Steven Coats

Grammatical feature frequencies of English on Twitter in Finland

1 Introduction

Technological change affects the parameters of language use, and as internet access has expanded rapidly in recent decades, communicative encounters resulting from online activity have begun to play an increasing role in daily life. Commercial social media platforms such as Twitter, whose content consists of millions of user messages with global extent, represent an important site of online language use. The use of English online has been subject to much attention in public discourse in mass media as well as in academic scholarship, and while research into online language has addressed a wide range of topical considerations, a recurrent typological interpretation of English as it is used in computer–mediated communication (CMC) is that it differs from traditional language varieties in terms of lexis, grammar, and pragmatic features. Crystal (2006: 18) uses the term *Netspeak* to refer to "a type of language displaying features that are unique to the Internet… arising out of its character as a medium which is electronic, global and interactive" (cf. Androutsopoulos 2006).[1]

At the same time, the global status of English as the world's lingua franca continues to evolve, with English now serving not only as the principal language of international communication in academia, business, media and diplomacy (Crystal 2003), but increasingly as an important language for online communication in informal and geographically localized communicative contexts, particularly in the European Union (European Commission 2011).

Despite widespread recognition of the prevalence of English in global CMC, there have been relatively few efforts to systematically investigate variation in the lexis or grammar of English on social media in international contexts.[2] Although studies have investigated the use of particular linguistic features in

1 However, the uniqueness of CMC/Netspeak in this respect has been disputed by e.g., Squires (2010).

2 Mocanu et al. (2013) provide a survey of language and geography for global Twitter data. Magdy et al. (2014) find that English is the predominant language on Twitter for 41 of 206 countries or territories.

Steven Coats, University of Oulu

various types of CMC, including Twitter, and the distribution of linguistic features in Twitter language in the United States (e.g., Eisenstein et al. 2014 and Grieve et al. 2015 for lexical innovation and the geographical distribution of lexical items; Bamann, Eisenstein and Schnoebelen 2014 for frequencies of selected lexical and grammatical items), to my knowledge, no corpus-based studies have analyzed Twitter English in non-L1 environments.

The present research characterizes the social media language variety "English on Twitter in Finland" as a Finland-based variety of online English that emerges as distinct when investigated on the basis of aggregate feature frequencies. The approach is not focused on close qualitative analysis of the discourse functionality of individual tweets or the linguistic behavior of individual Twitter users. Rather, it utilizes Exploratory Factor Analysis of aggregate feature frequencies as a means of identifying patterns of co-occurrence and distinguishing underlying functional or situational parameters that may contribute to register, genre or variety difference (Biber 1988, 1995, 2006). Grammatical features with information-marking functions, such as noun phrases, adjectives, complex verb forms, or numerals, tend to co-occur in many types of texts, as do features which may be utilized to negotiate interpersonal interaction, such as first- and second-person pronouns, present-tense verb forms or modal verbs. Factor analysis identifies communicative dimensions of discourse, such as an "Informational versus Involved Production" dimension that contrasts information-marking features with interactive features in some texts (Biber 1988: 107).

This study considers Twitter-specific discourse features (usernames, hashtags, and retweets, i.e., re-broadcastings of tweets by different users) and part-of-speech features, as determined by the output of a probabilistic tagger, in English tweets. In addition, tweet length is examined, as well as two features associated with many types of CMC: emoticons and the non-standard orthographical feature expressive lengthening. The findings are interpreted on the basis of a factor analysis as reflecting underlying communicative dimensions.

The dynamics of English use on social media such as Twitter may differ between traditional L1 English societies and societies in which English is not an official language ("core" and "expanding circle" contexts, in the terminology of Kachru 1990). To that extent, an analysis of English on Twitter in Finland may contribute to our understanding of the ways in which language interacts with the complex forces of globalization. Upon closer examination, the configuration of language feature frequencies most typical of the variety suggests that a characteristic communicative orientation may underlie the interactions of Finland-based users writing in English, who extract meaning-creating potential

from new technology at the interface of user functionality and medium constraints (cf. Hutchby 2001; Wikström 2014).[3]

2 Background

Twitter has become an important resource for online communication since its launch in 2006. Twitter platform users post public tweets of up to 140 characters and use the service to interact with other users by following or responding to their tweets and providing links to other online information. As of September 2015, the site reported more than 320 million active users monthly, 79% of whom were located outside the Unites States (Twitter 2015). In addition to the text (the "user message" field), the data structure of a tweet can contain additional metadata entities with information about language, location of the author, interaction of the author with other users, or other types of information. While tweets broadcast from desktop or laptop computers typically do not contain location information, a small but nonetheless substantial proportion of messages contain geo-coordinates corresponding to the location of the user (Morstatter et al. 2013 report 1.45–3.17% of tweets are geotagged; Leetaru et al. 2013 report 1.6%).

Twitter users employ platform-specific linguistic resources to interact with other users and situate their own messages within specific contexts of discourse on the platform. One such resource, the utilization of usernames or screen names with an affixed <@> symbol, is often used for direct exchanges between users. Honeycutt and Herring (2009) analyze a corpus of tweets in order to investigate the extent to which Twitter users engage in direct user-to-user exchanges. They find that the presence of <@> in tweet messages correlates with user interactivity, and suggest that microblogging may facilitate collaboration.

Dialogic participation and patterns of user interaction have also been the focus of studies by Ritter, Cherry and Dolan (2010) and Page (2012), who note that a significant proportion of tweets containing the <@> symbol consists of broadcast-style content with no explicit response to the tweet author from other users. Dialogues on Twitter (i.e., multiple-tweet content consisting of a message and at least one response directed to the message's author) tend to be short:

3 An example of meaning creation via unexpected user interaction with technology might be the development of the Short Message Service (mobile telephone texting) from a means for the automatic transmission of emergency broadcasts to an interpersonal communication resource (Hillebrand et al. 2010).

70% of conversations consist of one tweet and one response (Ritter, Cherry and Dolan 2010: 173).

Other features characteristic of Twitter user messages, such as the use of the hashtag (<#>), have been investigated. Zappavigna (2011, 2012) suggests that hashtags, originally employed on the Twitter platform as explicit topic or content markers, have taken on pragmatic functions. Due to the ways in which the Twitter interface allows users to search character strings preceded by the hashtag and interact with users utilizing specific hashtags, the symbol is now frequently used to show evaluative sentiment or broadcast subjective affiliation. Wikström (2014) analyzes several communicative tasks associated with the hashtag on Twitter, noting that in addition to marking topics and conversations, the symbol is used to participate in online communal games, mark metacommentary, or negotiate pragmatic categories such as self-presentation and maintenance of face. To that extent, hashtag use represents an example of how user interaction with communication technology interfaces can prompt the emergence of unexpected communicative behaviors (Wikström 2014: 148–150).[4]

Emoticons are "visual cues formed from ordinary typographical symbols that ... represent feeling or emotions" (Walther and D'Addario 2001, citing Rezabek and Cochenour 1998: 207; see also Dresner and Herring 2010 and Vandergriff 2014a). Non-standard features such as emoticons have not figured as prominently in corpus-based studies of language as have other units such as dictionary words or grammatical types. The relative lack of attention paid to the prevalence and communicative function of emoticons may reflect the somewhat restricted domains of use of these symbols, which are more frequently encountered in CMC text types such as chat, instant messaging, online message boards, or the anonymous imageboards known as "chans," but less frequently in blogs and the online equivalents of print media like news reports or academic writing (Ptaszynski et al. 2011).

Schnoebelen (2012) investigates the expression of affective content on Twitter, particularly through the use of emoticons and their co-occurrence with lexical items. He suggests that emoticons have broader discourse functionality than simply the representation of emotional states, and finds that on Twitter, use of particular emoticon types correlates with word choice.

Non-standard orthography, whether the result of error or used as an expressive resource, is another feature prevalent in CMC genres such as chat or Instant Messenger communication (Herring 2001; Paolillo 2001; Tagliamonte and Denis

4 For a discussion of the discourse features of Twitter, see further Squires (2016), Zappavigna (2011, 2012), and Page (2012).

2008) as well as on Twitter. In the linguistics literature, orthography has traditionally been considered from the perspective of the correspondence between characters and speech sounds, although more recent research has proposed a functional interpretation of orthographic variation (Sebba 2007). For Twitter, Callier (this volume) considers non-standard orthography corresponding to the fortition of dental fricatives in Twitter English as a style marker. Some research has examined *expressive lengthening*: non-standard orthography in which individual characters in a word string are repeated (e.g., *cooooooool, yesssssss, dumbbbb*). The feature has been interpreted primarily as an affective discourse marker (Rao et al. 2010; Bamann, Eisenstein and Schnoebelen 2014).

