Non-standard lexical and grammatical resources in Finland Twitter English

Steven Coats

English Philology, University of Oulu, Finland

19 September 2015
Poznań Linguistic Meeting

OULUN YLIOPISTO
UNIVERSITY of OULU

Outline

OULUN YLIOPISTO
UNIVERSITY of OULU

Contexts of the Present Research

- Increasing prevalence of Computer–Mediated Communication such as Social Media/Twitter

  > "Almost all humans today live in a textually mediated world, and the texts which mediate and impact on our lives are by no means all fixed in [physical] space" (Sebba 2010: 61)

- Categorization of discourse, language genres or varieties based on the principal communicative functions exemplified by configurations of linguistic features (Biber 1985, 1986, 1987, 1988, 1995, 2006; Biber and Conrad 2009)
- "Global Englishes" and the status of English in (traditionally) non–Anglophone societies (Kachru 1990)
- English as it is used on Twitter in Finland: Comparison of the frequencies of lexical and grammatical features of *Finland Twitter English* with a corpus of English Twitter messages with no specification of geographical provenance
- Non–standard features: Expressive lengthening, emoticons, and Twitter–specific grammatical types

Approaches to Twitter Language

- Hashtag functionality (Zappavinga 2011, Wikström 2014)
- Geographical distribution of lexical items (Eisenstein et al. 2012)
- Sociolinguistics of (American) English Tweets (Bamann et al. 2014)
- Regional variation of English use on Twitter internationally?

Characterization of Discourse by Lexical and Grammatical Frequencies

- Aggregate grammatical feature frequencies can tell us something about the nature of the discourse of a genre and can be used to distinguish varieties (Biber 1988, 1995, 2006).
- I compare aggregate grammatical feature frequencies of English-language Twitter discourse in Finland to English–language Twitter discourse overall and interpret the difference in terms of communicative function.
- Non–standard lexical and grammatical features are characteristic for Twitter and other CMC genres

OULUN YLIOPISTO
UNIVERSITY of OULU

Finland and Comparison Corpora Data Collection

- Finland data: Twitter Streaming API
  - Access levels
  - Extent of geo–encoded user messages (1.6% of tweets according to Leetahu et al. 2013)
  - Geo–coordinates bounding box
  - Filtering using data from GADM and packages in *R* (maptools, mapdata)

- Comparison data collected 2009 at Texas A&M Univ. from Twitter Streaming API with no geo–coordinates

OULUN YLIOPISTO
UNIVERSITY of OULU

Language Detection

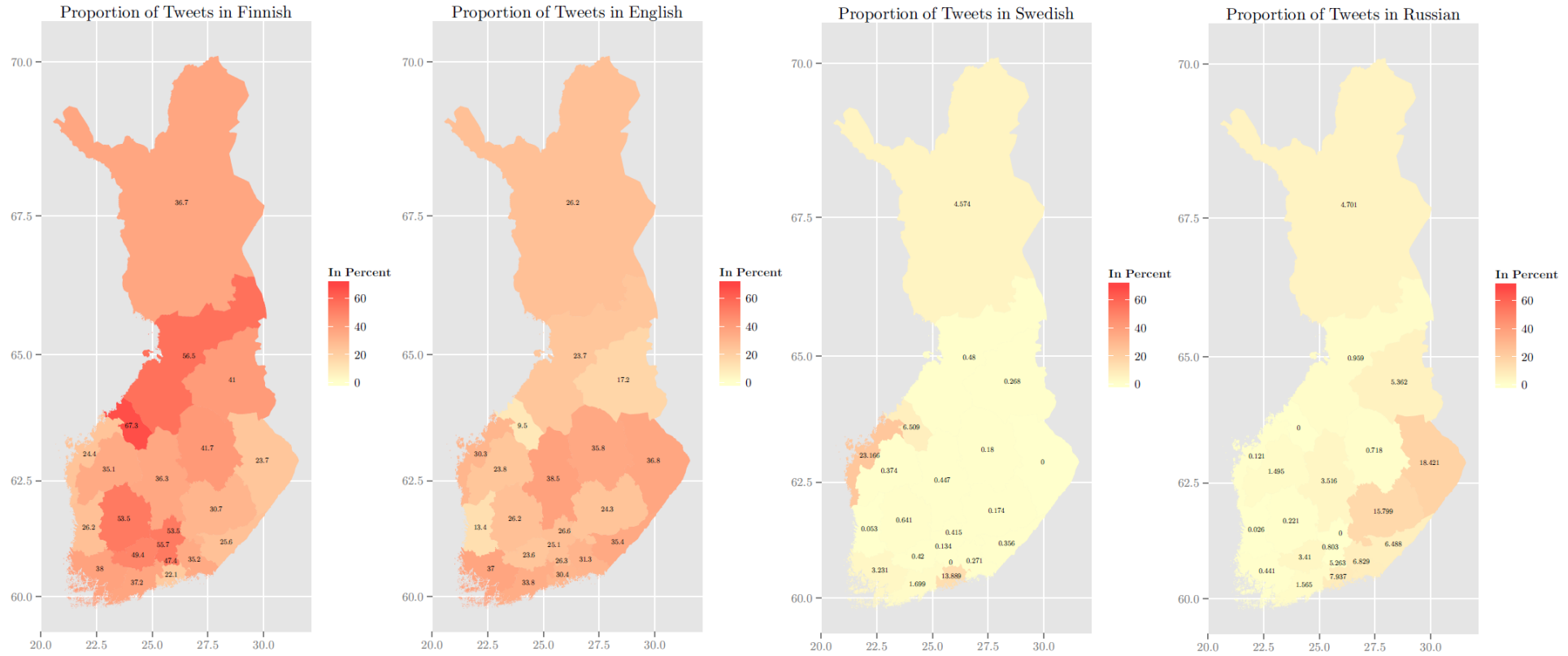
- Twitter provides a field in Tweet entity indicating language since mid–2013; prior to this have to detect language using own tools
- Automatic language disambiguation using langid.py (Lui and Baldwin 2012)

