Finland Twitter English:
Lexical, Grammatical, and Geographical Properties

by

Steven Slone Coats

(Under the direction of William A. Kretzschmar, Jr.)

Abstract

This project investigates the extent of English use in Finland on the Social Media platform Twitter and characterizes the properties of the variety Finland Twitter English. User messages in a novel, 1m–token corpus of Twitter data from Finland were disambiguated for language and language choice correlated with geographical and demographic factors. English user messages were annotated with part-of-speech tags and the principal lexical and grammatical characteristics of Finland Twitter English, including the relationship between feature frequency and gender, determined by comparing aggregate feature frequencies to those of a similarly processed non-Finland Twitter English corpus, using a statistical measure of association and multidimensional techniques. The lexical and grammatical features most characteristic of Finland Twitter English discourse reflect the primarily interactive communicative orientation of the userbase, which utilizes those language features most closely associated with the technological newness of the communication platform in order to establish and negotiate meaning.

Index words: Social Media, Finland, World Englishes, Corpus linguistics, Sociolinguistics, Multidimensional analysis

Finland Twitter English:
Lexical, Grammatical, and Geographical Properties

by

Steven Slone Coats

B.A., Oberlin College, 1996

Mag. Art., Ruprecht–Karls–Universität Heidelberg, Germany, 2007

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2015

Finland Twitter English:

Lexical, Grammatical, and Geographical Properties

by

Steven Slone Coats

Approved:

Major Professor: William A. Kretzschmar, Jr.

Committee: Lewis C. Howe
Christopher Eaket

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

Acknowledgments

I wish to extend thanks to my major professor Dr. Bill Kretzschmar for facilitating my progress in Georgia and Finland and for providing useful feedback on a draft version of this text. Thanks go also to my advisory committee members Drs. Chad Howe and Chris Eaket for feedback on the text and their flexibility in helping to organize the logistical aspects of the dissertation defense from afar. I've benefited from the knowledge and the administrative support of Dr. Jared Klein at UGA. Thanks very much to you all! In Finland I would like to acknowledge Dr. Elise Kärkkäinen, who supported my research and arranged my teaching responsibilities in such a way that allowed me to prepare the thesis. I've always felt supported in this academic endeavor and in all other endeavors by my family. Thanks go to Mum, Peter, and Lauren.

Contents

List of Figures

List of Tables

Chapter 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 informal language use. Online language 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) suggests that it may differ 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 com-

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

munication in academia, business, media and diplomacy (Crystal 2003), but increasingly as a dominant language for online communication in informal and geographically localized communicative contexts, particularly in the European Union (Paolillo 2005, European Commission 2011).

Despite widespread recognition of the prevalence of English in global CMC, there have been relatively few efforts to investigate the extent to which English comprises social media communication in regional or national contexts.[2] Although studies have investigated the use of particular linguistic features in various types of CMC, including Twitter, and the distribution of linguistic features in Twitter language in the United States (e.g. Bamann et al. 2014), few or no studies have analyzed the variety of English used in a social media environment such as Twitter in a specific regional or national context where English is widely used but does not serve as the official language or one of the principal languages of day–to–day communication in the speech community.

The present research characterizes the social media language variety "Twitter Finland English" as a Finland–based variety of informal online English that exhibits convergence features with other global Englishes as they are used online in informal contexts, but which emerges as distinct when investigated on the basis of aggregate lexical and grammatical feature frequencies and in consideration of the demographic characteristics and communicative situations typical of Finland–based users of the Twitter platform. In the context of conceptualizing English as it is used online as part of the expanding circle of users for whom English is not an official language in local communities and whose normative conceptions of appropriate English use, as evidenced by e.g. feature frequencies, may differ from those of users in traditional L1 environments (Kachru 1990), a characterization of Fin-

[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 in Twitter for 41 of 206 countries or territories.

land Twitter English may contribute to our understanding of the ways in which language interacts with the complex forces of globalization.

In a broader sense, the configuration of language feature frequencies and demographic characteristics most typical of Finland Twitter English suggest that the variety may serve as an example of the ways in which language users extract meaning-creating potential from new communication technology at the interface of user functionality and medium constraints (cf. Wikström 2014, Hutchby 2001: 206).

1.1 Organization of the Text

Chapter 2 situates the research project within the context of previous investigations of language variation by briefly reviewing the ways in which communicative parameters can determine the functional contexts of use of different registers or genres of language. A review of previous research demonstrates that CMC has most often been interpreted as a language variety that can be situated at an intermediate position along an axis whose poles represent spoken language and written language, and this may be the result of a configuration of communicative parameters typical for the variety. The parameters of communication most relevant for the study of English as it occurs on Twitter are shown to be similar, but not commensurate, with the set of parameters that are considered to be characteristic of CMC. Previous research into Twitter language has focused on, for example, natural language processing (NLP) methods for the automatic detection of sentiment and content, modeling patterns of user interaction, or investigating properties of Twitter text features such as the hashtag and modeling their geographical distribution.

Chapter 3 describes the techniques and methods used for the collection, storage, and analysis of the Twitter data used in the project. The open–source programming and statistical environment *R* is the primary tool utilized for data manipulation and analyses,

including frequency analyses of the lexical and grammatical items, associated statistical tests, and visualizations of the frequency data. Geo–encoded tweets from Finland were harvested by interacting with the Twitter API. A script prepared in the Python programming language was used to identify tweets suitable for harvesting and store them in a local database. Demographic data for the regions of Finland pertaining to income and educational levels was obtained from the Finnish Statistical Service in order to investigate the interaction of geography, income, and education within Finland and the features of Twitter English. One of the principal goals of the project is to characterize the properties of an emerging variety of Twitter English in Finland and investigate its similarity to other, non–Finland–specific Twitter English. For this purpose, a pre–existing corpus comprising tweets with no geo–coordinates was selected.

After filtering the two corpora for some non–renderable characters, the most likely language represented by each individual tweet was identified using a probabilistic algorithm; English–language tweets were subject to further analysis. Tweets determined to be the obvious result of scripts or automated tweeting tools and tweets sent multiple times were filtered from the corpora. For the Finland tweet data, a portion of the corpus was disambiguated for author gender using automated methods. Finally, each token of each English–language tweet in both the Finland data and the non–Finland data was annotated with a grammatical tag using a Twitter–specific part–of–speech tagger.

Chapter 4 describes the multilingual environment that is Finland Twitter and explores the distribution of the principal languages of the country as they are represented in the geo–encoded Finland Twitter data: Finnish, English, Swedish, and Russian. A principal finding concerns the extent of English use on Twitter in Finland; it is suggested that the decision to use English in Finland on Twitter is a result of the communicative functions most typical of Twitter as well as more broadly–based sociolinguistic and socio–economic factors.

Chapter 5 presents a frequency–based comparison of a large number of lexical and grammatical features for the two corpora: the Finland English Corpus and the Comparison English Corpus. The chapter establishes the fact that many characteristics are shared between the Finland English and Comparison English corpora, due to genre and communicative context considerations. Finland Twitter English can be identified as a distinctive variety, however, on the basis of feature frequencies. For most of the features, the extent to which the feature is differentiated by gender within the gender–disambiguated portion of the Finland English Corpus is considered as well.

Tweet length in characters and in tokens is the first feature considered in the chapter. The distributional profiles of the most frequent lexical features in the Finland English Corpus and Comparison English Corpus are compared with those of the Corpus of Contemporary American English and two components of the British National Corpus representing written and spoken language. The most frequent features of the Finland English Corpus and the Comparison English Corpus indicate that Twitter English, as discourse, is in some ways similar to spoken English.

The lexical items most overrepresented in the Finland data and in the data with no geographical constraints are examined: These provide some insight into culture– and location–specific discourse concerns, but to a certain extent are also artifacts of the data collection procedure. Considering the most "male" and the most "female" lexical types for the gender–disambiguated portion of the Finland English Corpus, however, shows how the emerging variety of Finland Twitter English reflects somewhat different communicative styles; the findings are interpreted (in Section 5.4) in the context of previous research into the relationship between lexical distributions and gender in CMC and other genres.

Lexical items that are used primarily (although not exclusively) as markers of affective orientation, such as profanity or taboo words and emoticons, are considered in some detail. The frequencies of these items in the Finland English data and their interaction with

non–language variables within Finland suggest that they may represent features that are characteristic of the Twitter platform in general. The analysis of emoticons considers a large number of emoticons and related symbol–based lexical types. The distributional profile for a subset of emoticon types is compared between the Finland data, the non-geographically-restricted data, and the findings of a large US–based study of emoticon use (Schnoebelen 2012). The distribution of emoticon types within Finland is examined more closely by correlating emoticon use and demographic variables such as income, educational level, or tweet language.

In Section 5.3, frequencies of grammatical features in the Finland English and Comparison English corpora are examined. Grammatical resources characteristic for certain types of CMC, such as the use of non–standard orthography and expressive lengthening of word types, are compared in the two principal corpora and within the gender–disambiguated section of the Finland English Corpus. The same frequency–based analysis is undertaken for grammatical part–of–speech features (as determined by the tagging algorithm) and for selected grammatical word class features consisting of invariant forms.

In the ensuing discussion in Section 5.4.1, the characteristic properties of Finland Twitter English begin to emerge as related to, but distinct from those of Twitter English in general. The distinctiveness of Finland Twitter English is presented in light of previous findings in the sociolinguistics and computational linguistics literature pertaining to Twitter use. Particularly for those lexical and grammatical features which can be considered most characteristic for the medium of Twitter, Finland Twitter English exhibits shared communicative functions, but a significantly different distributional profile. Despite some previous claims that online language is not substantively different from non–online language, the frequency data for Finland Twitter English suggest that the emergent properties of this specific global English variety have, in some ways, taken on an indexical function which is reflective of the process of technology–driven language change itself. In Section 5.4.2,

some previous work pertaining to language variation and gender, and particularly to online language variation and gender, are introduced. Feature frequencies and sex/gender interact in the Finland English data in ways that reinforce some of the principal findings of previous sociolinguistic research. The dynamics of interaction of sex/gender and lexical or grammatical variables in the Finland data suggest, however, that communicative orientations in Finland Twitter English can also be expressed using patterns of gendered expression that do not conform to the results of previous investigations.

Chapter 6 investigates the relationship between grammatical feature frequencies in the Finland English Corpus and the Comparison English Corpus by subjecting the data to quantitative multivariate analysis. Utilizing the basic conceptual framework of multidimensional analysis proposed by Douglas Biber in a series of influential articles and monographs (1985, 1986, 1988, 1995, 2006), the underlying communicative dimensions of Twitter English are interpreted in terms of pragmatic or functional contexts. By subjecting the individual grammatical part–of–speech features from the tweets of the Finland English Corpus and the Comparison English Corpus to a factor analysis and a principal component analysis, it can be demonstrated that certain grammatical part–of–speech features tend to co–occur in both corpora. Together with the findings from earlier chapters, this allows the Finland English and Comparison English corpora to be distinguished along an axis representing informational versus interactive communicative orientation.

Sequences of lexical and grammatical items (n–grams) are at the center of Chapter 7. The most frequent lexical and grammatical bundles in the Finland English and Comparison English corpora are analyzed according to communicative functions such as stance expression, discourse organization, and referential expression. It can be shown that the most frequent bundles in the corpora differentiate the two corpora in terms of informational versus interactive orientation.

