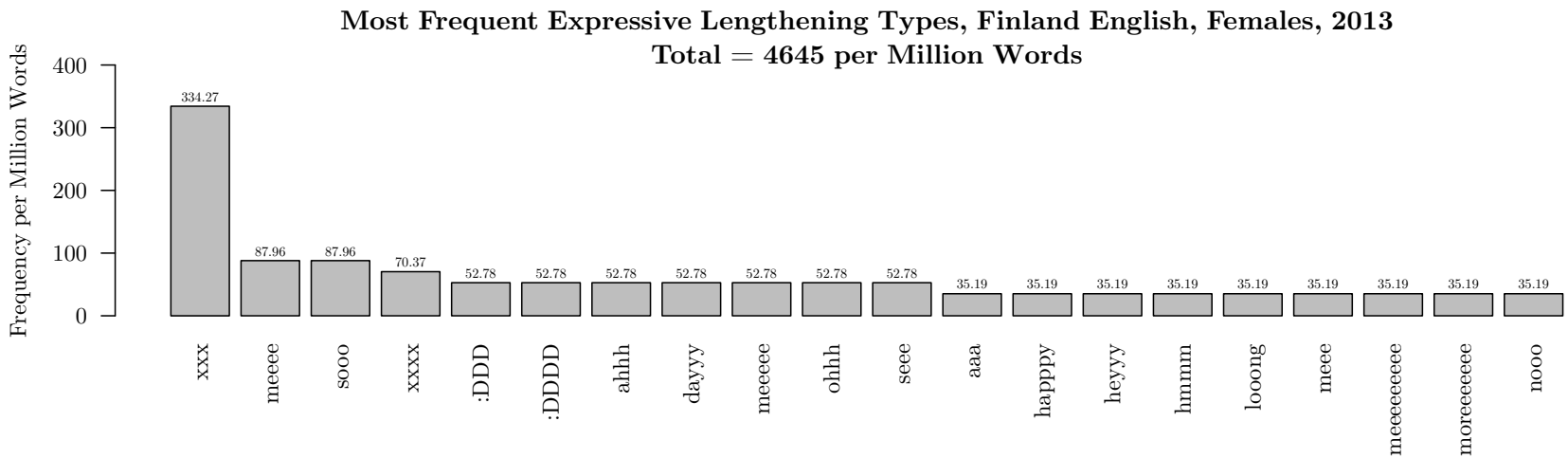
Expressive Lengthening by Gender, Global 2009