While the varied functionality of Twitter-specific discourse features and the use of non-standard items such as emoticons or expressive lengthening on the platform have been investigated in general, there have been relatively few studies of Twitter English in specific geographical contexts, perhaps due to the relatively small proportion of tweets that are geotagged. Some research has investigated aspects of regional differentiation of Twitter English within the United States. Alis and Lim (2013), for example, analyze the length (in characters) of geo-encoded user messages, and find that overall, tweet length in the US decreased slightly from 2009–2012. They regress tweet length with a number of demographic variables, and find the strongest correlation to be an inverse relationship between tweet length and proportion of African-American inhabitants for US states. Although this demographic parameter may not be relevant for an analysis of English-language Twitter in Finland, correlation of language variation and sociolinguistic identity may shed light on English on Twitter in Finland as well.

Eisenstein et al. (2014) explore the emergence of Twitter dialects, or geographically localized uses of particular word forms in the United States, by using location, population, and demographic identity as parameters in a statistical model of lexical diffusion. Bohmann (this volume) investigates the changing grammatical functions of the lexical item *because* in English-language Twitter data in different geographical contexts. Their findings reinforce a century of dialectological field work in which geographical distance and community size have been shown to be strong correlates of the diffusion of new language forms (Kretzschmar 2009).

Demographic information about Twitter users is limited, as the service, unlike some social media platforms, does not require users to provide real names, gender, or age. Pavalanathan and Eisenstein (2015) use automated methods to extract this information, and find that tweets with geographical coordinates tend to include more non-standard features and are more likely to be authored by young people and females.

Despite such work, research on the use of English on Twitter in non-L1 environments is not extensive. Investigation of individual linguistic behavior has been carried out: multilingual users' language choice on the platform reflects the predominant language of their social networks (Eleta and Golbeck 2014). Corpus-based studies comparing national varieties of English or core, outer circle, and expanding circle varieties (Kachru 1990) in terms of feature frequencies have yet to be conducted, as far as is known, either for standard grammatical classes such as parts-of-speech or for non-standard features such as emoticons or expressive lengthening.

This project looks at feature frequencies in English on Twitter in Finland by means of comparison to a reference corpus of global Twitter English messages with no geographic specification. Although the demographic characteristics of Finland-based users of Twitter can't be determined with any certainty, some inferences about persons writing English-language messages on Twitter in Finland may be made based on previous research into the use of English in Finland.

Taavitsainen and Pahta (2003, 2008), for example, discuss the use of English in Finnish daily life by examining Finnish print media advertisements and public signage that contain English words. They note that English has an "increasing influence" in Finland "in several fields of life" (2003: 12) and suggest that the use of English continues to increase; it may be entering a "new phase" (2008: 37). English lexical items are widely used in Finnish-language advertisements in print and television media (Paakkinen 2008). Leppänen (2007), in a conversation-analytic investigation of the use of English in four short excerpts from spoken and written online language samples by Finnish young people, attests a macro-scale language shift from Finnish to English among Finnish youth in certain contexts (167).

Most considerations of the role of English in Finland have been supported by qualitative analyses of a relatively small number of texts or recordings of spoken language, but there have also been efforts to compile larger-scale data on the use of English in Finland. Leppänen et al. (2011) present the results of a survey about the use of English in Finland administered to a sample of 1,500 Finnish respondents stratified by age, occupation, education, and gender. They find that Finns have good knowledge of English and a positive, pragmatic view towards the value of English skills in a globalized world: "by the 2000s, English had become not only an indispensable vehicular language in international interactions, but also a language used in many domains and settings within Finnish society" (16). The authors note that active users of English in Finland "are more likely to be youthful and involved in youth culture, have an interest in popular

culture, use the new media, and be alert to the demands/opportunities of an increasingly global economy" (166).

As English increasingly plays an important role in daily interaction in Finland, including in online communication, a characterization of English on Twitter in Finland in terms of its grammatical frequencies and underlying communicative dimensions contributes to the documentation and characterization of English as it continues to evolve as a global language (or set of global languages).

3 Data and methods

Approximately 93,000 tweets were collected from mid-March until early May 2013 via the Twitter Streaming API by selecting geo-tagged tweets originating from within a geographical box with the extent 60–70° N and 21–30° E, circumscribing the borders of Finland. To determine which tweets originated from within the borders of Finland, as well as in which region of Finland they originated, the latitude and longitude coordinates of each tweet were checked with the coordinates of the national and regional borders of Finland as encoded by GIS files publicly available through the Global Administrative Areas database GADM.[5] A comparison corpus, representing a random selection of approximately 305,000 tweets broadcast in late 2008 and early 2009, was downloaded via a commercial service in 2013.[6]

For both corpora, the language of each message was identified using the probabilistic language identification tool langid.py, which assigns language by comparing the frequencies of variable length n-grams (i.e., byte sequences that encode Unicode characters) in the text whose language is to be detected and comparing them with frequencies calculated from corpora in 97 languages, using a Bayesian classification algorithm (Lui and Baldwin 2012). The tool assigns a probabilistic value for the accuracy of the classification between 0 and 1. Because longer messages in a single language contain more byte n-grams that can be compared with the modeling data, they are typically assigned higher values, whereas language mixtures and extremely short user messages are typically assigned low values and sometimes misclassified. For that reason, only user messages determined to be in English with a probability of greater

5 Location disambiguation, factor analysis, and all other calculations were undertaken in *R*.

6 This data, collected at Texas A&M University, is no longer available for public download: Twitter policy since 2013 has discouraged public availability of Twitter corpora.

than 0.6 were retained in the two final corpora, the *Finland English Corpus* and the *Comparison English Corpus* (Table 1).

Table 1: Corpora size

	User messages	Tokens
All Finland tweets	93,451	1,039,865
Finland English Corpus	32,916	436,954
All Comparison tweets	305,310	3,361,444
Comparison English Corpus	181,861	2,864,798

The demographic identity or location of the authors of the messages in the Comparison English Corpus is unknown, but an examination of the messages suggests that a relatively high proportion of the tweets originate from the United States.[7]

Some filtering of tweets sent multiple times (often commercial advertisements generated automatically) was undertaken. Prior research has shown that broadcast-style tweets such as advertisements on Twitter do not figure prominently in conversational discourse (Ritter, Cherry and Dolan 2010). As such, messages are sometimes broadcast multiple times, their lexical and grammatical feature frequencies may be overrepresented in an analysis undertaken without filtering.

Twitter's default Streaming API access for end users is limited to 1% of the volume of traffic on the platform. As Twitter considers its proprietary data to have value to data miners, it provides higher levels of access primarily on a commercial basis. Given the high volume of messages broadcast by the platform, access limitations do not necessarily pose a practical problem for the compilation of a Twitter corpus. However, as noted above, only a small percentage of tweets include geographical coordinates, and data volumes from specific geographical locations are much more limited. Although Twitter is relatively popular in Finland, it is not among the countries with the highest per capita use of the platform (Mocanu et al. 2013). The size of the Finland English Corpus may not permit in-depth study of relatively rare grammatical or lexical phenomena. Nevertheless, large-scale trends in the frequency of grammatical features are evident in the data.

Automatic part-of-speech classification of the user messages in the two corpora was performed using the Carnegie-Mellon University Twitter Tagger

7 60% of the Comparison tweets are in English. As Twitter had less global penetration in 2008/2009 it is reasonable to assume that a relatively high proportion of the English tweets originate from the US.

(Gimpel et al. 2011; Gimpel et al. 2013; Owoputi et al. 2013). The 37 tags (Table 2) are applied according to a probabilistic model from a selection of tags from the Penn Treebank tagset (Marcus, Santorini, and Marcinkiewicz 1993). In addition to tags from the Penn Treebank set, the tagger applies distinct tags for the Twitter-specific types username, hashtag, and retweet, as well as a tag for URL addresses.