	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.

OULUN YLIOPISTO
UNIVERSITY of OULU

Percent of Tweets by Language and by Finnish Province

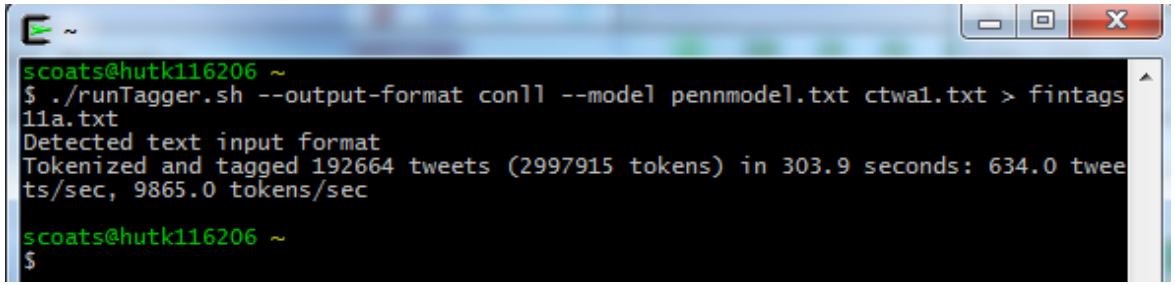


Languages of Finland Tweets: 44.8% Finnish, 35.7% English, 2.2% Swedish 2.1% Russian, 15.2% Other

OULUN YLIOPISTO
UNIVERSITY of OULU

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)
- Penn Treebank tags (Marcus et al. 1993)
- Additional tags for Twitter–specific types (retweet, username, hashtag)
- Output consists of tab–separated token/tag/prob tokens, including e.g. emoticons




```
1616  20→CD→0,9908
1617  degrees→NNS→0,9570
1618  outside→IN→0,9401
1619  :D→UH→0,8939
1620  now→RB→0,9528
1621  let's→VBZ→0,2536
1622  see→VB→0,9863
1623  what→WP→0,9712
1624  it's→PRP→0,7312
1625  like→IN→0,7036
1626  when→WRB→0,9884
1627  the→DT→0,9846
1628  cold→NN→0,6367
1629  weather→NN→0,9907
1630  arrives→VBZ→0,9720
```

Corpora Summary Statistics

Corpus	Tweets	Tokens	Types	TTR
Finland	125,117	1,225,215	294,966	4.15
Finland English	32,956	389,189	64,424	7.98
Comparison	305,310	3,358,788	421,147	6.04
Comparison English	197,310	2,677,580	236,082	11.34

Quantifying Similarity

- A relative frequency statistic was calculated for all lexical and PoS types in the two principal corpora (Evert 2004).

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

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

- Types with most extreme values represent the lexemes or PoS most distinctive for the discourse of each corpus.

OULUN YLIOPISTO
UNIVERSITY of OULU

Grammatical Features: Parts of Speech

- Of 37 grammatical tags assigned by the CMU Twitter Tagger, ten are used substantially more often in the Finland English Corpus and eleven more often in the Comparison English Corpus ($|\theta| \geq 0.182$).