Chapter 8 recapitulates the main findings and summarizes the discussion. The extent of English use in Finland Twitter, some of its sociolinguistic parameters, as well as the characteristic lexical and grammatical features of the variety are delineated in the context of an emerging paradigm for Global English. The frequencies of some lexical and grammatical features of Twitter English, and particularly of Finland Twitter English, are interpreted as representing an example of the ways in which specific modalities of the Twitter user interface take on unexpected communicative functions; in a broader sense, the English typified by the Finland messages indexes developments in a technological environment which continues to evolve rapidly. Some potential research perspectives on Twitter language and related varieties of online language are suggested.

In Appendix A, the code utilized to collect the data and conduct the analyses is presented. The code in *R* is mostly original, although much use of user–developed libraries has been made. The short Python scripts consist of slightly modified code by other authors. In Appendix B, data used for the disambiguation of gender within the Finland English Corpus is presented.

Chapter 2

Approaches to Online Language

Technological changes can gradually shift the frameworks in which communication takes place. Relatively recent developments such as increased access to the internet, the widespread availability of smartphones, and the establishment of global social media platforms have led to an increase in the use of English in online contexts with text as the principal communication channel. As Sebba remarks, "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" (2010: 61) — they are, increasingly, stored digitally on servers and spontaneously delivered anywhere in the world on demand.

In this chapter, some of the situational and textual parameters of communication and their interaction with technology are reviewed. In the history of the study of language, spoken and written language have often been differentiated on the basis of *a priori* ontological considerations, rather than the presence or absence of shared functional and communicative considerations, and are often described from the perspective of paradigmatic conceptions of differences between spoken versus written language. Like other language forms, CMC is constrained by inherent factors pertaining to the creation, transmission, and interpretation of discourse. As has been done for varieties of spoken and written English, Twitter

English can be described by making reference to the configuration of communicative parameters that are manifest in its use and which often correspond to characteristic feature frequencies. As will be shown in Chapters 4, 5 and 6, the specific characteristics of Finland Twitter English suggest that it emerges at the interface of functional considerations with technological and sociolinguistic factors. In order to demonstrate this, it is necessary to briefly consider the some of the communicative parameters that influence the choice of channel and how they are manifest in online language and Twitter.

2.1 Speech, Text, Technology

Theoretical characterizations of the dual nature of language as speech and written code have, in the past, tended to award primacy of form to either the spoken or written mode. As Biber (1988) remarks,

> There has ... been considerable disagreement concerning the need for a linguistic comparison of speech and writing. Historically, academics have regarded writing, in particular literary works, as the true form of language, while speech has been considered to be unstable, degenerate, and not worthy of study. (p. 5)

The status of written language as (semi–) permanent and the restriction of its primary domain of use for much of human history to individuals with education and power may provide a sociolinguistic explanation for the preference of writing to speech and the corresponding relative valuation of the two modes in previous eras.

A recognition of the authenticity of spoken language, particularly in its non-standard and dialectal forms, emerged as a concomitant of the late Enlightenment and Romanticist enthusiasm for folk and oral traditions, particularly as they lent authority to notions of national identity; this interest in spoken language coincided with a rapid rise in literacy

in European societies. In an English-language context, literary interest in traditional patterns of speech is evidenced in such works as Robert Burns' 1786 *Poems, Chiefly in the Scottish dialect*, or by Wordsworth and Coleridge's 1798 *Lyrical Ballads.* One might note in Germany the 1808 collection of folksongs *Des Knaben Wunderhorn* by Brentano and von Armin or the well–known compilation of German oral traditions *Haus- und Kindermärchen* published by the Grimm brothers in 1812. Similar literary works based on popular and peasant language exist in other European languages.

In the context of philology, systematic investigation of spoken language in the 19$^{\text{th}}$ century was motivated primarily by the desire to establish the historical genesis and development of language and language varieties. For example, work by philologists such as Rask, Grimm, Verner and Paul led to the postulation of regular, diachronic sound changes that could help explain the geographical dissemination and differentiation of the Germanic and Indo-Germanic dialects and languages. This "comparative method" in philology was developed further by the following generation of linguists in Germany, the *Junggrammatiker*, who recognized the value of this heuristic for comparative linguistic studies and the attempt to reconstruct language history. Their postulation of the homogeneity and regularity of sound change led to the study of living languages and dialects, ostensibly in order to demonstrate the diatopic regularity of sound change.[1] Most "traditional" philologists of this era considered the systematic study of spoken language to be a useful tool for the reconstruction of linguistic taxonomies, but not an objective in its own right.

This purely historical interest in the status of the spoken language began to change with the rise of dialectology and linguistic geography. German interest in the *Mundarten* culminated in the 19$^{\text{th}}$ century in Wenker's *Sprachatlas des deutschen Reiches* and related

[1]For example, the prominent philologist Hermann Paul claims: "Es ist eingewendet, dass es noch eine andere wissenschaftliche betrachtung der sprache gäbe, als die geschichtliche. Ich muss das in abrede stellen." ("It has been claimed that there is some other scientific approach to language, other than a historical one. I must refute this claim." 1886:19)

projects, whose publication began in 1878; similar endeavors were undertaken in France in the following decades. In England, Joseph Wright and Alexander Ellis began to investigate contemporary spoken language using many of the same methods as had German philologists.[2] Scientific apprehension of spoken language from a physiological perspective was undertaken in the same era in the emerging discipline of phonetics, pioneered in Britain by the works of philological-physiological researchers such as Henry Sweet.

By the 20th century, the conception of the relationship between speech and writing had shifted in the assessment of many philologists. No longer written language, nor language history, but spoken language was now accorded primacy of place. Saussure, for example, remarks that "the linguistic object is not both the written and the spoken forms of words, the spoken forms alone constitute the object" (1959 [1916]: 23f.). Biber cites Edward Sapir's claim that writing is a derivative form of language, amounting to "visual speech symbolism" (1921: 19, cited in Biber 1988: 6). In a much more recent assessment, Mark Aronoff writes that "spoken language is the 'true' language, while written language is an artifact" (1985: 28, cited in Biber 1988: 6).

The nature of the relationship between spoken language and writing (both the creation of written texts and the reading thereof) represents a complex overlap of cognitive, psychological and physical human experience in varying historical, social and situational contexts. Comparisons of speech with writing have therefore been undertaken by attempting to define the shared parameters of the two modalities in terms of broader categories of experience such as location in time and space, relationship of interlocutors, and communicative goal. Contemporary corpus linguistics tends to take a neutral position as to the "primacy" of the

[2] Wright had studied in Heidelberg, as had Henry Sweet, who remarked that not only had the Germans developed contemporary philological methods, but in the 1870s, "... it became too evident that the historical study of English was being rapidly annexed by the Germans, and that English editors would have to abandon all hopes of working up their materials themselves." (1885: v)

spoken or written mode of language; corpora have utilized both transcribed spoken and written components since the 1970s.

As far as CMC forms such as chat or social media messages are concerned, these have typically been postulated to demonstrate partial overlap with the parameters characteristic of typical written modes as well as with spoken language.

2.2 Language Online

Technological change has led to the emergence of new forms of media and possibilities for linguistic interaction in the context of the internet; it has been suggested that online language represents a combination of characteristics of written and spoken language. In this section, some previous research into the properties of online language is described and the contextual parameters which may affect the communicative functionality of Twitter more closely identified.

2.2.1 Computer–Mediated Communication

Much of the early scholarly interest in CMC took place in the context of theoretical interest into its informational properties or investigation of the psychology of communication behavior: Many scholarly articles containing the keywords "computer mediated communication" or its initialism CMC in the decades 1970–1990 appear in academic journals specialized in information science, psychology, behavior, or communication. For example, early research into CMC looked into, for example, the extent to which users utilized the new medium to exchange information, express opinions, develop new ideas, or establish and strengthen social contacts, compared to the existing modes of written (print) communication or spoken interaction (Hiltz and Turoff 1978). Linguistic aspects of CMC, although they were

recognized to exist by social scientists in early studies, do not seem to have been have been a primary focus of CMC research prior to the 1990s.[3]

Some early linguistic research on CMC such as chatrooms has been directed towards the classification of CMC and CMD (Computer–Mediated Discourse), typically by characterizing its properties in terms of the modalities of written and spoken language. Maynor (1994), for examples, suggests that CMC represents "written speech", whereas other analysts have suggested that CMC represents a type of language that is intermediate between spoken and written forms. For example, Tagliamonte and Denis (2008) analyze a large ($>$ 1m words) corpus of instant messaging (IM) communication by a group of American teenagers and compare it to a spoken–language corpus created from interviews with the same subjects. They find that IM language is in many ways (e.g. use of intensifiers or future reference) similar to spoken language, but in other ways (use of deontic modals) more similar to written genres.

There have been various interpretations among researchers of the typological coherence of CMC: Characterizations of internet chat as "conversation" have been disputed, and CMC has been variously characterized in the literature as a "text type", a "form of discourse", a "communicative genre" or a "form of communication" (Dürscheid 2005). Herring (1999) analyzes chat using methods developed by researchers studying face–to–face conversation, proposing that despite its status as an "incoherent medium" in terms of traditional stance– and adjacency–pair analysis, it offers communicators opportunities to exploit technological properties of the medium that result in heightened interactivity.

The meta-label "conceptual orality" has been suggested for chat and other forms of synchronous CMC (Androutsopoulos 2003). Analyses of code-switching between two or more languages in chat rooms have shown evidence for different patterns of code–switching

[3] See e.g. Hiltz and Turoff (1978) or Kochen (1978), who suggests that CMC may represent a "new linguistic entity with its own vocabulary, syntax, and pragmatics" (22, cited in Rice and Love, 1987: 86).

in CMC compared to those most common in face–to–face oral communication (Androutsopoulos and Hinnenkamp 2001).

"Emoticons" – typographical symbols that often represent facial expressions and are used to convey affective or other types of information – are another feature present in many different subgenres of CMC. A common interpretation of the linguistic function of emoticons is that they are meant to compensate for the lack of prosodic and other paralinguistic information in CMC environments (Hentschel 1998).

2.2.2 Textual Parameters of Variation

Some efforts to characterize the linguistic aspects of CMC have considered it as a language variety that can be interpreted according to functional and situational parameters that have been proposed in sociolinguistics, pragmatics and discourse studies. For example, Crystal (2001) extends an interpretation of functional text types originally developed to distinguish among the various genres of written and spoken communication to the analysis of CMC.[4] Beginning with the possible parameters of variation of the text itself, without considering communicative context, Crystal outlines the main parameters of variation for texts. Written language can vary in its graphic presentation, its orthographic features, its grammatical possibilities, its lexis, and its discourse features and functions (2001: 8ff.). The first two parameters apply only to written texts: Graphic variation would comprise the visual presentation of the text material in terms of typography as well as color or format such as newspaper, pamphlet, vellum scroll, computer screen, etc. Orthographic variation – a common feature of some types of CMC – refers to intratextual variability once the principal graphic presentation of a text is established. This includes conventions of writing such as capitalization, standard or non–standard spelling, punctuation variation, and the use

[4] Crystal utilizes the term "variety" to describe specific manifestations of language along the broad range of situational and functional variation.

of e.g. boldface, italics, or other modified visual forms. Several parameters apply to both written and spoken language: Grammatical variation comprises variation in morphosyntax and clause organization. Lexis, another parameter that has been intensively studied in linguistics, comprises the frequency and distributional characteristics of the vocabulary of a (spoken or written) text, whether individual words or multi–word expressions. Discourse features, according to Crystal, are governed by higher–level organizational parameters pertaining to the logical organization of semantic content and its varying pragmatic functions in different texts. The examination of Twitter English in Finland and Twitter English with no specific geographical restrictions shows the extent to which orthographic, grammatical, lexical, and discourse variation differentiates subvarieties.