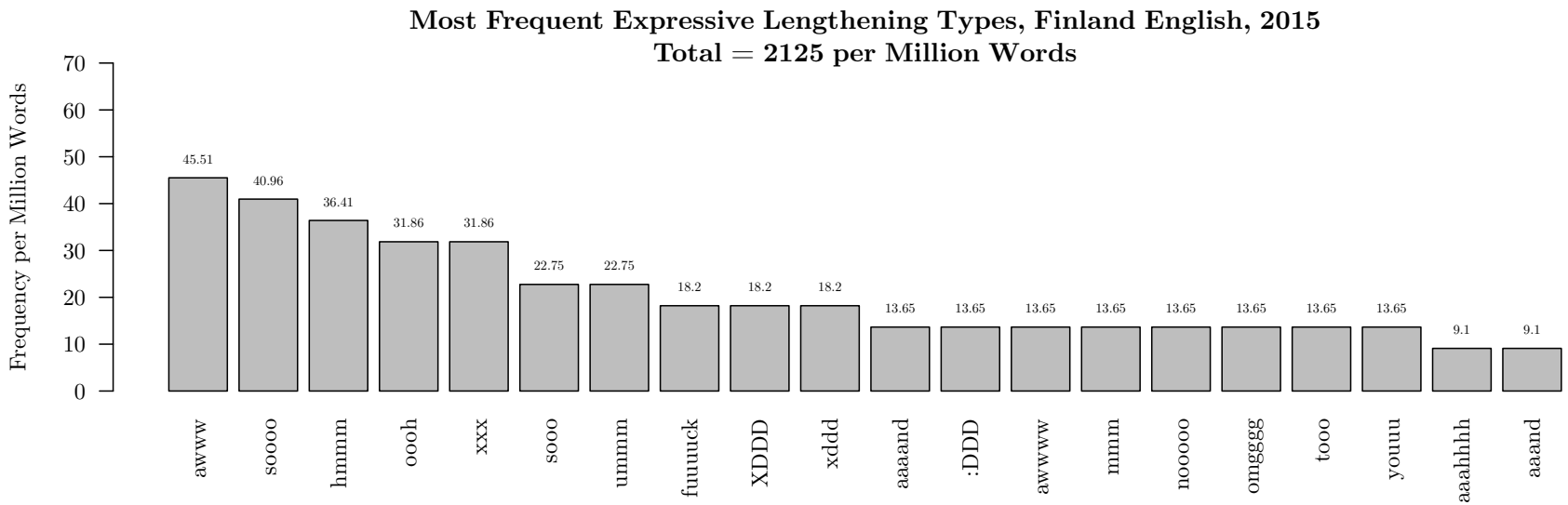


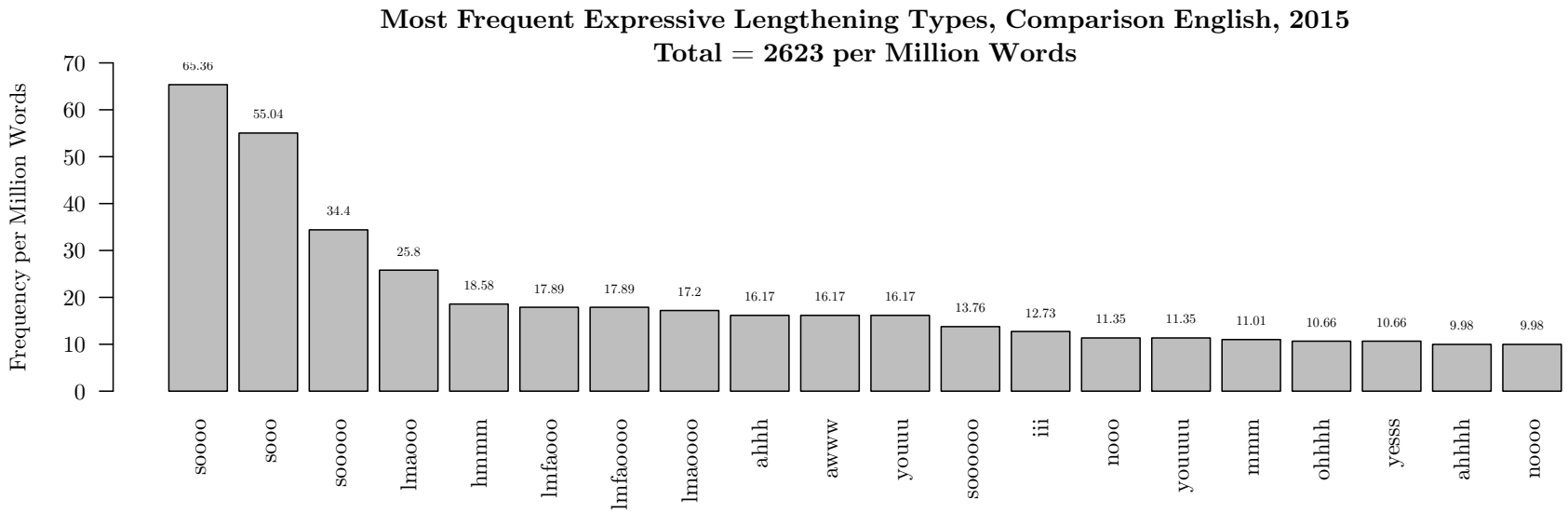
Expressive Lengthening by Gender, Finland, 2013