Table 2: Part-of-speech tags applied by the CMU tagger and used in the analysis

	Tag	Description		Tag	Description
1.	‘ ’	Quotation mark (“)	20.	RB	Adverb
2.	,	Comma	21.	RBR	Adverb, comparative
3.	.	Period (. *? !*)	22.	RBS	Adverb, superlative
4.	:	Punctuation (: ; . . . + − = < > / [ ] ~)	23.	RP	Particle
5.	CC	Coordinating conjunction	24.	RT	Retweet
6.	CD	Cardinal number	25.	TO	*to*
7.	DT	Determiner	26.	UH	Interjection
8.	EX	Existential *there*	27.	URL	Universal Resource Locator
9.	HT	Hashtag	28.	USR	Username (preceded by @)
10.	IN	Preposition or subordinating conjunction	29.	VB	Verb, base form
11.	JJ	Adjective	30.	VBD	Verb, past tense
12.	JJR	Adjective, comparative	31.	VBG	Verb, gerund or present participle
13.	JJS	Adjective, superlative	32.	VBN	Verb, past participle
14.	MD	Modal	33.	VBP	Verb, non-3rd person singular present
15.	NN	Noun, singular or mass	34.	VBZ	Verb, 3rd person singular present
16.	NNP	Proper noun, singular	35.	WDT	Wh-determiner
17.	NNS	Noun, plural	36.	WP	Wh-pronoun
18.	PRP	Personal pronoun	37.	WRB	Wh-adverb
19.	PRP$	Possessive pronoun			

Emoticons were detected in the corpora by filtering the output of the CMU Twitter Tagger for tokens that had been assigned the interjection tag.[8] Regular expressions were then used to select the subset of those tokens containing the characters that most frequently comprise emoticons, primarily non-letter ASCII characters as well as Unicode symbols. The 449 emoticon types determined in this manner were examined and types whose status as emoticons seemed questionable were removed, leaving a total of 240 emoticons for the ensuing

8 The Penn Treebank model uses the interjection tag for politeness forms, affective particles, and similar word types. The CMU Twitter Tagger, using the Penn Treebank model, applies the tag to emoticons as well.

analysis. Regular expressions were also used to capture expressive lengthenings in the Finland English and Comparison English data. All tokens containing at least three characters repeated in sequence were considered.[9]

Exploratory factor analysis of feature frequencies as determined by the tagger was then undertaken, allowing a preliminary characterization of the communicative and discourse dimensions of English on Twitter. These dimensions are used in the ensuing analysis, which focuses on the differences between English on Twitter in Finland and global Twitter English.

3.1 Exploratory factor analysis of feature frequencies

In order to conduct an exploratory factor analysis, mean feature frequencies were calculated from a merged dataset consisting of an equal number of 1000-token chunks from both the Finland English Corpus and the Comparison English Corpus. The frequencies were then used to construct a correlation matrix of the 37 individual features as variables.[10] A scree plot of eigenvalues for the correlation matrices suggested seven factors as optimal for the data. Factor loadings of the resulting factor analysis ≥ 0.3 (calculated using "varimax" rotation) are shown in Table 3.

If we consider the first two factors, shown in Figure 1, a viable interpretation of the communicative functionality of Twitter English features emerges. The first factor has a strongly positive loading on the features personal pronouns and non-3rd-person singular present verb forms (i.e., first- and second-person), while the features adverbs, base or infinitive verb forms, interjections, conjunctions, modal verbs, possessive pronouns, usernames, and Wh-adverbs have moderately positive loadings. There is a strongly negative loading on proper nouns and moderately negative loading on URLs, punctuation, and cardinal determiners (i.e., number words and numerals). This configuration suggests a functional separation between interacting with other users or situating one's own text in relation to discourse and supplying information in the form of specific reference. At one end of this dimension are interactive types and types pertaining to the

9 With three exceptions: tokens containing the sequence <www.> were excluded as URL addresses, and usernames and hashtags were not considered (multiple character sequences in these types are fixed and thus difficult to consider lengthenings in the same way as other lengthening types).

10 As the Finland English Corpus is shorter, the factor analysis was conducted upon 436 chunks of Finland English data (i.e., all of the tokens in the corpus) and an equivalent number of randomly selected 1000-token chunks from the Comparison English Corpus. See Biber (1988: 61–78, 1995: 85–140) for discussion of the methodology of exploratory factor analysis on textual material.

Table 3: Factor loadings for features in both corpora

	Factor						
	1	2	3	4	5	6	7
Proper noun, singular	−0.70				−0.62		
Personal pronoun	0.89	−0.32					
Adverb	0.56						
Verb, base form	0.57						
Verb, non-3rd person singular present	0.78						
Determiner		0.74					
Hashtag		−0.55					0.38
Preposition or subordinating conjunction		0.64					
Noun, singular or mass		0.67			0.38		
Interjection	0.30	−0.58	0.35				
to			−0.64				
Verb, gerund or present participle			−0.94				
Period (. ? !)				0.95			
Universal Resource Locator	−0.41	−0.31				−0.81	
Punctuation (: ; . . . + − = < > / [ ] ~)	−0.61			−0.33			−0.68
Quotation mark (“)							
Comma		0.36					
Coordinating conjunction	0.40						
Cardinal number	−0.31						
Existential *there*							
Adjective		0.43			0.32		
Adjective, comparative							
Adjective, superlative							
Modal	0.44						
Noun, plural		0.46					
Possessive pronoun	0.31						
Adverb, comparative							
Adverb, superlative							
Particle		0.43					
Retweet		−0.32					
Username (preceded by @)	0.41	−0.43	0.34				
Verb, past tense		0.39					
Verb, past participle		0.44					
Verb, 3rd person singular present		0.32					
Wh-determiner							
Wh-pronoun							
Wh-adverb	0.39						

Cum. variance = 0.41, $X^2 = 3198$, deg. fr. = 428, p-value < 10^{-232}

negotiation of stance expression, epistemic modality, affective orientation and discourse functionality, such as usernames, first- or second-person personal pronouns with present-tense verb forms, possessive pronouns, modal verbs, adverbs, and question words such as *who* or *why*. At the other end are types that specify entities, such as personal or place names, URLs, and numerical values (which have scalar/informational content but rarely organize large units of discourse), along with selected punctuation types.[11]

The second factor has a strong positive loading on determiners, slightly less strong positive loadings on prepositions and singular or mass nouns, and moderate positive loadings on commas, adjectives, plural nouns, prepositional components of phrasal verbs, past tense verb forms, past participles, and 3rd-person singular present verb forms. Moderately strong negative loadings are found on personal pronouns, hashtags, interjections, URLs, retweets, and usernames. This factor separates components of the nominal phrase and verb forms indicating past temporality from the discourse features specific to Twitter and another type used for discourse negotiation, interjections (which include emoticons). This dimension implies functional separation between types employed for reporting on events (nouns, determiners, past tense verb forms, phrasal verb particles) and types used to interact with other users of the Twitter environment and negotiate discourse concerns on the platform.

The interpretation of factors three through seven (not shown in Figure 1) is much more problematic: these factors have loadings with values ≥ 0.3 on only a few features or one feature. Factor three, which contrasts usernames and interjections with present participles/gerunds and *to*, may account for tweets sent by apps that automatically report user activity on a computer or smart device, such as listening to music, in the form: "listening to x,"[12] with no interactive or discourse organization types like emoticons or usernames. Factor five, contrasting singular or mass nouns and adjectives with proper nouns, is difficult to interpret. Factor six simply accounts for tweets with no URLs, and factor seven contrasts what may be alternative methods for the specification of tweet content or topicality: hashtags versus colons.

---

11 The punctuation types assigned this tag include those used to organize clause- or phrase structure, such as the colon, the semi-colon, and the ellipsis, as well as bracket types used for specification and types used to show relationships between numerals, such as the plus sign. For a discussion of the functions of punctuation types see Jones (1996), Nunberg (1990), and Quirk et al. (1985).

12 Although effort was made to filter for automated tweets (see above), many remain in the corpora.

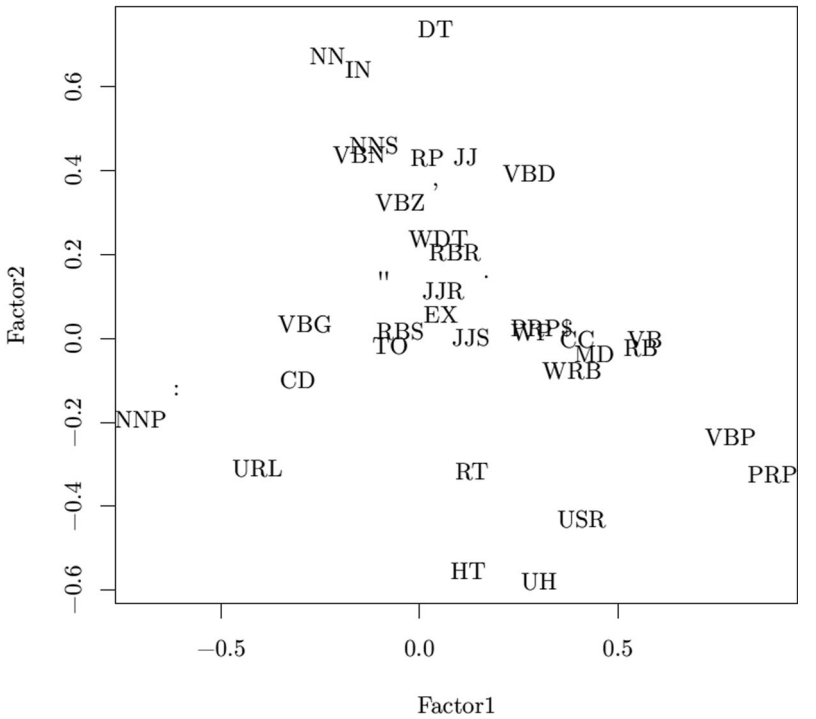

Figure 1: Factor loadings for 37 features (Factors 1 and 2 of the combined data)