2.2.3 Situational Parameters of Variation

The situational parameters of communication include the spatial and temporal configuration of the communicative act as well as contextual factors pertaining to interlocutor attitudes, intentions, and goals. These parameters are subject to somewhat different constraints in spoken language, written language, and written varieties such as CMC and Twitter.

2.2.3.1 Time and Space

The relation of interlocutors to one another in space and time can vary; some of the most salient differences between written and spoken language have been reflected in the configurations of interlocutors most common for these modes. Spoken communication typically (and is still most frequently) takes place in real time and with speakers in shared physical space, i.e. in a situation of synchrony and spatial proximity. Physical constraints imposed by human physiology have established these parameters: Intelligibility thresholds for are typically between 30 and 50 dB for speech sounds. The acoustic intensity of av-

erage conversation is approximately 50 dB, and as auditory signals lose 6 dB in intensity per doubling of distance, speech becomes unintelligible with increasing spatial distance (French and Steinberg 1947). Human cognition and memory limitations may constrain the functional coherence of asynchronous communication.[5] Written communication, on the other hand, represents a technology devised in order to overcome the constraint of spatial proximity, and as it developed historically, typically involved asynchrony and spatial distance (although neither condition is necessary in order for written communication to take place). Developments in technology such as the semaphore, telegraphy, telephony, computer networking and the internet have all, to varying degrees, made it possible to further decouple communication from physical and temporal proximity, allowing a broader range of situational parameters for communication. Crystal (2001: 136) suggests that the non–linear nature of asynchronous online communication such as moderated LISTSERV mailing lists has linguistic consequences that result in an increased diversity of forms. The temporal horizon of Twitter communication is not synchronous like CMC genres such as IRC chat, but may be more constrained than those of 1990s–era bulletin boards, LISTSERV communication, or 2000s–era Usenet groups: Yang and Leskovec (2011) find that for Twitter hashtags, a significant proportion of activity occurs within a 128–hour range centered on the hour with the largest number of tweets containing the hashtag. To that extent, Twitter may be considered a semi–synchronous medium. The horizons of temporality of Finnish English user messages and non–geographically–specified English user messages are discussed in Section 5.3. It can be shown that the temporal parameters of Twitter influence the communicative and discourse parameters of the medium by encouraging a more interactive functional orientation of users.

[5]For example, asynchronous communication often necessitates explicit anchoring of temporal deixis due to the non–overlap between the timeframes of the interlocutors. See Chovanec (2014: 33).

2.2.3.2 Other Situational Parameters

A number of other potential situational parameters that circumscribe the range of variation available to interlocutors for a communication event have been proposed. For example, the attitudes of interlocutors or communicative participants towards literacy and towards oral communication may be manifest in the frequencies of particular features they utilize: Someone who places great value on literacy may strive towards lexical variation, whereas someone more concerned with oral communication may pay close attention to prosodic features such as intonation, stress, duration, and volume. The extent to which interlocutors desire to maintain and mark differences in social status can also affect phonetic, phonological and prosodic variation in spoken communicative interaction, and grammatical or lexical variation in written or spoken communication. For social media such as Twitter, text phenomena such as emoticons and expressive lengthening have been interpreted as sociolinguistic variables (Bamman, Eisenstein and Schnoebelen 2014; see Sections 5.2.5 and 5.3.2.)

The relationship of the communicative participants to one another can vary along a number of functional parameters. Extent of interaction, extent of knowledge about other communicative participants, communicative goal and topic, effort required to establish or maintain relationship, and extent of shared cultural or world knowledge all affect the resulting language and thus the distributional profile for all types of features (Biber 1988: 40f., summarizing earlier work by various researchers). The degree of interactivity of communication can vary: Many genres (traditional print genres such as newspapers or magazines) exhibit very limited interactivity, whereas others (face–to–face speech, internet chat, use of online messenger services) are primarily interactive forms of communication.

Crystal identifies a number of internet contexts in which some of these specific situational parameters may influence the type of language used (2001: 12ff.). For example,

within the broader field of CMC, genres such as email may differ significantly (in terms of feature frequency) from, for example, chat language. The communicative goals for email interactions may be different from those of other online genres, for example to establish or maintain professional relations; this is a communicative context much less common in, for example, internet chatrooms.

Electronic bulletin boards or forums may exhibit language very different from the language of online gaming, as it is manifest e.g. in the text–based interaction of players of MUDs ("multi-user dungeons"). Even within a narrowly defined genre such as online bulletin boards that utilize the popular phpBB software, language use will vary considerably along the parameters described above. Informational bulletin boards devoted to personal finance and taxation, for example, can exhibit language very different from bulletin boards devoted to the discussion of popular music or popular internet culture. Spoken interaction in team–based online gaming, for example in the group interface in popular "first-person shooter" franchises such as Half–Life or Battlefield, or utilizing software such as Teamspeak, may exhibit characteristics that are very different from those of other gaming genres or other types of online language.

Situating Twitter discourse along the parameters described above can help to define what is particular to the language variety, but may be difficult, as its spatial, temporal and situational parameters are not homogeneous. Twitter communication is often associated with spatial distance and usually occurs in semi–synchrony, and some subsets of users may share situational parameters such as affective or knowledge orientation, but this is not categorically true for users of the service. The typological classification of CMC on the basis of communicative parameters has proven to be a challenging and somewhat contentious issue (Herring 2007, Squires 2010). Herring (2007), for example, utilizes earlier discourse function classifications in order to propose a faceted CMC classification scheme in which 18 factors pertaining to "medium" and "situation" are used to distinguish between

CMC discourse types, while Squires (2010) disputes the contention that CMC can be classified on the basis of situational parameters or feature frequencies. The utility of typological identification of CMC and comparison with well–studied media such as printed texts or spoken language has been called into question. Squires (2012), for example, in an analysis of apostrophe use and gender in a 9,600 word corpus of IM communication comprising messages by undergraduate students at an American university, suggests that "oversimplified inter–medium comparisons" such as those typified by Tagliamonte and Denis (2008) are less relevant than approaches influenced by social anthropology, in which the "social meaning" of variation is "both constructed and deployed in social interaction"; such an approach can also help to expose language ideologies (p. 295; p. 290).

Attitudes, shared knowledge or communicative goals and other situational parameters can vary for Twitter users interacting in English, whether in Finland or elsewhere. As the platform is theoretically accessible to anyone with internet access, generalizations about shared user attributes in terms of attitudinal or affective orientation may not be possible. If, however, shared communicative orientation can be inferred on the basis of aggregate use of lexical and grammatical features, as suggested by Crystal, it is possible to more narrowly describe Twitter English and particularly Finland Twitter English, as is demonstrated in Chapters 5 and 6.

2.3 Previous Twitter Research

Twitter has become an important resource for communication in online media since its launch in 2007. 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 mid–2014, the site reported more than 600 million registered users and a volume of more than 58 million tweets per day (Statis-

tic Brain 2014). Although the service was initiated in the United States, it has achieved significant global representation: Messages are broadcast in the platform in hundreds of languages from more than 100 countries worldwide (Mocanu et al. 2013). A small but significant proportion of tweets are broadcast with geo–coordinates corresponding to the location of the user (Morstatter et al. 2013). The large volume of the service, its public availability, its status as an emergent technology, and the geographic information associated with messages have resulted in significant interest being directed towards analyses of Twitter data in computer science and various subfields of the social sciences and the humanities in recent years (Boyd 2014).

Language–based studies utilizing Twitter data have approached the material from varying theoretical and practical perspectives. Lexical–item–based sentiment analysis and related types of information extraction have figured prominently in Natural Language Processing (NLP) research in the last decade (e.g. Turney 2002; Wiebe, Wilson and Cardie 2005; Pennebaker et al. 2007; Pang and Lee 2008). The Linguistic Inquiry and Word Count software (Pennebaker et al. 2001, 2007), aggregates frequency data for closed word classes such as articles and personal pronouns as well as lexical items that index sociality, negative emotion or positive emotion, and has been used widely in applied psychology settings as well as CMC analysis (e.g. Kapadzic and Herring 2011). The software has been utilized to study Twitter data in predictive contexts: For example, Tumasjan et al. (2010) use the software to predict German election outcomes. In a similar vein, Batra and Rao (2010) conduct sentiment analysis on named entities in a large Twitter dataset in order to gauge sentiment towards persons, places, and organizations. Bollen, Mao and Zeng (2011) assess the viability of using sentiment analysis on Twitter data to help predict stock market movements. The potential usefulness of Twitter as a source of real–time data for statistical inference in judging consumer or political sentiment has motivated much research activity.

Modeling social networks on Twitter in terms of dialogic participation and patterns of user interaction has informed a number of research approaches, represented by studies such as Honeycutt and Henning (2009), who 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 the @ symbol in the user message (typically used to indicate user names) correlates with user interactivity, and suggest that microblogging may facilitate collaboration. Wu, Hofman, Mason and Watts (2011), Murthy (2011) or Murthy (2013) represent similar, user-interactivity based studies of Twitter data.

Kelly (2009) shows that approximately 37% of Twitter posts can be considered conversational or dialogic, with the remainder comprising news reports, announcements, advertising or other one-way communication types. Ritter, Cherry and Dolan note that dialogue on Twitter tends to be extremely short in terms of turn sequences, with dialogic pairs of one tweet and one response comprising almost 70% of a dataset of Twitter dialogues (2010: 173).

Other analyses using Twitter data have investigated text features characteristic of Twitter user messages. Yang and Leskovec (2011) show that temporal patterns of hashtag use in Twitter are relatively short-term (several days), and reflect news events as they are reported in other online media. Zappavinga (2011) investigates the use of hashtags not only as explicit topic markers, but as markers of subjective user affiliation or orientation. Wikström (2014) undertakes a qualitative analysis of the diversity of pragmatic and communicative functions associated with the hashtag in Twitter messages, suggesting that hashtag use on Twitter may represent an example of how user interaction with technological interfaces can prompt the emergence of unexpected communicative functions. Eisenstein et al. (2012) utilize geo-encoded Twitter data to explore the emergence of Twitter "dialects" or geographically localized use of particular word forms in the United States, and model the dynamics of lexical diffusion based on geographical considerations and demographic

parameters. Their findings reinforce a century of dialectological field work in which geographical distance and community size have been shown to be the strongest correlates of the diffusion of new language forms,[6] but they also suggest that parameters such as racial or ethnic identity may correlate more strongly than expected with patterns of lexical diffusion. Alis and Lim (2013) utilize a similar approach in an analysis of tweet length; their findings reinforce those of Eisenstein et al. (2012). Gonçalves and Sánchez (2014) collect a large database of geo-encoded Spanish-language tweets and use k-means clustering to disambiguate macro-varieties of Spanish based on the frequencies of a set of synonyms that have typically had complementary geographical distributions in Spain and the Americas. They propose a new classification of two "super-dialects" of Spanish that reflect urban versus rural association rather than Europe versus the Americas.