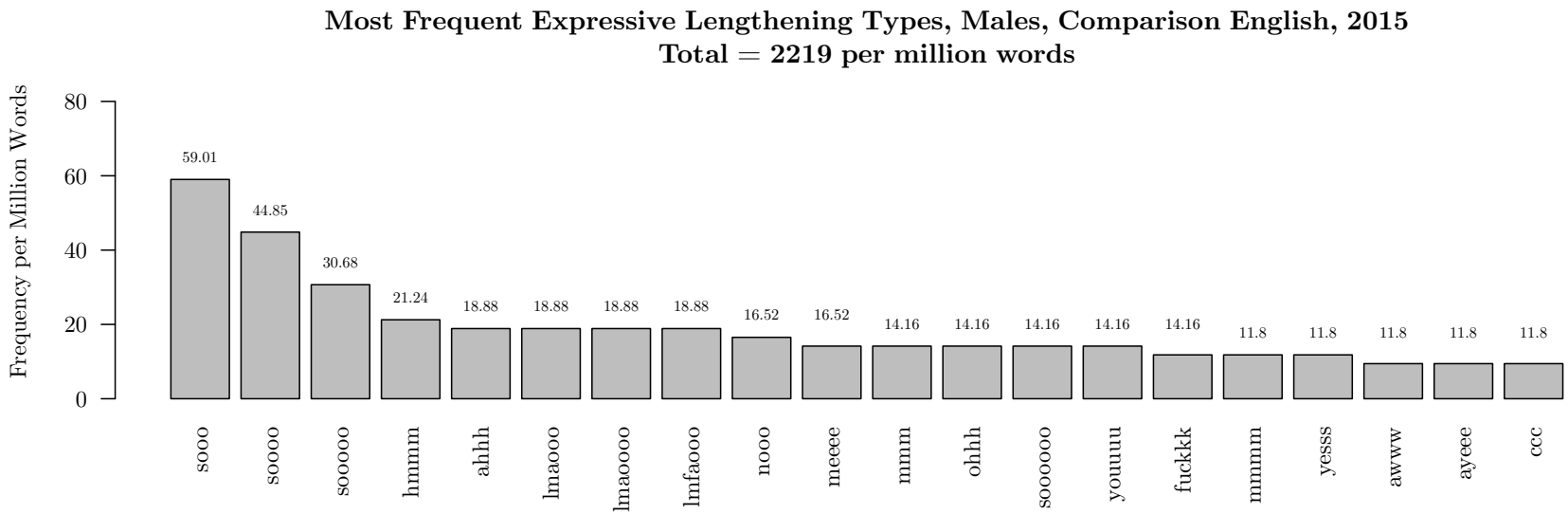


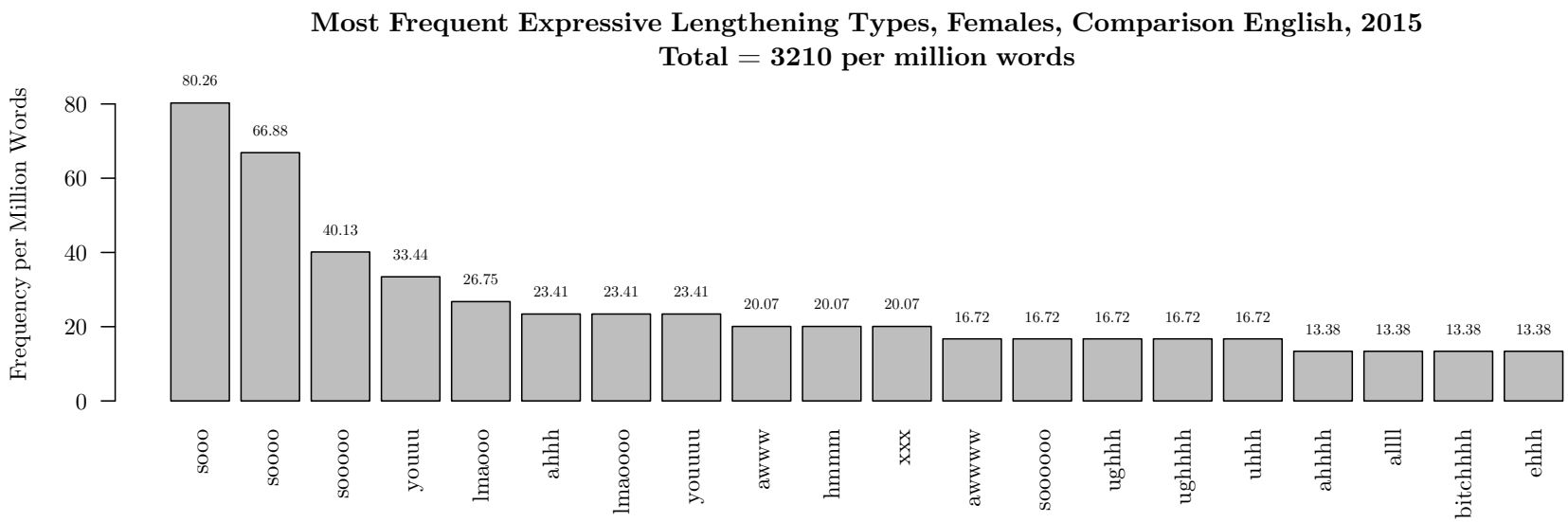
Expressive Lengthening by Geography, 2015