Exploratory factor analysis suggests that Twitter discourse (in this data) may be interpreted as variable along two main dimensions. The first dimension, *interaction – specification*, contrasts interactive, affective or stance orientation towards discourse-local entities such as the self or other users with reference to entities beyond the immediate discourse of Twitter. A second dimension, *narration – discourse negotiation*, contrasts features such as nominal phrase elements and past-tense verb forms with prominent discourse-organization features such as the Twitter-specific hashtag, username, and retweet as well as the similarly versatile interjection tag (which marks emoticons).[13] When the features (including those with factor loadings less than 0.3) are plotted along

13 These dimensions are analogous to the first two dimensions proposed in Biber's analyses: 'Informational versus involved production' and 'Narrative versus non-narrative concerns' (1988: 115; 1995: 141–155), suggesting that the patterning of grammatical features in Twitter English may be similar to that of other registers and genres.

the first two dimensions (Figure 1), the shared communicative functions of determiners, nouns and prepositions (at the top of the figure); non-3rd-person singular present verb forms and personal pronouns (on the right); hashtags, interjections, retweets and usernames (at the bottom); and proper nouns, punctuation, URLs, and numerals (on the left) become visually apparent by means of proximity.

Returning to differences between the Finland English Corpus and the Comparison English Corpus, relative feature frequencies can be used to situate the varieties along the proposed dimensions. A dimension score is calculated by summing the standardized difference for each unique feature on the first two factors between the individual corpus and the larger, merged dataset used for the factor analysis (Table 4). These aggregate values quantify the differentiation of communicative and functional properties that underlie the discourse of the two corpora.[14]

Table 4: Dimension scores for the Finland English Corpus and Comparison English Corpus

	Dimension 1: Interaction – Specification	**Dimension 2: Narration – Discourse Negotiation**
Finland English Corpus	3.27	–2.04
Comparison English Corpus	–3.19	1.77

4 Results

The data for the Finland English Corpus and the Comparison English Corpus show differences in average message length as well as differences in the frequencies of the features under consideration. Interpretation of the results reinforces the findings of the factor analysis, suggesting that differences in communicative orientation between the two groups of users may underlie the observed patterns.

4.1 Tweet and token length

The length of Twitter user messages is limited to 140 characters, but within the range of one to 140 characters, there is wide variation in tweet length. Tweet

14 See Biber (1988: 93–97) for a description of the steps involved.

lengths for the Finland English and Comparison English corpora, as measured by number of characters per tweet, are shown in Figure 2. The spike at n = 140 is due to the automatic shortening of longer tweets by the service; messages longer than 140 characters are automatically shortened to 120 characters and a 20-character url linking to the longer text is added.

Disregarding the spike due to addition of a URL, the most common tweet length for the Finland English data is 28 characters, whereas the mode for the Comparison English data is 58 characters; the corresponding mean values are 71.10 characters for the Finland data and 78.99 characters for the comparison data.

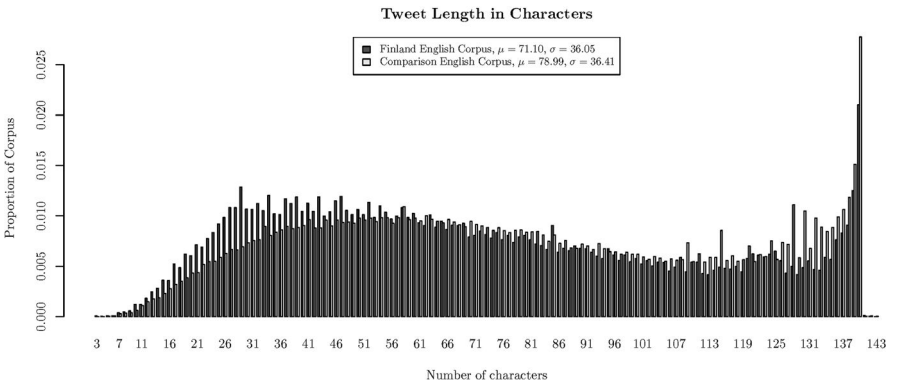

Figure 2: Tweet length in characters, Finland English and Comparison English corpora

Previous research has found that mean tweet length has been decreasing since the service was initiated in 2007. Alis and Lim show that mean tweet length for a Twitter user message corpus compiled between 2009 and 2012 decreased by approximately 8 characters, from ~85 to ~75 characters per tweet, values comparable to the mean tweet lengths in in the Finland English and Comparison English Corpora. They also find that for tweets that are geo-encoded, mean user message length for US states may reflect demographic characteristics of their populations (2013: 7).

Average token (word) length also differs between the Finland English and Comparison English corpora. The mean length of the tokens in the Finland English Corpus is 4.54 characters, whereas Comparison English Corpus tokens are on average 4.21 characters long. This is possibly due to much lower rates of article use in the Finland English data. There are no articles in Finnish: Grammatically, the function of providing information on the status of the referent as

known or not known in discourse typically falls in Finnish to demonstratives. Unsurprisingly, when L1 Finnish speakers write or speak in English, they tend to use articles less frequently than do L1 English speakers. The Finland English Corpus exhibits much lower frequencies of articles than does the Comparison English Corpus. Indefinite articles are used in the Finland English Corpus at a rate approximately 76% that of the Comparison English Corpus, but the definite article occurs only 64% as frequently.

4.2 Emoticons

The data show a large range of variation in the use and distribution of emoticons. In the Finland data, Twitter users who tweet in English are more likely to use emoticon symbols than those who tweet in other languages: 24.9% of English-language tweets from Finland contained at least one emoticon, and 56.1% of the users represented in the Finland English Corpus used at least one emoticon.[15] The prevalence of emoticons in the Comparison English Corpus was much more limited. Only 9.8% of the tweets in the Comparison English Corpus included at least one emoticon, and only 10.2% of the users represented in the Comparison English Corpus utilized at least one emoticon. In terms of regularized frequencies, the frequency of all 240 emoticon types considered is 23.87 per thousand tokens in the Finland English Corpus and 6.79 per thousand tokens in the Comparison English Corpus.[16]

An analysis of emoticon type relative frequencies from Schnoebelen (2012) was replicated in part using the Finland English and Comparison English data; the findings are summarized (along with Schnoebelen's results) in Figure 3. For the most part, the Finland English, Comparison English, and Schnoebelen data show a similar rank/frequency profile for some of the most widely used emoticons. Although Finland-based users employ emoticons far more frequently overall, their proportional use of different emoticon types is similar to that of users elsewhere:

15 Interestingly, the overall rate of emoticon use in the entire Finland corpus (i.e., in all languages) is lower – Finland-based tweeters use more emoticons when writing in English.

16 It may be the case that this large difference results from an increase in use of emoticons overall on Twitter in 2013 compared to 2008–9: There seems to be no research into the prevalence of emoticon use on Twitter over time.

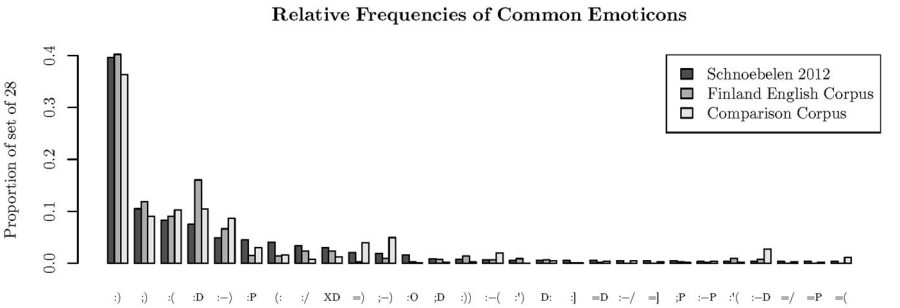

Figure 3: Relative frequency of 28 emoticon types in Schnoebelen 2012, Finland English Corpus and Comparison English Corpus

The relative frequencies for this specific set of 28 emoticons are much the same.[17]

4.3 Expressive lengthening

Expressive lengthening is also much more prevalent in the Finland English Corpus. The feature occurs at a rate of 5.00 per thousand tokens in the Finland English data and 1.86 per thousand tokens in the Comparison English data. The most frequent lengthened types and their rates of occurrence are shown in Figure 4.