Some sociolinguistic studies utilizing Twitter data have been undertaken. For example, Schnoebelen (2011, 2012) investigates the expression of affective content in tweets, particularly through the use of emoticons and emoticon co-occurrence with lexical items. He finds that emoticons are used as an expressive resource that pattern with linguistic and non-linguistic variables such as social network and gender. Bamann, Eisenstein and Schnoebelen (2014) focus on gender variation in Twitter messages. They conduct cluster analyses of word forms used by 14,000 Twitter users in the United States and note that users typically cluster into groups that share topical interests or writing styles, many of which have a strongly gendered component.

An attempt to situate Twitter along the parameters described above may be difficult. Twitter communication is global and thus often characterized by spatial distance between interlocutors, and it can occur with asynchrony. Twitter users often provide explicit topic indicators to their messages in the form of hashtags. As such, the parameter settings spatial distance and asynchronicity are not necessarily absolute for Twitter.

[6]See e.g. Chambers and Trudgill (1998), Sonderegger (1983) or Kretzschmar (2009).

Chapter 3

Data Collection and Processing

In this chapter, the methodology employed for the collection, manipulation, and analysis of the data is described. The tools and programming languages used for data collection and processing are briefly introduced and some characteristics of the Twitter API are discussed. Additional demographical data from official Finnish sources are introduced. The two principal corpora created, the Finland English Corpus and the Comparison English Corpus, are introduced, along with some summary statistics. Then, details concerning the tokenization procedure, aspects of character encoding, the disambiguation of tweet language for the corpora, the removal of automated tweets, and the creation of a smaller Finland English Corpus in which gender is disambiguated are discussed. Finally, the procedure used to annotate the data with grammatical information is described.

3.1 Data Collection and Processing Framework

A frequency–based linguistic analysis is necessarily reliant on data processing software. A large number of computerized tools have been developed for different platforms that facilitate the collection and analysis of digitized texts. Automated tools for calculating word and word class frequencies include various dedicated and online concordance and colloca-

tion software, such as the Mike Scott's WordSmith Tools (Scott 2012 [1996]) or Lawrence Anthony's AntConc (2014 and earlier versions). These tools can calculate type–token statistics for words and n–grams in single texts as well as provide some co–occurrence information in the form of e.g. a Mutual Information statistic, but offer limited support for statistical analysis and data visualization. Variation analysis of language has often been performed with the Varbrul or GoldVarb software packages; these allow regression modeling of multiple variables (Cedergren and Sankoff 1974; Sankoff, Tagliamote and Smith 2005) but not other multivariate approaches such as ANOVA, factor analysis, or principal component analysis.

3.1.1 R–based Tools

The open-source programming platform for statistical analysis *R* provides access to user-created libraries that have been developed for specific data-processing and analysis functionality in most scientific fields (*R* Core Team, 2014). The core functionality of the software allows many types of multivariate analysis to be undertaken; in addition, libraries have been specifically developed for lexical statistics, corpus analysis, and text mining, such as *languageR* (Baayen 2013), *tm* (Feinerer, Hornik and Mayer 2008; Feinerer and Hornik 2014), *zipfR* (Evert and Baroni 2007), and others.[1] The flexibility and extensibility of the *R* core package makes it a good choice for frequency–based corpus analysis.

Visualization and geographical visualization of data in *R* can be facilitated with libraries such as *ggplot2* (Wickham 2009) and the mapping libraries *rgdal* (Bivand, Keitt and Rowlingson 2014), *sp* (Pebesma and Bivand 2005), *mapproj* (McIlroy 2014), and *maptools* (Bivand and Lewin-Koh 2014), among others. For this project, processing and analysis steps were achieved using the functionality provided in the sore *R* packages and autonomous

[1]For a general overview of some approaches to data mining and text analysis using *R* and various packages, see Zhao (2012).

code, presented in the Appendix (Section A), as well as code developed in the above *R* packages. As the project was undertaken primarily using *R*, some packages for direct access to and manipulation of Twitter data were evaluated.

Twitter–oriented packages in R include *twitteR*[2] (Gentry 2012) or *xml*[3] (Lang 2012; the package is focused on metadata and tag processing but has some functionality for dealing with Twitter data); these are libraries of functions written in *R* that automate some of the steps for interaction with the Twitter Application Protocol Interface (API) and/or allow manipulation of Twitter data.

The *R* packages *twitteR* and *xml* proved to be not well suited to the research plan for two main reasons. First, although the packages utilize the Twitter API, their interaction is limited to the REST API, not the Streaming API. The difference between the two interfaces is that REST API involves a single request to the Twitter server (with various search parameters), whereas the Streaming API opens a continuous connection in a single process that continually delivers data to the user until the process is closed.[4]

The REST API typically provides access to a maximum of ~2000 tweets that have been created within two weeks of the request time. If the parameters of the request are highly specific (i.e. tweets that are geotagged and that originate from Finland), much less data will be available for a single request. Although it would be theoretically possible to compile a corpus by sending requests to the REST API in individual sessions at regular intervals, then removing duplicate tweets, utilizing the Streaming API is much more straightforward: the end user simply sends a single request to the API, then lets the data accumulate in a database or local file (in our case a text file) until the process is closed.

At the time the data was collected, Jeff Gentry (the author of the *twitteR* package) had not yet implemented code that would allow storage of tweets collected via *twitteR*

[2]https://github.com/geoffjentry/twitteR
[3]http://www.omegahat.org/RSXML/
[4]For an overview of the differences between the APIs see https://dev.twitter.com/streaming/overview.

in a relational database management system such as MySQL; this functionality, which would allow online access and manipulation of the dataset and thus facilitate processing of data from the Twitter REST API, was partially implemented in early 2014.[5] The lack of Streaming API access, however, outweighs the utility of relational database support for our purposes.

The second reason for not utilizing the *twitteR* package is the limited native support for Unicode (UTF-8) characters provided by *R* when it is run under system locales for which Unicode is not the native encoding scheme (i.e. Windows). Twitter user messages are encoded in Unicode UTF-8 characters. *R* packages (such as *twitteR*) can handle Unicode characters in Windows after appropriate processing steps have been undertaken, but will not correctly display some Unicode characters if they are retrieved directly from the Twitter API without being subject prior conversion. This fact can make it difficult to maintain character set integrity when data is retrieved from the Twitter API and converted between multiple formats for subsequent processing (manual inspection, language detection and tagging, part-of-speech tagging) using other software, as it has been for this project. More details pertaining to character set and encoding issues are discussed below in Section 3.4.2

3.1.2 Python

Functionality has been developed in Python, an object–oriented programming language, for statistical analysis (van Rossum and Drake 2006; Ascher et al. 2001). NLP–specific functionality has been developed in the form of the Natural Lanuage Processing Toolkit module, or nltk (Bird, Loper and Klein 2009).[6] Python offers full support for Unicode (UTF-8), which makes it an appropriate choice for interacting with the Twitter API. The present research utilized a Python script to access the Twitter Streaming API and store

[5] http://www.r-bloggers.com/twitter-now-supports-database-persistence/
[6] http://nltk.org/

data in a local file. The data–harvesting code was based on the Python module *Tweepy* (Roesslein 2013), which provides functionality for interacting with the Twitter Streaming API.[7]

3.1.3 The Twitter Streaming API

Twitter 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 (the 20% "Gardenhose" and the 100% "Firehose") on a commercial basis only to data resellers or "market-tested leaders" which "make Twitter more valuable to businesses, encourage their use of Twitter, and bring Twitter to new users" (Twitter 2015).

Access limitation does not necessarily pose a practical problem for the compilation of a tweet corpus, given that the platform broadcasts more than 400 million messages per day, as of late 2014. However, if one is interested only in a specific subset of tweets, much less data is potentially available to someone attempting to compile a sufficiently large dataset. For example, a highly specific filter interacting with the Twitter Streaming API, in which only those tweets are returned that match a specific term in metadata and include geographical coordinates, would result in a significantly smaller potential data set of interest. If that data is further limited by restricting the set of tag-matching, geo-encoded tweets to a small geographical range (such as latitude and longitude coordinates that encompass Finland), and then only 1% of matching tweets are available to the researcher, far fewer messages are available. Although Social Media and Twitter are relatively popular in Finland, with the country's Prime Minister (as of early 2015) Alexander Stubb being an enthusiastic user of the service, it is not among the countries with the highest per–capita use of the platform (Mocanu et al. 2013).

[7] https://github.com/tweepy/tweepy

These factors resulted in a relatively slow rate of accumulation of data for the compilation of the Finland English Corpus. The corpus data was collected from March until of May 2013, typically during the hours of ~3 PM – 10AM the following day (GMT). Data was output by the Python script into a local .txt file. At several times server connection or other errors resulted in an interruption of data delivery; in these cases the script was simply re-started and the data added to the existing file.

3.1.4 Additional Demographic Data for Finland

For the Finland Corpus, data from the Finnish National Statistical Office was used to investigate some correlations between geographical region in Finland and English use in tweets according to demographic variables such as income and education level. In addition to GDP per capita of the region under consideration, several educational variables were examined: First, the proportion of the regional population enrolled in upper-secondary education in 2012; then several statistics derived from that proportion, such as the proportion of the population of the region passing the school–leaving exam and the regional proportions of the total population granted qualifications granted in universities, in polytechnic universities, or in vocational skills exams for the year 2012 (Official Statistics of Finland 2013a, 2013b, 2013c, 2013d, 2103e).

3.2 Finland English Corpus

The Python script used to collect tweets from the Twitter Streaming API can be found in Section A.1 of the Appendix. The script imports relevant modules into python, including *tweepy* and a module for rendering Unicode characters. The script then authenticates the user in order to access the Twitter API. The script is designed to collect the following six components of Twitter messages ("tweets"): the "status", i.e. the user message of

maximum 140 characters; the author name; the time of posting; the source of the post (software used to create the tweet, such as android app, iphone app, web interface, etc.); the posting coordinates as latitude and longitude values; and the place name. The sixth component, place, reflects changes in the way the Twitter API communicates geographical information that were implemented in the update from version 1.1 to version 1.2. Tweepy was written for the older (1.1) version of the API, and thus the script could only access latitude/longitude coordinates, not other place information. Latitude and longitude were narrowed down by applying a filter box to the set of coordinates that the script harvests. For Finland tweets, a box with the coordinates 60–70°N and 21–30°E was implemented; this encompasses the borders of Finland except for a very small portion of the Åland islands and Northern Karelia. Only tweets originating within these coordinates were collected. The coordinates also encompass portions of Sweden, Norway and Russia (notably including the densely populated St. Petersburg region). GIS (geographic information system) processing at a later stage removed all tweets from the data set that did not originate from within the borders of Finland (see Section 4.2).