Feature	odds ratio θ	Feature	odds ratio θ
Hashtag	3.360	Particle	-0.562
Retweet	1.678	Other punctuation (: ; ... + - = <> [ ])	-0.494
Username (preceded by @)	0.770	Verb, gerund or present participle	-0.494
Interjection	0.751	Verb, past participle	-0.385
Personal pronoun	0.350	Wh-determiner	-0.314
Wh-adverb	0.350	*to*	-0.301
Verb, non-3rd person singular present	0.336	Proper noun, singular	-0.301
Universal Resource Locator	0.223	Determiner	-0.235
Adjective, superlative	0.215	Noun, singular or mass	-0.235
Coordinating conjunction	0.207	Period (. *?* *!*)	-0.210
		Verb, 3rd person singular present	-0.198

- Most "Finnish": Non-standard Twitter/CMC–specific types, personal pronouns, 1.p. and 2.p. verb forms. Least "Finnish": Phrasal particles, punctuation, NP elements (nouns, determiners), punctuation, compound VP elements (gerunds, participles, *to*).

Grammatical Feature Findings

- Grammatical feature (part–of–speech) frequencies suggest that overall, Finland Twitter English is more **interactive** and less **informational** than English on Twitter overall
- Grammatical feature frequencies in Finland Twitter English tend to index Twitter/CMC–specific grammatical functions (i.e. index technological interaction)

Lexical Features: Most and Least "Finnish" Types

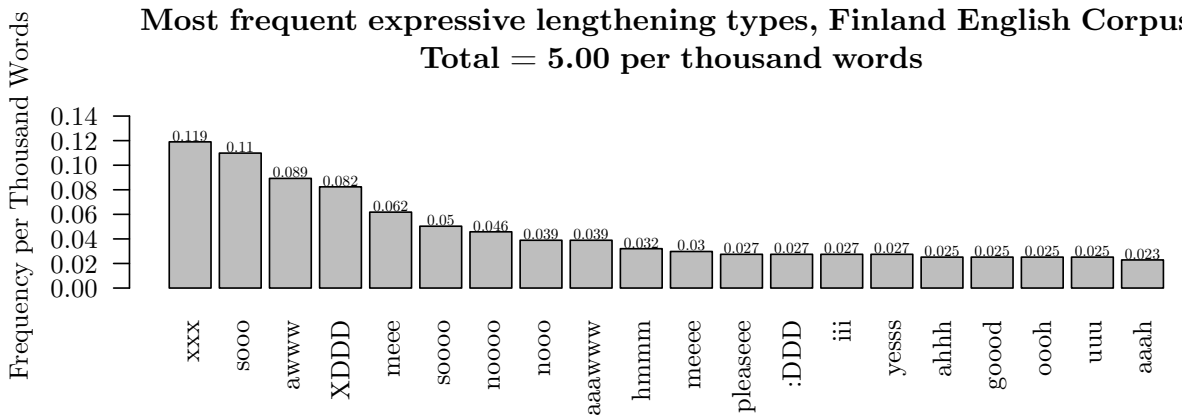
- Lexical type frequencies reflect discourse–related topicality specific to the varieties
- Some highly frequent types reflect the temporal and geographical parameters of the data collection

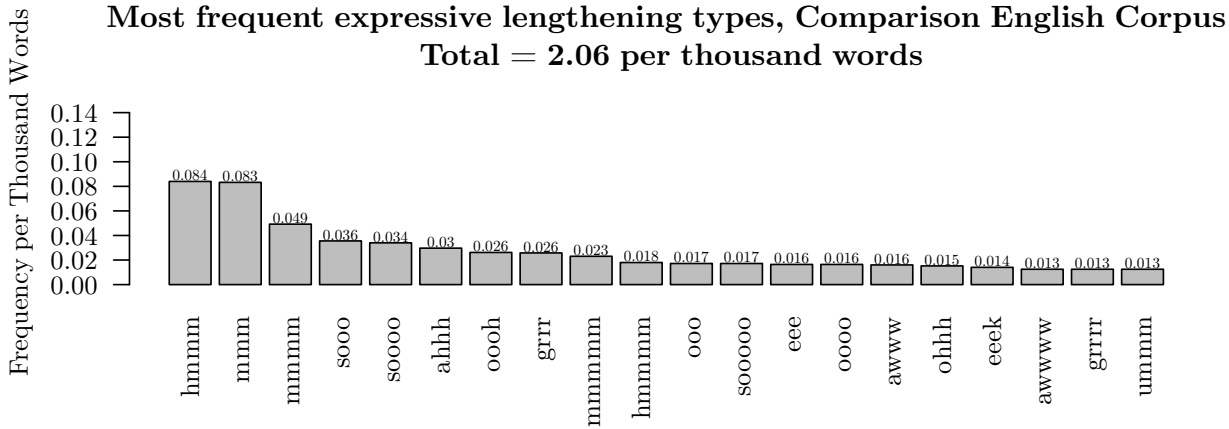
	Word	odds ratio θ	Word	odds ratio θ
1	finland	7.870	steelers	-4.976
2	❤	6.978	m4	-3.985
3	niall	6.124	scout	-3.953
4	helsinki	5.989	»	-3.920
5	finnish	5.590	flickr	-3.829
6	xx	5.377	herbal	-3.776
7	sweden	5.204	obama	-3.693
8	hel	5.084	palin	-3.671
9	#party	4.673	a5	-3.493
10	#food	4.673	xbox	-3.455
11	:DD	4.505	nfl	-3.417
12	ikr	4.495	reader	-3.361
13	apparatus	4.495	firefox	-3.332
14	justin	4.493	ebay	-3.310
15	rn	4.467	twittering	-3.188
16	;))	4.445	entry	-3.063
17	gaga	4.392	blog	-3.029
18	casually	4.392	blogging	-2.991
19	youu	4.392	site	-2.957
20	<3<3	4.392	lane	-2.931