The types *awww, sooo, soooo, ahhh*, and *oooh* are among the most frequent lengthened types in both corpora. The most frequent type in the Finland data is the non-pronounceable non-dictionary word *xxx*, usually interpreted as kiss symbols. Two types among the most frequent Finnish lengthenings, *XDDD* and *:DDD*, can be interpreted as emoticons with multiple mouths. The other most frequent types in the Finnish data consist of lengthened dictionary words (*meee, meeee, sooo, soooo, nooo, noooo, pleaseee, iii, yesss, goood*) and lengthened interjections or pronounceable non-dictionary words (*hmmm*, *mmm*, *mmmm*, *awww, aaawww, ahhh, uuu, aaah*). Eighteen of the twenty most frequent lengthenings consist of three letters in succession; two types (*soooo* and *noooo*) contain 4-character lengthenings.[18]

17 Wilcoxon signed-rank tests show no significant difference between the median ranks of the relative frequencies of the 28 emoticons for the three corpora: For Finnish and Comparison data V = 222, p-value = 0.68; for Finnish and Schnoebelen data V = 161, p-value = 0.35, and for Comparison and Schnoebelen data V = 174, p-value = 0.52.

18 The distinction between pronounceable and non-pronounceable dictionary and non-dictionary words is from Bamman, Eisenstein and Schnoebelen (2014).

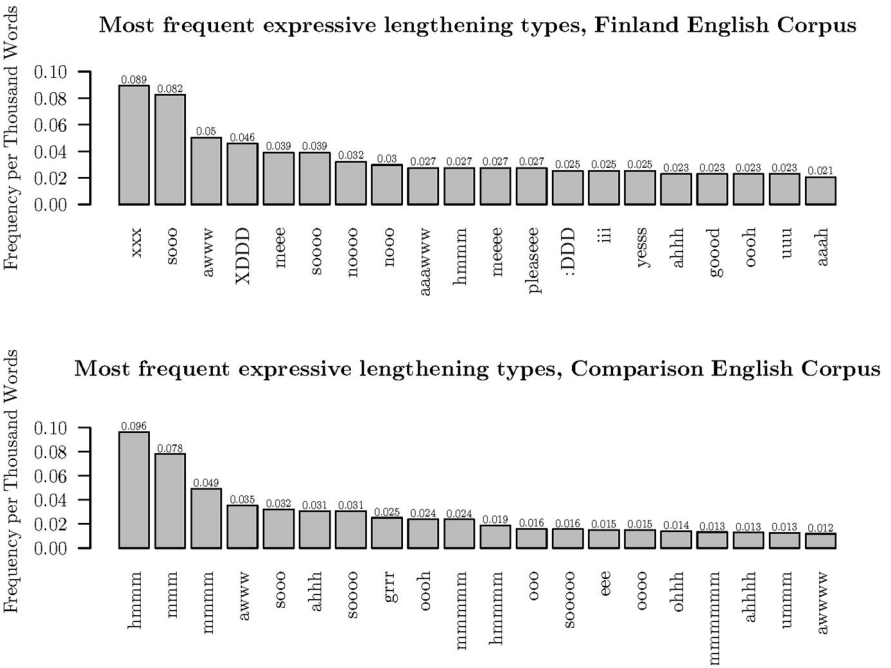

Figure 4: Most frequent expressive lengthening types, Finland English and Comparison English corpora

The most frequent types in the Comparison English Corpus are almost all interjections or pronounceable non-dictionary words, most of which correspond to discourse-marking sounds used in spoken conversation (*hmmm, mmm, mmmm, awww, ahhh, grrr, oooh, mmmmm, hmmmm, ooo, eee, oooo, ohhh, mmmmmm, ahhhh, ummm, awwww*). Although lengthenings are much less frequent in the Comparison English Corpus, they are more likely to consist of longer sequences of characters: Thirteen of the twenty most frequent lengthening types contain three-character sequences, four four-character sequences, two five-character sequences, and one a six-character sequence.

An interesting pattern emerges when one considers the proportion of all lengthenings by character in the two corpora (Figure 5). Again, the profiles are similar, but there are some differences.

For the Finland English data, the letter *o* comprises 22% of the lengthenings in the corpus. In general, vowel characters are more subject to lengthening than are consonants, with characters representing open and mid vowels *o*, *e*, and *a* more likely to be lengthened than those that represent close vowels *i* and *u* or

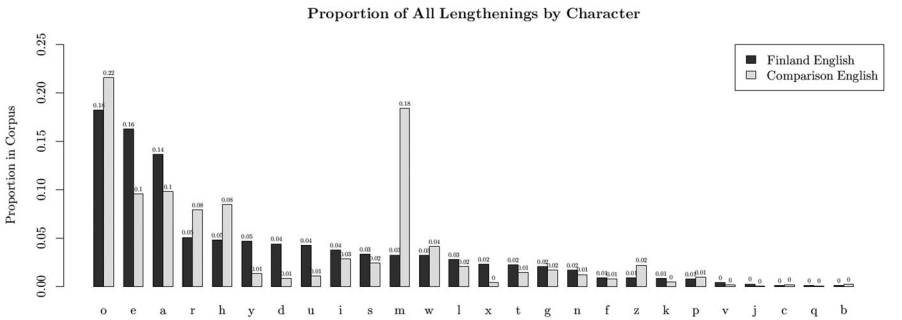

Figure 5: Expressive lengthening sequences by character, Finland English and Comparison English corpora

the semi-vowel *y*. Among non-vowel characters, *r* is most subject to lengthening, followed by *h*, *d*, *s*, *m*, and *w*. The characters *l*, *x*, *t*, *g*, *n*, and *f* are slightly less likely to be subject to lengthening, and the characters *z*, *k*, *p*, *v*, *j*, *c*, *q*, and *b* are the least likely to be lengthened.

The Comparison English Corpus data shows overall lower lengthening frequencies, but also a somewhat different distribution of lengthening types and frequencies. The character most susceptible to lengthening is again *o*. Vowels are also in the Comparison English data somewhat more likely to be lengthened than consonants: Again characters representing open and mid vowels (*o*, *a*, *e*) are lengthened more often than characters representing close vowels (*i*, *u*). The characters *h* and *r* are among the more frequent targets for consonant lengthening. The most striking difference is in the lengthening of *m*, which is proportionately six times more likely to be lengthened in the Comparison English Corpus.[19] The consonants *w*, *s*, z, *l*, *g*, *t*, and *n* follow in the ranking, with the consonants *p*, *d*, *f*, *k*, *x*, *c*, *b*, *v*, *q*, and *j* the least likely to be lengthened in the Comparison English data.

Summarizing the results pertaining to emoticons and expressive lengthening, the two features are far more prevalent in the Finland English Corpus. Although the distribution of emoticon types in the two corpora are similar, the character targets for expressive lengthening differ.

4.4 Part-of-speech features

While emoticon and expressive lengthening frequencies were calculated by using regular expressions to retrieve tokens from the corpora, the frequencies

19 Written forms of common disfluency particles or fillers in Finnish include types such as *ööh*, *siis*, and *niinku*, but not *hmm* or related types.

for other part-of-speech or discourse types were based on tags applied by the CMU Twitter Tagger. The Finland English and Comparison English data exhibit different distributional profiles for the relative frequencies for 37 grammatical feature tags. The relative frequency of each feature is shown in Table 5 as the logarithmic odds ratio θ of frequency in the Finland English Corpus to frequency in the Comparison English Corpus: Features in the left-hand column are overrepresented in the Finland English Corpus; those on the right in the Comparison English Corpus.[20] Differences in frequencies are significant according to the results of a chi-squared test of independence at $p < 0.001$, except for those features marked with an asterisk.