As noted above, most tweets do not have geographical coordinates: This is an option that can be accessed by the user via his or her Twitter interface. Those tweets with geographical coordinates typically originate from smartphone devices, for which the default settings of Twitter apps such as Twitter for iPhone typically communicate geo–location, unless the user turns it off. In terms of data collection, this resulted in approximately 3500 tweets per day that were harvested during the two months of data collection. The corpus of all tweets harvested from Finland is referred to in the following as the **Finland Corpus**, a subset of this corpus, processed to eliminate non–English tweets, is referred to as the **Finland English Corpus**. Basic summary statistics are shown in Table 3.1.

3.3 Comparison English Corpus

A principal motivation for the current study is to investigate lexical, grammatical, and discourse properties of English as it is used in Social Media (Twitter) in Finland, both to shed light on how English is used in Social Media and other online domains in general, but also to investigate what role English plays on Social Media in Finland and how it is similar (or dissimilar) to English used in the same domain elsewhere, in terms of quantifiable lexical, grammatical, or syntactic features. For this reason, a contrastive approach has been adopted, in which a Twitter corpus compiled of messages with no geographical restrictions was subject to the same processing and filtering steps. The lexical and grammatical features of the two corpora, their distributions, and the communicative functions which these features may represent can then be compared, contrasted, and analyzed together. A number of Twitter corpora have been created for purposes of linguistic or other analysis. Yang and Leskovec (2011) created a large Twitter database to investigate properties of temporal diffusion of lexical items and topical content. Eisenstein et al. (2012) created a large corpus of US Twitter messages in order to research the diffusion of novel word types. Other researchers associated with the Carnegie-Mellon University Natural Language Processing lab have created or utilized similar corpora (Gimpel et al. 2011; Bamman, Eisenstein and Schnoebelen 2014). For this research, a Twitter corpus compiled by researchers at Texas A&M University in 2009 was utilized. The corpus was available as a freely downloadable file through the data aggregation and big data company infochimps.com in 2013. However, a change in Twitter policy in 2013 has led to copyright claims and requests that have resulted in Twitter corpora no longer being made publicly available; formerly publicly available Twitter corpora such as the corpus compiled at Stanford and used by Yang and Leskovec (2011) or the Texas A&M Corpus are no longer accessible. Twitter policy does not prohibit the compilation of corpora, but the company regards its data as proprietary

content. The Comparison English Corpus has no geographical limitations; it was compiled from the unfiltered Twitter Streaming API. As with the Finland data, **Comparison Corpus** refers in the following to the unfiltered, multilingual corpus, and **Comparison English Corpus** refers to a subset filtered to exclude non-English tweets. Summary data is provided in Table 3.1.

TABLE 3.1: Finland Corpus and Comparison Corpus Summary Data

Corpus	Tweets	Tokens	Types
Finland Corpus	101,612	1,039,865	251,606
Finland English Corpus	32,916	436,954	53,863
Comparison Corpus	305,310	3,361,444	467,493
Comparison English Corpus	181,861	2,864,798	155,043

According to Twitter and the Finnish state broadcasting company, in early 2013 there were 63,000 active Twitter users in Finland (Yleisradio 2013). Approximately 102,000 of the Finland Corpus tweets originated from within the borders of Finland. If the rough estimate that Twitter users send, on average, $^1/_2$ tweet per day (Sass 2011), holds true, the collected data corresponds to approximately 5% of the tweets from Finland in this time period.

3.4 Data Processing

Various approaches to the creation of corpora from web language have been suggested. Biemann et al. (2013) note that corpus construction procedures can vary according to to the goals of the research project; a corpus created to e.g. serve as a representative sample of text discourse functions in English writing from a certain domain would be compiled and processed differently than a corpus constructed for automatic translation purposes. They note that in "empirically-oriented theoretical linguistics, carefully selected sampling procedures and non-destructive cleaning is important, while for many tasks in

computational linguistics and language technology, aggressive cleaning is fundamental" (23). The same basic principles are valid for the assembling of a Twitter corpus. The Finland Corpus is not intended for specific NLP tasks such as machine translation or syntactic treebank modeling. As such, an attempt was made to limit the extent to which the data was subject to manipulations prior to analysis. For example, stemming of word tokens, described below in Section 3.4.1, was not undertaken. Some removal of Unicode characters, however, was necessary. This is described below in Section 3.4.2. Filtering the tweets from the defined geographical range for only those tweets that originated from within the borders of Finland was accomplished with GIS packages in *R*. The procedure is described in Section 4.2.

3.4.1 Tokenization

In corpus-based studies that utilize NLP techniques, feature counts can be somewhat dependent on the data processing procedures applied prior to analysis. For example, the tokenization of a text, or converting long character strings consisting of letter, punctuation, or other characters and blank spaces into separate tokens, can be undertaken in various ways. A common simple tokenization procedure consists of converting word forms to lower case (to achieve equivalency between the distinct forms *The* and *the*, for example), stripping a text of all punctuation, and counting the resulting types. While removing punctuation may make subsequent processing easier (taggers have sometimes had poor accuracy in interpreting word forms that contain various types of punctuation), the method is problematic for three main reasons. The first reason is that removing punctuation may cause the frequencies of some types to be under– or overestimated. Contracted word forms such as *we're*, for example, would be counted as the single word form *were*, and the resulting analysis would overestimate the ratio of verbal to pronominal forms. There are NLP stemmers that can recognize morphological variants of underlying verb froms,

including contractions, and these can be used to attain more accurate frequencies.[8] Due to the inherent noisiness of Twitter data, however, use of tools such as taggers which do not remove punctuation may represent a better choice.

The second reason pertains to the role of punctuation itself in written language. In corpus–based studies, and in descriptive text linguistics in general, relatively little attention was paid to the frequencies of different punctuation types for much of the 20th century, causing Gleason (1965, qtd. in Nunberg 1990: 9) to remark that "very little descriptive data [is available] on how the English, or any other, punctuation system is actually used. The large volume of published material which is available is predominantly normative." The historical lack of interest in punctuation may reflect the development of a conception of language in which primacy of place is awarded to spoken, rather than written language. Punctuation, according to such a conception, merely represents the transcription of prosodic elements with little semantic or grammatical content. As Albert Markwardt (1942, qtd. in Nunberg 1990: 11) remarks, punctuation indicates "those elements of speech which cannot be conveniently set down on paper: chiefly pause, pitch, and stress." Descriptive grammars of English have typically devoted little to no space to the role of punctuation, leaving the topic to language mavens and other arbiters of prescriptive norms.[9] Punctuation has increasingly been considered an integral part of the written language, and linguistic considerations of punctuation, including from a corpus–based or computational perspective, reflect this.[10] For CMC, realtively few studies of punctuation types have been undertaken.

[8] A well–known stemmer is the Lancaster stemmer, described in Paice (1990). See http://www.comp.lancs.ac.uk/computing/research/stemming/index.htm. The popular Python package *nltk* offers implementations of various stemmers including the Porter stemmer (Porter 1980) and the Lancaster stemmer.

[9] Truss (2003) can be considered exemplary.

[10] For example, Nunberg (1990); Jones (1996) is a corpus–based study utilizing much of the theoretical treatment of Nunberg. Say and Akman (1997) give an overview of computational work on punctuation. Bayraktar, Say and Akman (1998) undertake a functional discourse analysis of the role of the comma in a text corpus. Say and Akman (1998) is a similar analysis of the em–dash using corpus methods. Biber et al. (1999) consider the role of the comma as a disambiguator of restrictive vs. non–restrictive post–modification (pp. 602–658), but otherwise do not consider punctuation.

Squires (2012) examines the role of the apostrophe in an IM corpus and finds females more likely to use this punctuation type.

There is good reason for including punctuation types in a corpus–based study of type distributions, particularly for data from social media such as Twitter. Punctuation serves an important function as an element of meaning, grammatical or discourse organization in written language. A more practical reason has to do with the nature of internet and Twitter language. The present analysis focuses on the frequencies of dictionary word forms, including common contracted forms, but also on non–standard orthographical forms and pseudo-word forms that are highly characteristic for Twitter and other CMC, such as multi–character emoticons. Removing punctuation would effectively remove those types that are among the most characteristic for the language being studied. In order to preserve these types, the data collected from the Twitter API and the Comparison Corpus has not been stemmed or subject to removal of punctuation. Punctuation items are thus present in the lists of most frequent lexical items presented in the following sections.

3.4.2 Text Encoding Issues

A rather significant issue for Natural Language Processing has to do with the encoding of text characters and their interconvertibility between applications. Converting text from one program or operating system to another can introduce errors when codeset uniformity or compatibility has not been ensured. For text that is presented graphically, the phenomenon has come to be known as "Mojibake"[11] from the Japanese word for "character transformation", which typically results when attempting to display characters encoded under a different encoding scheme, such as Japanese encodings. It will produce garbled text for those characters that are not in the local character set. In Table 3.2, common Nordic-language characters such as *ä* and *ö* are not represented correctly.

[11] An adequate overview is provided at http://en.wikipedia.org/wiki/Mojibake

Table 3.2: Finland Corpus Messages Showing "Mojibake" Due to UTF-8 and CP-1252 Incompatibility

	Text
1	@AinoEmine EmmÃ¤Ã¤huomannu mitÃ¤Ã¤kirjoitusvihreitÃ¤. #sokea
2	@unipala oukkidoukki! Teen parhaani˜!
3	Ja toi oli belgialainen gull joka ei ees tiedÃ¤tuksuu. Wth tl.
4	Milo enjoying his evening walk... @ RÃ¶hkÃ¶mukka http://t.co/sdFCrGuBtX
5	Tulin mun lÃ¶kÃ¤reihin excuse me RT "@nakatsu_fin: @ninnumon tuksu on niin hawt oh god"

Table 3.3: Finland Corpus messages with Correct Rendering

	Text
1	@AinoEmine Emmää huomannu mitään kirjoitusvihreitä. #sokea
2	@unipala oukkidoukki! Teen parhaani˜!
3	Ja toi oli belgialainen gull joka ei ees tiedä tuksuu. Wth tl.
4	Milo enjoying his evening walk... @ Röhkömukka http://t.co/sdFCrGuBtX
5	Tulin mun lökäreihin excuse me RT "@nakatsu_fin: @ninnumon tuksu on niin hawt oh god"

For the processing of Twitter data, several encoding-based issues can arise. The Twitter API utilizes Unicode 8-bit (UTF-8) encoding, but the native encoding scheme for Windows (and thus for programs running under Windows) is not UTF-8, but rather the 16-bit encoding scheme CP-1252, based on ISO 8859-1. This is also true for programs that run under Windows (including *R*), although both Windows and *R* offer some degree of Unicode support via various scripting fixes.

Our data, prior to processing and filtering steps, includes graphemes from tens of different writing systems from almost 100 different languages, all of which was encoded in UTF-8. For this reason, and also because the analysis involves steps in which the data was processed using programs with UTF-8 internal support or run from Unix locales (notably the part-of-speech tagging step but also the language identification step), the sensible approach was to maintain Unicode (UTF-8) encoding when possible at all intermediate processing steps.

The processing and analysis in *R*, however, relied on the utilization of the local Windows character sets. For this reason, accurate conversion of characters needed to be checked and maintained in all intermediary steps. Conversion of ASCII characters between encodings is generally unproblematic, but problems can arise when Latin characters with diacritics (such as *ä, ö, å, ü,* etc.) are used or characters from other alphabets.