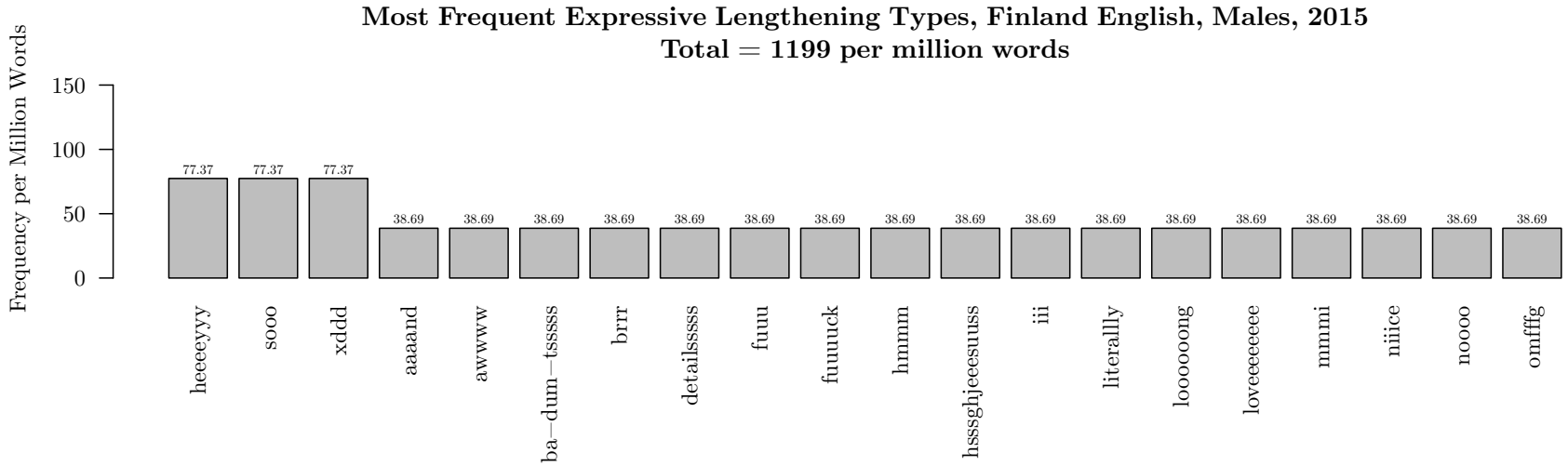


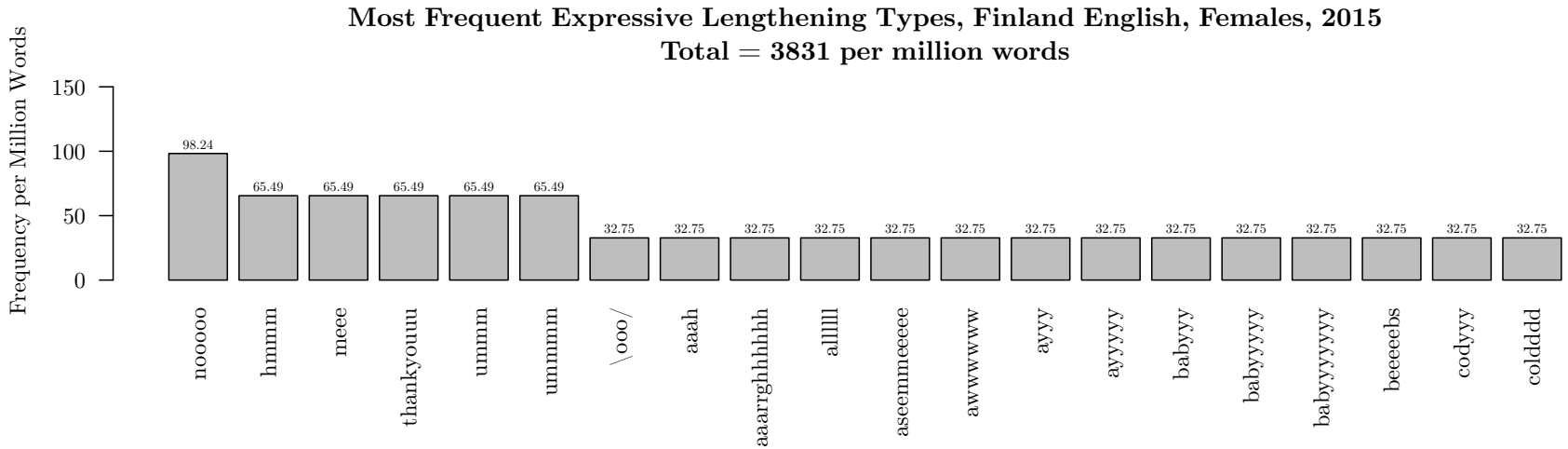
Expressive Lengthening by Gender, Global 2015





Expressive Lengthening by Gender, Finland 2015





Non-standard Lexical Features: Emoticons

- *Emoticon*: sequence of text characters (typically ASCII) that can represent facial expression or gesture
- Affect indicators that convey contextual information corresponding to spoken-language cues such as prosodic, stress, and intonation features (Herring 1999, 2013; Ptaszynski 2007):
  :) :-) ;D XD :(:-(:^/ ;_; :| :'| o.O -__-
- Discourse organizers with little fixed affective content, used to position audiences around propositions (Schnoebelen 2012, Vandergriff 2014)
- In general, emoticons are more frequently used by females in CMC and on Twitter (Baron 2004, Bamann et al. 2014), but emoticons associated with sarcasm, flirting behavior, or negative affect may be used more by males (Wolf 2000, Herring 2013).

Non-standard lexical Features: Emoticons

- PoS tags used to identify all emoticon tokens, regex used to filter out interjection word forms, 218 most frequent emoticon types considered
- Tweets from Finland and Finland English tweets **more** likely to have an emoticon; Comparison/US tweets and Comparison/US English messages **less** likely to have an emoticon.

Corpus	Emoticons/1m Tok
Comparison English 2009	4533
Finland English 2013	18616
US English 2015	1441
Finland English 2015	3463

Non-standard lexical Features: Emoticons

- Tweets by females **more** likely to have an emoticon; Male/female discrepancy more pronounced for Finland English.