Table 5: Logarithmic odds ratios (Finland English Corpus vs. Comparison English Corpus) for 37 features

	Feature	θ		Feature	θ
1	Hashtag	3.36	1	Phrasal particle	−0.56
2	Retweet	1.68	2	Verb, present participle or gerund	−0.49
3	Username (preceded by @)	0.77	3	Other punctuation	−0.49
4	Interjection	0.75	4	Verb, past participle	−0.39
5	Personal pronoun	0.35	5	Wh-determiner	−0.31
6	Wh-adverb	0.35	6	Proper noun	−0.30
7	Verb, non-3rd-person singular present	0.34	7	*to*	−0.30
8	URL	0.22	8	Noun, singular or mass	−0.24
9	Adjective, superlative	0.22	9	Determiner	−0.24
10	Coordinating conjunction	0.21	10	Period, question mark, exclamation mark	−0.21
11	Adverb	0.17	11	Verb, 3rd-person singular present	−0.20
12	Modal verb	0.17	12	Comma	−0.17
13	Wh-pronoun	0.16	13	Verb, past tense	−0.16
14	Verb, base form	0.10	14	Noun, plural	−0.15
15	Existential *there**	0.09	15	Adverb, comparative	−0.15
16	Adverb, superlative*	0.08	16	Preposition	−0.10
17	Possessive pronoun	0.07	17	Adjective	−0.08
18	Quotation mark*	0.02	18	Adjective, superlative*	−0.07
			19	Cardinal number*	−0.02

The three features most overrepresented in the Finland English data correspond to the Twitter-specific tags applied by the CMU Twitter tagger. Hashtags are used in the Finland English Corpus at a rate almost 29 times that of the Comparison English Corpus. Retweets, or re-broadcastings of a tweet by a different

20 The statistic is calculated according to the formula $\log O_{11}O_{22}/O_{21}O_{12}$, where O_{11} and O_{21} represent the number of occurrences of the feature in the Finland English and Comparison English Corpora, respectively, and O_{12} and O_{22} the corresponding number of tokens that do not represent the feature.

user, are more than five times more common in the Finland English Corpus. Finland-based Twitter users tweeting in English are more than twice as likely as Comparison English users to utilize usernames preceded by <@>.

The frequency of interjections in the Finland English Corpus is more than twice that of the Comparison English Corpus. Tokens assigned the interjection tag include emoticons, non-standard initialisms such as *lol*, non-dictionary pronounceable types such as the hesitation marker *ummm* or the laughter indicator *haha*, as well as lexical items such as profanity and politeness markers. The discrepancy between the corpora reflects the high rate of use of emoticons in the Finnish data: As noted above, emoticons are used approximately 3.5 times more often in the Finland English data.

Personal pronouns and Wh-adverbs are both used in the Finland English Corpus at a rate 1.42 times that of the Comparison English Corpus. Non-3rd-person singular present verb forms are more common in the Finland English Corpus than in the Comparison English Corpus by a factor of 1.4. The final three categories for which the Finland English Corpus has a substantially higher rate of use than does the Comparison English Corpus are URLs, superlative adjectives, and coordinating conjunctions, used in the Finland English Corpus at rates 1.25, 1.24, and 1.23 times that of the Comparison English Corpus.

As can be seen in Table 3, the particle component of phrasal verbs is the most overrepresented part-of-speech in the Comparison English, occurring at a rate 1.75 times that of the Finland English Corpus. Participles, gerunds, and punctuation are approximately 63% more common in the Comparison English data.[21] Types overrepresented by 30%–50% include past participles, the words *what* and *which*, proper nouns, and *to*. Singular or mass nouns, determiners, the punctuation types < . *?* *!* >, and 3rd-person singular present verb forms are overrepresented by 20%–30%. Past tense verbs, plural nouns, comparative adverbs, and prepositions are more than 10% overrepresented. Adjectives are slightly (8%) overrepresented. Superlative adjectives and numbers are also more common, although the difference is insignificant.

5 Analysis and discussion

The findings from the analysis of grammatical features as they are manifest in the two principal corpora help to situate English on Twitter in Finland within the communicative dimensions *Interaction* versus *Specific Reference* and *Narration* versus *Discourse Negotiation*. Specifically, they allow a preliminary assessment of the extent to which Twitter English in Finland differs from global Twitter

21 Punctuation marks given this tag are those in the set <*: ; … + − = < > / [ ] ~*>.

English and how an emergent Finland-based Twitter English variety could be characterized.

5.1 Tweet length

Studies of established corpora have documented average sentence lengths of between 17 and 22 word tokens for sentences from the Brown Corpus, the British National Corpus or the London-Oslo-Bergen Corpus (Ellegård 1978: 23; Fengxiang 2007: 129). As to be expected for a medium with an upper limit on the number of characters per message, Twitter message lengths are much shorter: the mean lengths for the Finland English and Comparison English corpora are 13.27 and 15.75 tokens, respectively. If punctuation characters are not considered tokens (a common approach in corpus-based lexical studies), the mean message lengths are 9.66 and 12.59 tokens, respectively. These values correspond to mean message lengths of 11.9 tokens and 10 tokens found for other corpora compiled from Twitter or from SMS messages (Walkowska 2009: 149; Xu, Ritter and Grishman 2013). They are slightly longer than mean message lengths reported for instant messaging corpora: Baron reports an average IM message length of 5.4 words (2004: 409); Squires calculates an average IM message length from a different IM corpus of 6.18 words (2012: 299).

Finland-based messages in English on Twitter are significantly shorter than non-Finnish English Twitter messages in terms of number of characters per tweet and number of tokens per tweet, and Finland English messages utilize significantly fewer long (≥ 6 characters) words. Zipf noted the inverse relationship between word length and frequency of use, suggesting that a "principle of least effort" optimizes expression length according to communicative efficiency considerations (1949: viii).

Sigurd, Eeg-Olofsson and van Weijer (2004) confirm the inverse relationship between word length in characters or syllables and frequency for English, Swedish and German texts, and observe that sentence length exhibits a similar distributional profile, best approximated mathematically by the Gamma distribution. Agreeing with Zipf, they suggest that communicative economy concerns govern the relationship between length and informational content of words and sentences.

The tweet length findings from the Finland and Comparison English corpora can be interpreted in the context of the communicative economy observations of Zipf and others as indications of the functional–pragmatic dynamics of language use online. Shorter words and shorter tweets generally contain less information than do longer tweets and longer words. In aggregate, Twitter discourse

contains less information and is more interactive than the discourse of text types such as news reports, academic writing, or fiction. The shortness of tweets corresponds to communicative functions typical of Twitter, which include self-representation, often in abbreviated form, negotiation of discourse concerns, and interactivity. In this respect, the Finland English Tweets are even less informational and more interactive than the Comparison English Corpus tweets: they are shorter and contain fewer long words. Non-Finland English tweets, although similar to Finland English tweets in many ways, reflect a slightly broader range of communicative functions pertaining to the presentation of information.

These results suggest that language use may differ systematically between Finland-based persons writing on Twitter and other users of English on the platform, and that the difference, at least in part, may reflect language interference phenomena. Shorter words contain less information (Zipf 1949), as do shorter sentences. Tweet length differs between the Finland English and the Comparison English corpora in a way that suggests English on Twitter in Finland may be less information-oriented and more interactive.

5.2 Emoticons

Overall, the Finland English Corpus is rich in emoticon usage, and Finland-based Twitter users writing in English are enthusiastic users of emoticons: The Finland English Corpus exhibits much higher rates of use for emoticon types per tweet and per user than does the Comparison English Corpus. For Finland, the relative proportions of all emoticon use comprised by certain specific emoticon types are similar in the Finland English and the Comparison English corpora and comparable to those reported by Schnoebelen (2012), suggesting that whatever the evolution of the communicative or discourse-organization functions of emoticons may be, their type distributions have been somewhat stable across cultural boundaries in English-language Twitter from late 2008–2013.

The interpretation of emoticons as direct reflections of the emotional state of the user or as written equivalents of prosodic features is problematic. In light of a recent study in which emoticons on Twitter are analyzed as discourse markers with various functional roles, it may be the case that emoticons in the Finland data are "interactive in nature, positioning audiences around propositions" (Schnoebelen 2012: 118). The interpretation of emoticons as a linguistic resource whose meaning is contextualized by discourse considerations is reinforced by research into non-L1 use of English on instant messaging, where emoticons may serve as contextualization cues and compensatory gestures for non-native-speaker competence (Vandergriff 2014b).

The idea that emoticons are used for multiple communicative functions is somewhat similar to pragmatic interpretations of hashtag functionality of Zappavigna (2011) or Wikström (2014). According to these analyses, the status of the hashtag, the <@> symbol, or emoticons in online discourse on platforms such as Twitter continues to evolve.

The interpretation of emoticons as symbols with discourse organization functions is strengthened by the results of the factor analysis, which shows a shared communicative space for emoticons and hashtags. Emoticon use in English-language Twitter may correspond to an evolving youth-based communicative functionality pertaining to discourse negotiation strategies which has developed on Twitter and in other social media.

5.3 Orthography and expressive lengthening

Widespread orthographical variation in Twitter may represent individuals and groups utilizing non-standard language variants to create social meaning. Non-standard orthography in the form of expressive lengthening is a frequent feature in both the Finland English and the Comparison English Corpus, but the feature is much more extensive in the Finland English Corpus.