In addition, some non-language UTF-8 characters are not (yet) supported in the default Unicode libraries of many programs, including those developed under Unix locales. This is commonly the case for UTF-8 characters that have been introduced relatively recently into the Unicode coding scheme, such as pictograph characters referred to as "emojis" that have become popular in Social Media.[12] These UTF-8 emoji characters, introduced into the Unicode standard in 2010, are mostly in the byte range U+1F300 and above.[13] They are not supported by Windows 7 or *R*, and the full range of characters is, as of November 2014, also not supported by most common text processing programs, including those that have extensive UTF-8 support, such as Notepad++. Even recent versions of browsers such as Firefox 31.2 and Chrome 38 do not yet support all UTF-8 characters in the U+1F300 range. These characters are, however, supported in the Android OS operating system that runs on many mobile devices, and are integrated in the Twitter environment. Some users make frequent use of these emojis, and they are present in our data set as it was compiled.

To deal with these issues, the following steps were taken. First, Unicode is the emerging default standard for text encoding. Its range and extendibility are suitable for capturing a very large number of characters and graphemes from the world's writing systems. As the default encoding of the file connection in the *R* implementation used for the project is not Unicode, but rather that of the System locale (i.e. Windows 7), care was taken to

[12]Emojis (see http://en.wikipedia.org/wiki/Emoji) are sometimes also called "Emoticons". For the purposes of this project, "Emoticon" will refer to multi-character emotional content symbols created using the extended ASCII character set, such as :-) or :D and also to emojis.

[13] See http://www.utf8-chartable.de/unicode-utf8-table.pl

systematically convert all files into UTF-8 or UTF-8 compatible encodings upon reading them as input and prior to writing them to file. This was done mainly by utilizing function command parameters in *R* for data input and output, as well as native encoding-conversion functions within the language itself (notably `iconv`)[14] when appropriate.

Secondly, verification of the encoding integrity of the data was done with the Notepad++ program, which supports Unicode (UTF-8) text as well as many other encoding paradigms, including the default Windows character sets. The Python and Unix-based text processing and analysis steps were less problematic, as Unicode is the native character set for Unix systems and Unicode compatibility is well developed in Python.

Third, all emojis in the byte range U+1F300 and above were removed utilizing regex expressions. Although the loss of information is unfortunate, leaving emojis in the corpora leads to errors in the intermediate steps that ultimately result in errors when calculating the frequencies of lexemes, parts of speech, and other linguistic features using *R*. An unsupported emoji described as "GRINNING CAT FACE WITH SMILING EYES"[15] in the Unicode scheme, for example, is represented variously as a series of question marks in ASCII (??), as box symbols (□), or as Unicode mojibake (ï¿½ï¿½) if the local character set does not support the character, or it will appear in a text file as the corresponding Unicode code sequence (<U+1F638>), as a UTF8-hexadecimal byte sequence (\xF0\x9F\x98\xB8), as a 16-byte sequence under Windows (\x3D\xD8\x38\xDE) or as ASCII HTML entities (&#55357;&#56888;), depending on the settings and command parameters in the software and system being used.

The *R* processing steps for the corpora involve typical string manipulations such as substitutions, conversions between upper– and lower cases, string splitting, and concatenation of substrings, at the most basic level. Unsupported characters in mojibake or byte

[14]https://stat.ethz.ch/R-manual/R-devel/library/base/html/iconv.html
[15]http://www.unicode.org/charts/PDF/U1F600.pdf

sequences typically break even simple code and result in errors – for example, Unicode mojibake symbols cause errors when strings are put in lower case. Hex sequences can be difficult to distinguish from emoticons or written text without writing code exceptions for every possible hexadecimal sequence, and html entities cannot always be automatically disambiguated from authentic use of constituent symbols such as the ampersand and the semicolon.

The relative proportion of the total corpora content consisting of emojis was quite small in our data prior to their removal: using a regular expression[16] to tabulate the number of non-printable characters in the raw data set showed that of the total 13.6 million characters of the original Finland tweet dataset, 7878 (approximately 0.057%) consisted of unprintable characters; these include characters not supported in the Windows character sets as well as recent UTF-8 characters such as emojis. The comparison data set makes even less frequent use of non-printable characters: approximately 0.029% of the characters in the raw data were unprintable. This may reflect the fact that the comparison data was compiled in 2009, before Unicode characters in the range U+1F300 and above came into widespread use. However, an examination of the unprintable characters in question for the comparison data reveals that many of them are the ♫ character, which suggests that the non-printable characters in the Comparison Corpus data are the result of a UTF-8 to Windows character conversion issue, rather than the presence of recent Unicode characters such as emojis. In any case, the regular expressions used to remove the emojis are based on hexadecimal UTF-8 sequences, not the criteria of POSIX-printability. This means that in effect, even less than 0.05% or 0.03% of the characters were removed from the two corpora in this processing step. Although this filtering makes a specific analysis of some of the grammatical properties of emojis impossible, based on the low percentage of emojis

[16] The regular expression [^[:print:]]+ in Notepad++ captures non-printable characters according to POSIX.1-2008 (see http://pubs.opengroup.org/onlinepubs/9699919799/); that is, all characters that are not ASCII and also do not exist in the current system locale.

in the data, their removal does not significantly affect the ensuing analysis of lexical and grammatical types.

For an analysis of English use in Finnish Twitter messages, the question presents itself: Why not analyze only those tweets that consist of only ASCII characters, as standard English is written only using those characters? As has been suggested, doing so would eliminate much data of linguistic interest, particularly language mixing phenomena or English-language tweets containing proper nouns with non-English characters (such as many place names in Finland, e.g. Jyväskylä), as well as many of the emoticons and other extended types based on UTF-8 characters that figure prominently in Twitter discourse.

The Comparison Corpus data was subjected to some additional processing steps. Examination of the text material brought some obvious data format conversion or encoding artifacts to light. Examples include character sequences such as *<* and *>*: these are HTML/ISO Latin-1 entity codes that represent the symbols < and >.[17] Such errors were corrected by using regular expressions to substitute the Unicode characters for the HTML entity equivalent.

Finally, any encoding artifacts that did escape notice during file conversions were filtered out using a catalogue of regular expression search/replace functions in the text editor *Notepad++*. *Notepad++* was frequently used to visually examine the text between steps to ensure processing was proceeding smoothly.

3.4.3 Language Identification and Classification

The language of each individual tweet used in the study was identified using the probabilistic language identification tool *langid.py* in a Python shell (Lui and Baldwin 2012). The code utilizes a Naïve Bayes Classification algorithm based on variable length n-grams. The langid.py classifier is distributed with an embedded language identification model which

[17]See, e.g. http://www.utexas.edu/learn/html/spchar.html.

consists of documents in 97 different languages drawn from the domains of government documents, software documentation, news reports, an online encyclopedia, and an "internet crawl" (27).

Lui and Baldwin compare the language detection accuracy of langid.py with that of three other language classification tools: *LangDetect*,[18] *TextCat*,[19] and *CLD*.[20] The authors find that langid.py is somewhat more accurate than the other tools, attaining classification accuracy of between 90 and 99 percent on, for example, news articles, European Union government documents, or Wikipedia articles.

Lui and Baldwin note that the new text domain of Twitter and microblogging services "presents a significant challenge for automatic language identification, due to much shorter 'documents' to be classified, and is compounded by the lack of language-labelled in-domain data for training and validation⋯Despite the recently published results on language identification of microblog systems, there is no dedicated off-the-shelf system to perform the task" (29).

They test the accuracy of langid.py using two microblogging datasets that have been used in previous research on language identification, and report between 88 and 94 percent correct classification on two Twitter datasets in five and six European languages (29).

3.4.4 Problems with Automatic Language Classification Tools

Test runs undertaken on the dataset showed that both langid.py and CLD provide generally accurate language classification, although there were some errors. CLD classifies a text's language in two steps. In the first step, the possible languages for a classification are determined based on the number of languages that use the characters found in the text. In a second step, word-internal 4-gram character sequences from the text under consider-

[18] http://code.google.com/p/language-detection/
[19] http://odur.let.rug.nl/vannoord/TextCat/
[20] http://code.google.com/p/chromium-compact-language-detector/

ation are compared with probability tables from the model; this is used to determine the probability of the text being in a certain language.

Thus, if a text is comprised only of ASCII characters, it can possibly be in any of the world's languages which utilize ASCII characters. However, classification problems arise when texts include characters which are conventionally restricted to use in specific languages. Twitter and other online communication often show extensive use of language mixture and emoticons, many of which contain Unicode symbols that were developed to represent non-Western character glyphs. For example, the Unicode range 0D00–0D7F contains characters for the Malayalam language, a Dravidian language of South Asia. If Malayalam characters from this range are included (as elements in emoticons, for example) in very short English or other-language tweets, the classifier may incorrectly characterize the language of the tweet as Malayalam.

The langid.py tool takes a somewhat different approach. Instead of filtering the set of possible languages for classification based on the character set of the message, n-gram frequencies of variable length byte sequences are compared with probability tables derived from n-gram frequencies of known models.[21]

Neither of the methods is ideal for handling language mixtures. Short texts consisting of language mixtures can sometimes be classified as a third language, especially if they contain tokens with non-standard orthography such as usernames or urls. For example, consider a short text from the Finland data such as *@shaidafaiqi haha faktiskt*, consisting of a username, an English non-dictionary word, and a Swedish word, meaning approximately "@shaidafaiqi haha, that is true" in English. For the CLD classifer, whose classifications are based on 4-gram character sequences, the individual word tokens in such a message do

[21] The "multinomial event model" used by Lui and Baldwin does not consider only the dependent probabilities of distinct word type or text character sequences as they are rendered by the system interface, but rather looks at the byte sequence of the text to be classified and calculates 1-to-4-grams for a mixture of byte lengths (25). As some non-western Unicode characters are encoded with up to four bytes, this method may introduce error if the classification is based on n-grams with short byte length.

not have enough 4-gram sequences that are of high probability in Swedish or English to allow classification as either of those languages. The classifier will assign the tweet to a language based on the relative probabilities of sequences such as *haid*, *aida*, *idaf*, *dafa*, or *afai* from *@shaidafaiqi* as well as *akti*, *ktis*, and *tisk* from *faktiskt*. The assigned language is likely to be incorrect; in this case Albanian, due to the relative frequencies of the above character 4-grams in that language.[22]

TABLE 3.4: Inaccurate Language Attributions by Langid.py (Probabilistic Indication of the Accuracy of the Assignation is Provided

	Text	Lang	Prob
1	@seanhackbarth An Ole Miss fan perhaps?	de	0.5173
2	@katiewitt Good luck!	de	0.4721
3	@tehduh Indeed!	de	0.2036

For short texts, langid.py sometimes can also provide inaccurate language classification, due to the presence of proper nouns and non-standard orthography or capitalization. For example, consider the tweet messages in Table 3.4, incorrectly classified as German. The classification is due to the presence of proper nouns (usernames) and the non-standard capitalization of *an, good*, or *indeed*; when the sentences are analyzed with these words written in lower case, they are correctly classified as English.[23]

Inaccurate classifications do not present an insurmountable problem for the automatic disambiguation of English tweets, however, for two main reasons. First, there are relatively few false positives for English, due to its character set of 26 one-byte Unicode characters. There are, however, many false positives for tweets in languages that utilize multiple-byte non-Western characters but also contain one-byte characters. For this project, as we use the langid.py classifier to identify English tweets, and not e.g. Hindi or Oriya tweets, there is some risk of missing tweets that are actually in English but have been misclassified as

[22]The username suggests the user may be of Albanian extraction.