OULUN YLIOPISTO
UNIVERSITY of OULU

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





OULUN YLIOPISTO
UNIVERSITY of OULU

Non-standard Lexical Features: Emoticons and Emojis

- *Emoticon*: series of text characters (typically punctuation or symbols) meant to represent a facial expression or gesture

:) ;-DD 0_0

- *Emoji*: A pictograph that can be used inline in text, internally represented as either an image or an encoded character (Davis and Edberg 2015)

OULUN YLIOPISTO
UNIVERSITY of OULU

Non-standard lexical Features: Emoticons

- PoS tags used to identify all emoticon tokens; Regex used to filter out interjection word forms; 240 most frequent emoticon types considered
- Tweets from Finland and Finland English tweets **more** likely to have an emoticon; Comparison tweets and Comparison English messages **less** likely to have an emoticon.

	Type	Percent		Type	Percent
1	:)	26.5	11	:P	1.0
2	<3	11.0	12	;))	1.0
3	:D	10.6	13	(:	0.9
4	;)	7.8	14	:))	0.9
5	:(	6.0	15	:'(	0.6
6	:-)	4.4	16	;-)	0.6
7	:3	2.3	17	:')	0.6
8	XD	1.5	18	(;	0.6
9	^^	1.3	19	:-D	0.5
10	xD	1.1	20	:o	0.5

Corpus	% messages with emoticons	emoticons per 1000 tokens
Finland	19.0	21.7
Finland English	22.1	23.2
Comparison	9.7	9.0
Comparison English	6.7	5.6

OULUN YLIOPISTO
UNIVERSITY of OULU

Summary

- Finland Twitter English emerges as a distinct CMC English variety on the basis of aggregate grammatical feature frequencies.
- Non–standard features such as Twitter–specific grammatical tags, expressive lengthening and emoticon use are characteristic of Finland Twitter English and serve to demonstrate the primarily interactive communicative function of its discourse for much of the userbase.
- Feature frequencies, particularly for items associated strongly with CMC and the Twitter platform, may index the adoption of technological innovation such as smartphone use for text–based communication.
- Finland Twitter English users use these features to construct and negotiate meanings at the interface of online interactivity and technological change (Hutchby 2001, Wikström 2014).

OULUN YLIOPISTO
UNIVERSITY of OULU

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.


Biber, D. (1985). “Investigating macroscopic textual variation through multi-feature/multidimensional analyses”. *Linguistics* 23: 337–360.


Biber, D. (1986). “Spoken and written textual dimensions in English: Resolving the contradictory findings”. *Language* 62: 384–414.


Biber, D. (1987). “A textual comparison of British and American writing”. *American Speech* 62: 99–119.


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


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.

OULUN YLIOPISTO
UNIVERSITY of OULU

References II


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*. Stroudsburg, PA: ACM, pp. 42–47.


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.


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).


Lui, M. and T. Baldwin. "Langid.py: An off-the-shelf language identification tool". *50th Proceedings of the Association for Computational Linguistics*. Stroudsburg, PA: ACM, pp. 25–30.


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

OULUN YLIOPISTO
UNIVERSITY of OULU

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 NAACL-HLT*, pp. 380-390.


Sebba, M. (2010). "Discourses in transit". In: A. Jaworski and C. Thurlow (eds.), *Semiotic Landscapes: Language, Image, Space*. London: Continuum, pp. 5976.


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


Zappavinga, M. (2011). "Ambient affiliation: A linguistic perspective on Twitter". *New Media and Society*, 13(5): 788–806.


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.

OULUN YLIOPISTO
UNIVERSITY of OULU

References IV

- 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.

OULUN YLIOPISTO
UNIVERSITY of OULU