Overall, vowel characters are the most likely to be lengthened, but Finland Twitter users writing in English tend to lengthen different somewhat different consonant characters than do non-Finland Twitter users writing in English. This may reflect L1 interference phenomena for the Finland English users: For example, voiced plosives are uncommon in Finnish.

Considering the distributions of lengthening sequences according to character, the phenomenon may reflect phonological and prosodic considerations as well as discourse and pragmatic factors. Phonological and phonetic experiments have shown that longer vowel duration can be perceived by listeners as marked for affect or emotional content (Klatt 1976). Vowels and other characters that correspond to segments in speech with higher sonic prominence, such as the sonorant nasals and approximant laterals, seem more likely to be lengthened than characters corresponding to obstruents such as stops. Morphological considerations such as segment- and word boundaries undoubtedly also play a role in this complex patterning. The extent to which L1 Finnish may play a role in the choice of characters to be lengthened deserves further investigation.

Expressive lengthening was not considered as a variable in the factor analysis used to identify dimensions of functional variation of Twitter language – it is not identified by the CMU Twitter Tagger and may have a status that is not equivalent to that of parts-of-speech or discourse markers (e.g., a string such as

yesss is both an interjection and an example of expressive lengthening). Nonetheless, as expressive lengthening may mark affective orientation, its prevalence in the Finland English Corpus can be tentatively interpreted as contributing to the interactive nature of English on Twitter in Finland, a variety in which expression of affective stance plays an important role.

5.4 Part-of-speech frequencies

Exploratory factor analysis of aggregate part-of-speech frequencies was used to identify two dimensions of functional variation in Twitter English. The examination of feature occurrence ratios between the Finland English and Comparison English data provides further insight into the dynamics of English on Twitter in Finland. Hashtags and retweets, features associated with discourse organization and orientation, are the two most overrepresented features in the Finland English Corpus. Of the four most overrepresented features, three are unique to the Twitter language ecosystem (the hashtag, the retweet, and the username), and one feature, the interjection, is associated with emoticons, another discourse-organization type.

It should be noted that the relative lack of use of hashtags in the Comparison English Corpus may be due in part to the fact that the comparison data was collected in late 2008–2009, prior to the introduction of the "Trending Topics" feature in Twitter which highlighted the most-used hashtags on the homepages of Twitter uses. This interface change by Twitter prompted an increase in the prevalence of hashtags on the service. Still, the extent to which hashtags are overrepresented in the Finland English Corpus (29 times more common than in the Comparison English Corpus) is remarkable.[22] As Wikström (2014) notes, the changing nature of hashtag use on Twitter may represent an example of the ways in which functions originally envisaged for an innovation within the framework of a technological medium are utilized in an unexpected manner by members of a user community and evolve to become emblematic of the medium itself.

Language use online may not differ too dramatically from linguistic behavior under other circumstances, and it would be unwise to consider technological developments to be the sole force driving changes in language use online – what Squires terms "technological determinism" (2010). Nonetheless, it may be

22 A preliminary analysis of smaller but similarly processed data sets collected in 2015 again find a higher rate of use of hashtags in Finland-based English tweets, albeit by a smaller factor than in the data in this study.

the case that the evolving norms of language use on Twitter, as they are manifest in frequency data for Finland, exhibit a technological moment.

6 Summary, outlook, and conclusion

English is increasingly used online in societies where it has not traditionally played a large role in daily communication. Factor analysis was used to identify two dimensions of variation in data consisting of a corpus of Finland-based English-language Twitter messages and English-language Twitter messages with no geographical location. The dimensions "Interaction versus Specification" and "Narration versus Discourse Negotiation" best capture the co-occurrence of grammatical and discourse features in the data and clearly distinguish English on Twitter in Finland from global Twitter English: The former is more interactive and its authors make more use of discourse-referential types, whereas the latter is more informational and narrative.

Emoticons are used far more frequently in English on Twitter in Finland compared to global Twitter English, although the proportional use of common emoticon types is comparable. Given previous findings as to the diverse functionality of emoticons in CMC and on Twitter, and in light of the association of emoticons with hashtags, usernames and retweets, according to an exploratory factor analysis, emoticons are best interpreted as types with various functions, including discourse organization, affective stance orientation, and evaluation.

The non-standard feature expressive lengthening is overrepresented in English on Twitter in Finland. Standard word forms and emoticon types are more likely to be lengthened in English from Finland, whereas pronounceable non-dictionary words are more common lengthening targets globally. There is evidence that somewhat different letters are typically lengthened in Finland-based English compared to global English. Expressive lengthening may be a means of imbuing word forms with affective content, but a closer examination of the phenomenon is needed in order to confirm this hypothesis. A consideration of expressive lengthening in different languages and its relationship to the phonological characteristics of those languages would also be informative.

Part-of-speech frequencies for individual features can be interpreted according to the findings of the factor analysis: they suggest that Finland-based users of Twitter writing in English exhibit a more interactive communicative orientation and make particular use of language features on Twitter associated with the organization and negotiation of discourse: hashtags, retweets, usernames, and interjections, many of which are emoticons.

While the present study proposes differences in group communicative behavior based on aggregate feature frequencies, further research is needed to establish the identity of Finland-based persons tweeting in English, especially in light of recent findings that young people in Finland are the most likely to report high levels of fluency in English (Leppänen et al. 2011) and that young female users are overrepresented in tweets that include latitude and longitude metadata (Pavalanathan and Eisenstein 2015). A user base for the Finland tweets skewed towards a younger and more female demographic may account for higher frequencies of some features, such as non-standard orthography and emoticons.

In an era when an increasing proportion of English-language communication is mediated by technology and internet-based services such as Twitter, a survey of the extent of use of English as it continues to evolve globally must take into account local use of English in online contexts. For Finland, English as it is used on Twitter is characterized by shorter message length, high frequencies of non-standard language features such as expressive lengthening and emoticons, as well as a specific configuration of part-of-speech and discourse item frequencies. The study suggests that English on Twitter in Finland emerges as a distinct variety on the basis of the high frequencies of features that are primarily used to interact with others; indicate evaluative, epistemic, or affective stance; and, situate these elements in discourse. In a broader sense, the analysis suggests that users of Twitter utilize non-standard and platform-specific features to construct and negotiate meanings at the interface of online interactivity and technological change.

7 References

Alis, Christian & May Lim. 2013. Spatio-temporal variation of conversational utterances on Twitter. *PLoS ONE* 8(10). http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077793 (accessed 20 March 2016).

Androutsopoulos, Jannis. 2006. Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics* 10(4). 419–438.

Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2). 135–160.

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

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Crystal, David. 2003. *English as a global language*. 2nd edn. Cambridge, UK: Cambridge University Press.

Crystal, David. 2006. *Language and the internet*. 2nd edn. Cambridge, UK: Cambridge University Press.

Dresner, Eli & Susan C. Herring. 2010. Functions of the non-verbal in CMC: Emoticons and illocutionary force. *Communication Theory* 20(3). 249–268.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(11). http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0113114 (accessed 20 March 2016).

Eleta, Irene & Jennifer Golbeck. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior* 41. 424–432.

Ellegård, Alvar. 1978. *The syntactic structure of English texts: A computer-based study of four kinds of text in the Brown University Corpus*. Göteborg: Acta Universitatis Gothoburgensis.

European Commission. 2011. *Flash Eurobarometer 313: User language preference online*. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf (accessed 20 March 2016).

Fengxiang, Fan. 2007. A corpus based quantitative study on the change of TTR, word length and sentence length of the English language. In Peter Grzybek & Reinhard Köhler (eds.), *Exact methods in the study of language and text: Dedicated to Gabriel Altmann on the occasion of his 75th birthday*, 123–130. Berlin & New York: Mouton de Gruyter.

Gimpel, Kevin., Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan & Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. Human Language Technologies: North American Chapter of the Association for Computational Linguistics *(NAACL-HLT 2011)* 12. 42–47. Stroudsburg, PA: Association for Computational Linguistics. www.ark.cs.cmu.edu/TweetNLP/gimpel+etal.acl11.pdf (accessed 20 March 2016).

Gimpel, Kevin, Nathan Schneider & Brendan O'Connor. 2013. Annotation guidelines for Twitter part-of-speech tagging version 0.3 (March 2013). Computational Science Department, Carnegie Mellon University, Pittsburgh, PA. http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf (accessed 20 March 2016).