[23]In standard German orthography, the first character of certain word classes is always capitalized.

some other language. There is relatively little risk, however, of including tweets in Farsi or Arabic that have been misclassified as English.

Secondly, the langid.py classifier provides a probabilistic indication of the accuracy of the language assignation, which tends to be rather low for extremely short tweets that have non-standard orthography or capitalization (as in Table 3.4). For this study, only those tweets that were shown to have a probabilistic value of greater than 0.6 in the assignation of English as the language of the tweet are considered in the data sets. Manual examination of the data showed that this value is likely to include all incontestably English tweets as well as a fair number of tweets that contain English elements but exhibit language mixing phenomena of linguistic interest. The procedure also has the advantage of filtering out a large number of tweets that consist of only one or two tokens and are thus of limited linguistic interest for many potential language features; such tweets are inevitably assigned low probabilistic accuracy values.

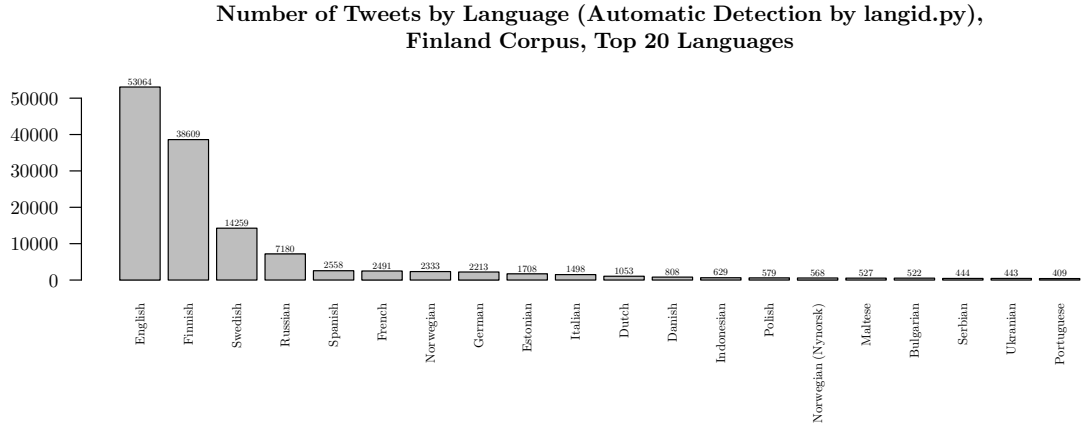

FIGURE 3.1: Languages Detected in the Finland Corpus: 20 Languages with the Highest Number of Tweets (Prior to Probabilistic Accuracy Filtering)

A total of 94 languages (many of which represent classification errors) were detected prior to subsequent filtering based on probabilistic accuracy values. Figure 3.1 shows the number of tweets per language for the 20 languages with the highest number of tweets from

within the defined geographical boundaries of 60°–70°N and 21°–30°E, prior to accuracy filtering. The numbers of Indonesian, Maltese, and Bulgarian tweets likely represent automatic classification errors: the "Bulgarian" tweets are mainly extremely short tweets with Russian words in non–standard orthography and other, non-Russian elements. The tweets classified as Maltese and Indonesian are mainly short texts with non-dictionary words such as usernames or non-standard orthography in Finnish, English, or other languages.

The fact that English is the most represented language in the data is immediately apparent. The relative prominence of English in Twitter in Finland reflects its status as the primary global language as well as social and communicative factors that are discussed in more detail in Chapter 4.

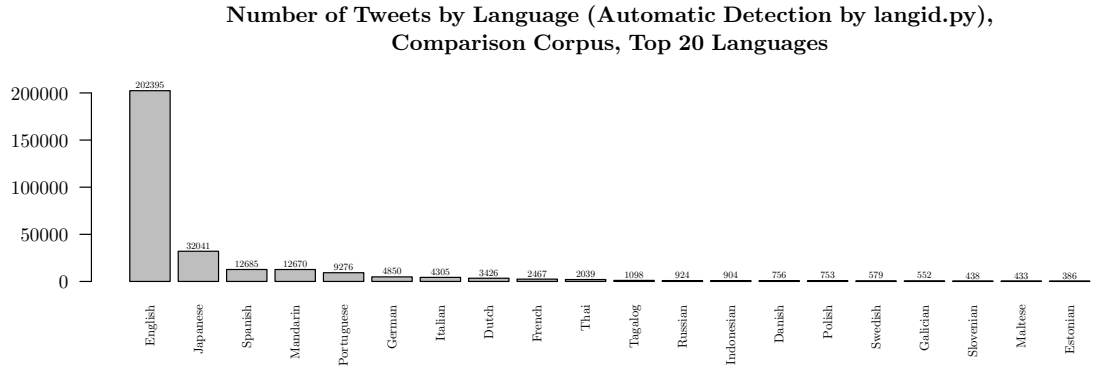

FIGURE 3.2: Languages Detected in the Comparison Corpus: 20 Languages with the Highest Number of Tweets

In the Comparison Corpus, English is by far the most detected language (Figure 3.2). The other detected languages are, for the most part, languages with large numbers of native speakers that are used in relatively developed societies. Some languages with large numbers of speakers, such as Arabic, are underrepresented in the Comparison data. This may reflect socioeconomic factors and Twitter adoption rates at the time of compilation (2008–9). Seshagiri (2014) asserts that as of 2008, more than 80% of Twitter user messages

were in English.[24] In general, the data in the Comparison Corpus can be considered representative of language use in Twitter in 2008–9, the time of compilation.

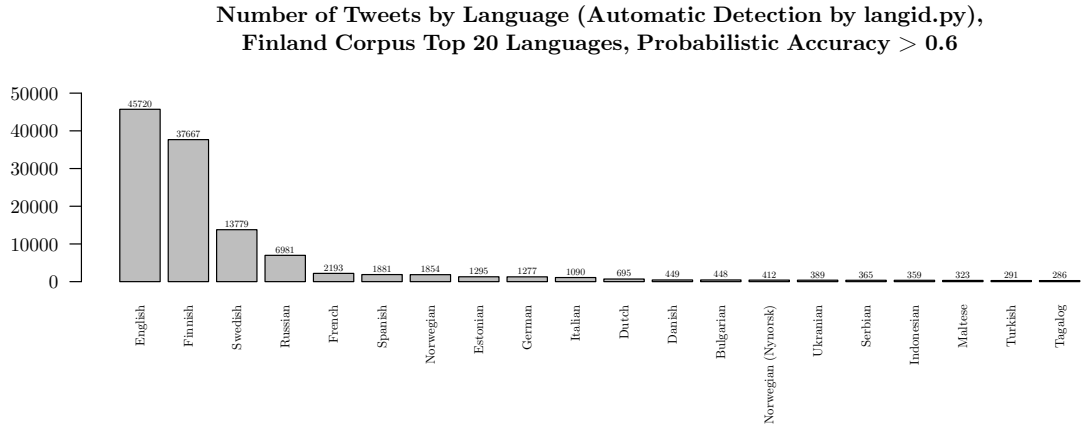

FIGURE 3.3: Languages Detected in the Finland Corpus: 20 Languages with the Highest Number of Tweets (after Probabilistic Accuracy Filtering)

In Figure 3.3 the language distribution for the Finland Corpus is shown after filtering the tweet messages for only those whose language probabilistic accuracy as determined by langid.py is greater than 0.6. As can be seen, the relative number of tweets classified as languages whose use may be unusual for Northern Europe, such as Indonesian or Maltese, has dropped compared to Figure 3.1.

The Twitter data was subject to further processing to identify those tweets that originated not from within the pre–defined latitude/longitude boundaries, but from within the borders of Finland.[25] The Finnish, English, Swedish and Russian tweets identified with a probability of greater than 0.6 and that also originate from within the borders of Finland comprise 84.7% of the data. For this reason, it is these languages that will be considered in Section 4.2 below, concerned with the distribution of tweets by language in Finland.

For the Comparison Corpus, the English tweets with ≥ 0.6 probabilistic values were used in the analysis as the Comparison English Corpus. Lexical and grammatical features

[24]The data is based on a corpus compiled by the social media company *Gnip*, but no descriptive statistics as to the content of the corpus or the methodology used to identify tweet language are provided.

[25]The procedure is described in Section 4.2 and the corresponding code is found in the Appendix, A.2.

from the Finland English Corpus are compared with those of the Comparison English Corpus in Sections 5.2 and 5.3.

3.4.5 Filtering for Automated Tweets

An additional data processing step consisted of filtering the English data in the corpora in order to remove tweet user messages that resulted from automated scripts or from users re-sending tweets a large number of times. A significant proportion of Twitter data comprises tweets that are sent out multiple times by an individual user or texts that are generated entirely by automated scripts (Kelly 2009). As many of these tweets may disappear unnoticed into archival server storage, and do not comprise dialogic or conversational discourse (Ritter, Cherry, and Dolan 2010) an attempt was made to remove them prior to further analysis. Recognition of automated tweets in the corpus was accomplished via various means. Metadata tags in the "source" variable in the original tweet corresponding to apps that generate automated tweets such as Foursquare and Endomondo were used to identify some of the data to be removed. As automated tweets exhibit stable collocational word patterns, creating a frequency table of word 3– and 4–grams allowed the most common automated or massively re–sent tweets to be identified and then removed with regular expressions. As described above, these corresponded to messages sent multiple times by individual users (often attempts to capture the attention of popular culture celebrities active on Twitter) as well as automated tweets generated by scripts. The script–generated tweets included, for example, a tweet sent every hour by the Helsinki Lutheran Cathedral with content indicating the number of times the church bell has struck, automated weather tweets sent by the home weather station software Sandaysoft Cumulus, or automated tweets sent by Instagram users alerting followers to page activity (Table 3.5).

Some manual analysis was necessary in order to ensure that only duplicate tweets or fully automated tweets were removed. For example, all of the tweets from the source

Table 3.5: Automated Tweet Examples from the Finland English Corpus

	Text	User	Time	Source	Coordinates
1	BONGI BONGALAINEN	tuomiokirkko	2013-05-05 11:00:02	Tuomiokirkko Kello	[24.95027304, 60.17036263]
2	Wind 0.5 m/s WNW. Barometer 1023.1 mb, Rising slowly. Temperature -2.5 °C. Rain today 0.0 mm. Humidity 84%	Borgaensis	2013-03-28 21:40:14	Sandaysoft Cumulus	[25.7775, 60.42]
3	Just posted a photo @ Porvoon tuomiokirkko / Borgå domkyrka http://t.co/QDCou9DNBo	juhakatila	2013-04-14 11:22:33	Instagram	[25.65786123, 60.3971832]

"Sandaysoft Cumulus" represent automated weather reports, but only a relatively small proportion of the tweets from the source "Instagram" represent automatically generated updates about photo uploads — most Instagram tweets contain user–generated text as well as several word tokens and a url link to an image file appended by Instagram. For this reason filtering was done by first identifying sources that send automated tweets, then using n–gram matching to remove obvious multiple or automated tweets. Massively re–sent user messages such as entreaties for the attention of the entertainer Justin Bieber were identified on the basis of n–gram frequencies and all instances but one removed from the data if the corresponding word n–grams appeared among the twenty most frequent word 3– and 4–grams (see chapter 7).