Corpus	Emoticons/1m Tok
Comparison English 2009 Male	4206
Comparison English 2009 Female	6907
Finland English 2013 Male	11916
Finland English 2013 Female	24243
US English 2015 Male	1317
US English 2015 Female	1816
Finland English 2015 Male	3404
Finland English 2015 Female	8907

Lexical Features Global: Most Male and Female Emoticons, Global, 2009

- Males: 6 Japanese-style emoticons, 9 non-positive emotional affect emoticons
- Females: 1 Japanese-style emoticon, 12 non-positive emotional affect emoticons

	type	θ	type	θ
1	):	3.37	(;	-4.80
2	^^	2.33	=)	-4.59
3	xD	1.88	:-D	-4.56
4	:o	1.87	(:	-3.48
5	O.o	1.87	=(	-3.05
6	:-P	1.66	=]	-2.91
7	o-o	1.39	:-)))))	-2.91
8	:-))	1.02	=D	-2.70
9	:]	1.02	:-))))	-2.64
10	o_O	1.02	=[	-2.47
11	;(	1.02	]:	-2.27
12	D:	1.02	:-)))	-2.27
13	O_o	1.02	=/	-2.02
14	:-\	1.02	:O	-1.68
15	^__^	1.02	;o	-1.68
16	;)	1.01	;P	-1.54
17	:P	0.80	:((	-1.17
18	:)	0.64	:'(	-1.17
19	:|	0.44	;))	-1.17
20	:D	0.39	>_<	-1.17

Lexical Features Global: Most Male and Female Emoticons, Finland, 2013

- Males: 7 Japanese-style emoticons, 8 non-positive emotional affect emoticons
- Females: 4 Japanese-style emoticon, 10 non-positive emotional affect emoticons

	type	θ	type	θ
1	=)	3.76	(;	-2.87
2	;3	3.62	:p	-1.67
3	=))))	3.24	:o	-1.36
4	O_o	3.24	(:	-1.21
5	;_;	2.99	:-(	-0.91
6	^_^	2.65	:((	-0.91
7	=)))	2.65	:)))	-0.91
8	x_x	2.15	</3	-0.91
9	XDDD	2.14	:|	-0.91
10	=DD	2.14	:DDD	-0.91
11	^_____^	2.14	:DDDD	-0.91
12	=)))))	2.14	<3	-0.81
13	=/	2.14	:(	-0.68
14	;(((	2.14	:D	-0.67
15	).:	2.14	;__;	-0.57
16	]:	2.08	XDD	-0.57
17	:3	1.90	;(	-0.57
18	=D	1.89	-__-	-0.57
19	^^	1.75	-_-	-0.57
20	0_0	1.63	-,-	-0.57

Lexical Features Global: Most Male and Female Emoticons, Global, 2015

- Males: 5 Japanese-style emoticons, 7 non-positive emotional affect emoticons
- Females: 4 Japanese-style emoticon, 12 non-positive emotional affect emoticons

	type	θ	type	θ
1	(=	2.38	o.O	-1.64
2	-.-	2.03	:////	-1.64
3	:D	1.86	:')	-1.57
4	(':	1.59	:'(	-1.13
5	^.^	1.59	((:	-1.13
6	D:	1.59	;-;	-1.13
7	D;	1.59	;/	-1.13
8	:o	1.59	=(	-1.13
9	:3	1.45	:((((	-1.13
10	:P	1.28	^^	-1.13
11	:))))	1.09	</3	-1.13
12	-_-	1.08	///	-1.13
13	xD	1.08	-,-	-1.13
14	:p	1.08	;))))	-1.13
15	^_^	1.07	:DD	-1.13
16	<333	1.07	:|	-1.13
17	:]	1.07	>:(	-1.13
18	>.<	1.07	;_;	-1.13
19	]:	1.07	O.o	-1.13
20	o:	1.07	;))	-1.13

Lexical Features Global: Most Male and Female Emoticons, Finland, 2015

- Males: 5 Japanese-style emoticons, 5 non-positive emotional affect emoticons
- Females: 3 Japanese-style emoticon, 7 non-positive emotional affect emoticons

	type	θ	type	θ
1	;_;	3.11	:')	-2.77
2	XD	2.95	:p	-2.61
3	;))	2.76	:-D	-1.74
4	>_<	2.23	</3	-1.46
5	^^	2.23	:O	-1.46
6	._.	2.23	;p	-1.29
7	:333	2.23	^_^	-0.83
8	(=	2.23	):	-0.83
9	XDDD	2.23	(:	-0.83
10	xddd	2.23	D:	-0.83
11	:P	1.14	-_-	-0.49
12	:-(	1.13	:DD	-0.49
13	;D	1.13	:((	-0.49
14	:))	1.12	;3	-0.49
15	<333	1.12	o.O	-0.49
16	\o/	1.12	:D	-0.45
17	;P	1.12	<3	-0.32
18	xD	1.12	:3	-0.11
19	:o	1.12	;-)	-0.10
20	;)))	1.12	<33	0.02