Grieve, Jack, Andrea Nini, Diansheng Guo & Alice Kasakoff. 2015. Big data for the analysis of language variation and change. Paper presented at From Data to Evidence: Big Data, Rich Data, Uncharted Data, University of Helsinki, 19–22 October. https://dl.dropboxusercontent.com/u/99161057/D2E_GRIEVEETAL.pdf (accessed 20 March 2016).

Herring, Susan C. 2001. Computer-mediated discourse. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *Handbook of discourse analysis*, 613–634. Oxford: Blackwell.

Hillebrand, Friedhelm (ed.). 2010. *Short message service (SMS): The creation of personal global text messaging*. New York: Wiley.

Honeycutt, Courtenay & Susan Herring. 2009. Beyond microblogging: Conversation and collaboration via Twitter. *System Sciences (HICSS)* 42, 1–10. http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf (accessed 20 March 2016).

Hutchby, Ian. 2001. *Conversation and technology: From the telephone to the internet*. Cambridge, UK: Polity.

Jones, Bernard. 1996. *What's the point? A (computational) theory of punctuation*. Edinburgh, UK: University of Edinburgh dissertation.

Kachru, Braj. 1990. *The alchemy of English: The spread, functions, and models of nonnative Englishes*. Urbana, IL: University of Illinois Press.

Klatt, Dennis H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59. 1208–1221.

Kretzschmar, William A. 2009. *The linguistics of speech*. Cambridge, UK: Cambridge University Press.

Leetaru, Kalev H., Shaowen Wang, Guofeng Cao, Anand Padmanabhan & Eric 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 20 March 2016).

Leppänen, Sirpa. 2007. Youth language in media contexts: Insights into the functions of English in Finland. *World Englishes* 26(2). 149–169.

Leppänen, Sirpa, Anne Pitkänen-Huhta, Tarja Nikula, Samu Kytölä, Timo Törmäkangas, Kari Nissinen, Leila Kääntä, Tiina Räisänen, Mikko Laitinen, Heidi Koskela, Salla Lähdesmäki & Henna 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, Finland: Varieng. http://www.helsinki.fi/varieng/series/volumes/05/evarieng-vol5.pdf (accessed 20 March 2016).

Lui, Marco & Timothy Baldwin. 2012. Langid.py: An off-the-shelf language identification tool. *Association for Computational Linguistics* 50, 25–30. http://www.aclweb.org/anthology/P12-3005 (accessed 20 March 2016).

Magdy, Amr, Thanaa M. Ghanem, Mashaal Musleh & Mohamed F. Mokbel. 2014. Exploiting geo-tagged tweets to understand localized language diversity. *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data (GeoRich '14)*, 7–12. New York, NY: Association for Computing Machinery. http://dl.acm.org/citation.cfm?id=2619114&CFID=772834627&CFTOKEN=94166201 (accessed 20 March 2016).

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics* 19(2). 313–330.

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang & Alessandro Vespignani. 2013. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE* 8(4). http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061981 (accessed 20 March 2016).

Morstatter, Fred, Jürgen Pfeffer, Huan Liu & Kathleen M. Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Association for the Advancement of Artificial Intelligence International Conference on Weblogs and Social Media (AAAI-ICWSM 2013)* 7. 400–408. http://www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013.pdf (accessed 20 March 2016).

Nunberg, Geoffrey. 1990. *The linguistics of punctuation*. Palo Alto: CSLI.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider & Noah. A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. Human Language Technologies: North American Chapter of the Association for Computational Linguistics *(NAACL-HLT 2013)* 14. 380–390. http://www.ark.cs.cmu.edu/TweetNLP/owoputi+etal.naacl13.pdf (accessed 20 March 2016).

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

Paakkinen, Terhi. 2008. Coolia englantia suomalaisissa mainoksissa [Cool English in Finnish advertising]. In Sirpa Leppänen, Tarja Nikula & Leila Kääntä (eds.), Kolmas kotimainen: Lähikuvia englannin käytöstä Suomessa [The third domestic language: Close-ups of the use of English in Finland], 299–331. Helsinki, Finland: Suomalaisen Kirjallisuuden Seura.

Paolillo, John. C. 2001. Language variation on Internet Relay Chat: A social network approach. *Journal of Sociolinguistics* 5(2). 180–213.

Pavalanathan, Umashanthi & Jacob Eisenstein. 2015. Confounds and consequences in geotagged Twitter data. arXiv:1506.02275v2 [cs.CL]. http://arxiv.org/pdf/1506.02275v2.pdf (accessed 20 March 2016).

Ptaszynski, Michal, Rafal Rzepka, Kenji Araki & Yoshio Momouchi. 2011. Research on emoticons: Review of the field and proposal of research framework. Association for Natural Language Processing (NLP-2011) 17, 1159–1162. http://arakilab.media.eng.hokudai.ac.jp/~ptaszynski/data/E5-6.pdf (accessed 20 March 2016).

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive grammar of the English language*. London: Longman.

Rao, Delip, David Yarowsky, Abhishek Shreevats & Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, 37–44. New York, NY: Association for Computing Machinery. http://dl.acm.org/citation.cfm?id=1871993&CFID=772834627&CFTOKEN=94166201 (accessed 20 March 2016).

Rezabek, Landra L. & John J. Cochenour. 1998. Visual cues in computer mediated communication: Supplementing text with emoticons. *Journal of Visual Literacy* 18. 201–215.

Ritter, Alan, Colin Cherry & Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. Human Language Technologies: North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010) 11. 172–180. http://www.aclweb.org/anthology/N10-1020 (accessed 20 March 2016).

Schnoebelen, Tyler. 2012. Do you smile with your nose? Stylistic variation in Twitter emoticons. University of Pennsylvania Working Papers in Linguistics 18(2). 115–125. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1242&context=pwpl (accessed 20 March 2016).

Sebba, Mark. 2007. Spelling and society: The culture and politics of orthography around the world. Cambridge, UK: Cambridge University Press.

Sigurd, Bengt, Mats Eeg-Olofsson & Joost Van Weijer. 2004. Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica* 58. 37–52.

Squires, Lauren. 2010. Enregistering internet language. *Language in Society* 39. 457–492.

Squires, Lauren. 2012. Whos punctuating what? Sociolinguistic variation in instant messaging. In Alexandra Jaffe, Jannis Androutsopoulos, Mark Sebba & Sally Johnson (eds.), *Orthography as social action: Scripts, spelling, identity and power*, 289–324. Berlin: De Gruyter.

Squires, Lauren. 2016. Twitter: Design, discourse, and the implications of public text. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), The Routledge handbook of language and digital communication, 239–256. London and New York: Routledge.

Taavitsainen, Irma & Päivi Pahta. 2003. English in Finland: Globalisation, language awareness and questions of identity. *English Today* 19(4). 3–15.

Taavitsainen, Irma & Päivi Pahta. 2008. From global language use to local meanings: English in Finnish public discourse. *English Today*, 24(3). 25–38.

Tagliamonte, Sali & Derek Denis. 2008. Linguistic ruin? Lol! Instant messaging and teen language. *American Speech* 83(1). 3–34.

Twitter. 2015. Company facts. https://about.twitter.com/company (accessed 20 March 2016).

Vandergriff, Ilona. 2014a. Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics* 51. 1–12.

Vandergriff, Ilona. 2014b. A pragmatic investigation of emoticon use in nonnative/native speaker text chat. *Language@Internet* 11(4). http://www.languageatinternet.org/articles/2014/vandergriff (accessed 20 March 2016).

Walkowska, Justyna. 2009. Gathering and analysis of a corpus of Polish SMS dialogues. In M. Alojzy Kłopotek, Adam Przepiórkowski, Sławomir T. Wierzchoń & Krzysztof Trojanowski (eds.), *Recent advances in intelligent information systems*, 145–157. Warsaw: Exit.

Walther, Joseph B. & Kyle P. D'Addario. 2004. The impacts of emoticons on message interpretation in computer mediated communication. *Social Science Computer Review* 19. 324–347.

Wikström, Peter. 2014. #srynotfunny: Communicative functions of hashtags on Twitter. *SKY Journal of Linguistics* 27. 127–152. http://www.linguistics.fi/julkaisut/SKY2014/Wikstrom.pdf (accessed 20 March 2016).

Xu, Wei, Alan Ritter & Ralph Grishman. 2013. Gathering and generating paraphrases from Twitter with application to normalization. *Building and Using Comparable Corpora (BUCC-2013)* 6. 121–128. https://www.cs.nyu.edu/~xuwei/publications/ACL2013_BUCC.pdf (accessed 20 March 2016).

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

Zappavigna, Michele. 2012. *How we use language to create affiliation on the web*. London: Bloomsbury.

Zipf, George. K. 1949. *Human behavior and the principle of least effort: An Introduction to human ecology*. New York: Hafner.