ULEÅBORGS UNIVERSITET
УЛЕАБОРГСКИЙ УНИВЕРСИТЕТ

Table of Contents

1 Introduction: Twitter English and Finland Twitter English

2 Data Collection and Processing

- Collecting and Processing the Data
- Language Detection
- Geographical Distribution of Language
- Automatic PoS Tagging
- Gender Disambiguation

3 Selected Lexical Features

- Lexical Features

4 Conclusion and Summary

Summary

- Finland Twitter English makes slightly less use (at present) of expressive lengthening than global/US English
- Females use expressive lengthening more than males, especially in Finland Twitter English
- Expressive lengthening letter choice may be influenced by L1 phonological considerations
- Overall, expressive lengthening on Twitter may be decreasing in Finland, but increasing elsewhere/in the US
- Finland Twitter English makes more use of emoticons
- Females use emoticons more than males, especially in Finland Twitter English
- Overall, emoticon use on Twitter seems to be decreasing somewhat over time (emojis?)

Summary

- Lexical frequencies suggest that Finland Twitter English, especially that of females, may be more **interactive** and less **informational** than English on Twitter overall
- Authors use non-standard lexical features to construct and negotiate meanings at the interface of online interactivity and technological change (Hutchby 2001, Wikström 2014).
- Larger corpora are needed!

References I

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

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

Davis, M. and P. Edberg. (2015). *Unicode Emoji* (Technical report UTR No. 51). Unicode Consortium. http://unicode.org/reports/tr51/#Emoticons. (Accessed 15 October 2015)

Eisenstein, J., B. O'Connor, N. A. Smith and E. P. Xing. (2012). "Mapping the geographical diffusion of new words". *Computing Research Repository*. http://arxiv.org/abs/1210.5268. (Accessed 15 October 2015)

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*, 42–47. Stroudsburg, PA: ACM. www.ark.cs.cmu.edu/TweetNLP/gimpel+etal.acl11.pdf. (accessed 15 October 2015)

References II

Gimpel, K. N. Schneider and B. O'Connor. (2013). "Annotation Guidelines for Twitter Part-of-Speech Tagging Version 0.3". Computational Science Department, Carnegie Mellon University. http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf. (accessed 15 October 2015)

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

Leetaru, K. H., S. Wang, G. Cao, A. Padmanabhan, and E. Shook. (2013). "Mapping the global Twitter heartbeat: The geography of Twitter". *First Monday* 18(5/6). http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654. (accessed 15 October 2015)

Lui, M. and T. Baldwin. "Langid.py: An off-the-shelf language identification tool". *Proceedings of 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 25–30. Stroudsburg, PA: ACM. http://www.aclweb.org/anthology/P12-3005. (accessed 15 October 2015)

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

References III

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 8.4*.

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*, 380-390. Stroudsburg, PA: ACM. http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.naacl13.pdf. (accessed 15 October 2015)

Page, R. (2012). "The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags". *Discourse and Communication* 6(2), 181–201.

Squires, L. (2012). "Whos punctuating what? Sociolinguistic variation in instant messaging". In: A. Jaffe, J. Androutsopoulos, M. Sebba & S. Johnson (eds.), *Orthography as Social Action: Scripts, Spelling, Identity and Power*, 289–324. Berlin: De Gruyter.

Wikström, P. (2014). "#srynotfunny: Communicative functions of hashtags on Twitter". *SKY Journal of Linguistics* 27, 127–152.

References IV


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


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.


Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. Thesis, University of Stuttgart.


Hutchby, I. (2001). *Conversation and Technology*. Cambridge, UK: Polity.


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