Although this method may not have removed all automated and re–sent tweets from the data, systematically applying these filtering steps resulted in corpora that better represent the communicative interactivity that lies at the heart of social media and is manifest in the lexical and grammatical frequency distributions as they are analyzed in the following chapters.

3.4.6 Filtering for Author Gender (Finland English Corpus)

Using a list of the 200 most common names for Finnish females and the 231 most common names for Finnish males, the Finland English Corpus was filtered into a smaller dataset

of gender-tagged tweets. To assign gender to individual users, the usernames of Twitter users in our data whose tweets originated from within the borders of Finland were tagged according to whether or not they included character strings corresponding to the most common male or female Finnish names.[26] This method is not infallible, as it does not consider those Finnish users who have usernames that do not include a common Finnish name, and there is no guarantee that the presence of a male– or female–name character string within a username will always correspond to the gender of the user. However, as Bamann, Schnoebelen and Eisenstein (2014) remark, large–scale analysis is based on aggregate trends, not on individual exceptions. It is assumed that for our data, the overwhelming majority of usernames that include common Finnish names correspond to persons with the gender to which that name is most commonly assigned.

For the set of 7,362 tweets identified in this manner (83,562 words), there are 628 distinct male users and 507 distinct female users. Females in this set are more active Twitter users: Tweets by females outnumber those of males by a ratio of 2.27 to 1. Males, however, are slightly more prolix: The female/male ratio for total number of tokens is 2.04 to 1, slightly lower than the female/male tweet ratio. This corresponds to a significant difference in average tweet length for males of 13.76 tokens and females of 12.3 tokens.[27]

For selected lexical and grammatical features, the comparison of Finland English and Comparison English Corpora in the following sections was extended to the category of gender within the Finland English Corpus. However, the relatively small number of gendered tweets (and the lack of certainty in the assignation of gender to particular usernames) should be kept in mind in the following analysis.

[26]The list of names, from http://www.sci.fi/ kajun/finns/ can be found in the Appendix, Section B. The method was inspired by Bamann, Schnoebelen and Eisenstein 2014.

[27]A t-test of population means gives a value of t = 8.09, p-value = $7.96e^{-16}$.

3.4.7 Grammatical Feature Tagging

Automatic assignation of part–of–speech tags to tokens from CMC texts can be problematic due to its "noisiness", i.e. the widespread use of features such as non–standard sentence structure and orthography and punctuation symbols (emoticons); this is also true for Twitter messages. Widely used taggers such as the CLAWS tagger (Leech, Garside and Bryant 1994) or the Stanford PoS tagger (Toutanova et al. 2003) exhibit poor performance on Twitter data, with tag accuracies of approximately 80%, compared to accuracies of 97% for text types such as news articles (Ritter et al. 2011). Various taggers have been developed for the annotation of CMC and Twitter data. Derczynski et al. (2013) report tag accuracies of 88.7% on a Twitter dataset. Owoputi et al. (2013) report accuracies of 93% for the Carnegie Mellon University Twitter Tagger. The CMU Twitter Tagger was used for the annotation of the Finland English and Comparison English data with grammatical part–of–speech tags.

3.4.7.1 Twitter–NLP Carnegie Mellon Tagger

The Twitter tagger used for the material from the Finland English Corpus and the Comparison English Corpus was developed by the Carnegie Mellon University NLP research unit and is described in Gimpel et al. (2011) and Owoputi et al. (2013). The tagger was run from a Cygwin Linux shell on a Windows 7 system.[28]

The tagger tokenizes a text and probabilistically assigns a part of speech tag according to a model based on manually tagged texts. For Twitter user messages, a standard tokenization procedure in which non–Western characters are removed from a text prior to tagging could remove significant information, as tweets often include language mixtures, character sequences with punctuation or non–Western characters such as emoticons, or url

[28]Cygwin (https://cygwin.com/index.html) is a set of tools that provides a POSIX environment for Microsoft Windows systems; making it possible to run Linux-based scripts on Windows machines.

addresses. The Twitter tagger is sensitive to punctuation, numerals, non–Western Unicode characters, and tokens that consist of combinations thereof, typically assigning such tokens specific punctuation tags (Owoputi et al. 2013: 1). The probabilistic tag assignation model is based mainly on the token–tag assignments of the Penn Treebank, a large manually annotated corpus comprised primarily of a marked–up version of the Brown Corpus and an annotated corpus consisting of articles from the Wall Street Journal (Marcus, Santorini, and Marcinkiewicz 1993), although the authors note that they depart from the Penn Treebank convention for the assignation of some tags where they note "inconsistencies in the annotation" (Gimpel, Schneider and O' Connor 2013: 2). The CMU Twitter Tagger model has been augmented with additional data meant to assist in recognition of urls, emoticons, and proper nouns such as personal, place, celebrity and video game names (Owoputi et al. 2013: 2). Hierarchical word clustering of a large (847 million tokens) corpus provided data to improve performance of the model (3).

The probabilistic algorithm used to assign tags is based on a a Markov chain implementation of the log–likelihood of token transition probabilities in the training data.[29] Owoputi et. al describe how the parameters were trained for the model using the Penn Treebank and additional data, based on a log–linear optimization algorithm (9). The authors report the accuracy of the tagger to be between 90% and 93% in trial runs on various Twitter and chat corpora.

3.4.7.2 Annotation: Grammatical Feature Tags Used

The tags used to markup the Finland English and Comparison English corpora consist of most of the Penn Treebank tags, as well as additional tags used by the CMU Twitter Tagger for url addresses, hashtags, usernames, and retweets (Gimpel, Schneider and O'Connor 2013). The tagset is shown in Table 3.6.

[29]For a brief discussion of the log–likelihood association measure, see Section 5.2.2.

TABLE 3.6: Tagset Used in the Research

No.	Tag	Description	No.	Tag	Description
1	-LRB-	Left-hand bracket	24	NNP	Proper noun, singular
2	-RRB-	Right-hand bracket	25	NNPS	Proper noun, plural
3	``	Quotation mark (“)	26	PDT	Predeterminer
4	,	Comma	27	POS	Possessive ending
5	.	Period (. *?* *!*)	28	PRP	Personal pronoun
6	:	Punctuation *(: ; ... + - = < > / [ ] ˜)*	29	PRP$	Possessive pronoun
7	HT	Hashtag	30	RB	Adverb
8	RT	Retweet	31	RBR	Adverb, comparative
9	URL	Universal Resource Locator	32	RBS	Adverb, superlative
10	USR	Username (preceded by @)	33	RP	Particle
11	CC	Coordinating conjunction	34	SYM	Symbol
12	CD	Cardinal number	35	TO	*to*
13	DT	Determiner	36	UH	Interjection
14	EX	Existential *there*	37	VB	Verb, base form
15	FW	Foreign word	38	VBD	Verb, past tense
16	IN	Preposition or subordinating conjunction	39	VBG	Verb, gerund or present participle
17	JJ	Adjective	40	VBN	Verb, past participle
18	JJR	Adjective, comparative	41	VBP	Verb, non-3rd person singular present
19	JJS	Adjective, superlative	42	VBZ	Verb, 3rd person singular present
20	LS	List item marker	43	WDT	Wh-determiner
21	MD	Modal	44	WP	Wh-pronoun
22	NN	Noun, singular or mass	45	WP$	Possessive Wh-pronoun
23	NNS	Noun, plural	46	WRB	Wh-adverb

Of the 46 tags in the Penn Treebank tagset, 37 are used in the analysis of the Finland English and Comparison English data. The tags *LS* (for list item markers), *PDT* (for predeterminers such as *both* in *both the boys*), *POS* (possessive endings with apostrophe–s), and *WP$* (the Wh-possesive pronoun *whose*) were not applied by the CMU Twitter Tagger. Lists of items are uncommon in the data due to the 140–character length constraint. Predeterminers are tagged as determiners (*DT*) in our data. Possessive endings are not treated separately, as word tokens in the corpus are not stemmed (which can lead to information loss). However, as the ensuing analysis is not primarily focused on morphological considerations, it was felt that the non–use of this tag does not present a significant prob-

lem.[30] *Whose*, in the Finland English and Comparison English data, receives the pronoun, the Wh–pronoun, or the determiner tag, depending on syntactic context.

After manual inspection and preliminary statistical analysis of the data, it was determined that five additional tags from the tagset are either applied extremely infrequently by the CMU Twitter Tagger or represent features that may have limited communicative range and are thus less interesting from the standpoint of a linguistic discourse analysis. These tags are retained in the data, but their frequencies are not considered in the analysis sections or as variables in the multi-dimensional analysis chapter (see Chapter 6).

Frequencies of the tags *FW* (foreign word), *SYM* (symbol), and *NNPS* (plural proper noun) were close to zero in both the Finland English and Comparison English corpora.[31] The tags -*LRB*- and -*RRB*-, for left– and right–hand parentheses, were found in a preliminary multi-dimensional analysis run to cluster strongly with other punctuation tags such as the tag for colons, semi-colons and other punctuation or the tag for periods, question marks, and exclamation marks. As the present study is not focused specifically on an analysis of variation in punctuation, and the parentheses feature frequencies contribute little to the interpretation of potential discourse properties of Twitter language as it differs according to geographical provenance, they are not discussed in the analysis.

[30]The data could be stemmed prior to tagging for closer examination of e.g. variation in the syntax of possession by comparing frequencies of prepositional attribution versus genitive–s forms; structural differences between English and Finnish may result in language interference phenomena. However, this would be the subject for a future investigation.

[31]The CMU Twitter Tagger assigns relatively few *FW* tags due to the risk of miscategorizing orthographical variation as non–English words. The *SYM* tag was extremely infrequent in our data set due to the pre–processing removal of most Unicode symbols, described above. The *NNPS* tag is assigned very infrequently by the CMU Twitter Tagger.

Chapter 4

Distribution of Geo–encoded Tweets by Language and Geography

Prior to investigating the lexical and grammatical properties of Finland Twitter English, the extent to which English is represented in Twitter user messages originating from Finland was investigated. This was done using the set of 101,612 geo–tagged tweets that originated from within the borders of Finland in all languages. As described and shown above in Table 3.1, 32,196 of these tweets were determined to be in English.

4.1 Extent of Geo–encoded Data

The Python Twitter API script collected for each tweet the following six types of information: (1) Text of the tweet, (2) Author (username), (3) Date and time of creation (4) Source, and (5) Latitude/longitude coordinates.[1] Fields 1, 2, 3, and 5 are self-explanatory. Field 4 provides information about the software or client application used to create the tweet. The 20 most frequent sources, accounting for 94.3% of all the Finnish Tweets, are shown in Figure 4.1. The Twitter for iPhone and Twitter for Android apps were the two

[1]The code can be found in the Appendix, A.

most frequent sources for geo–encoded tweets, suggesting that a significant proportion of the Finland data is from smartphone users.

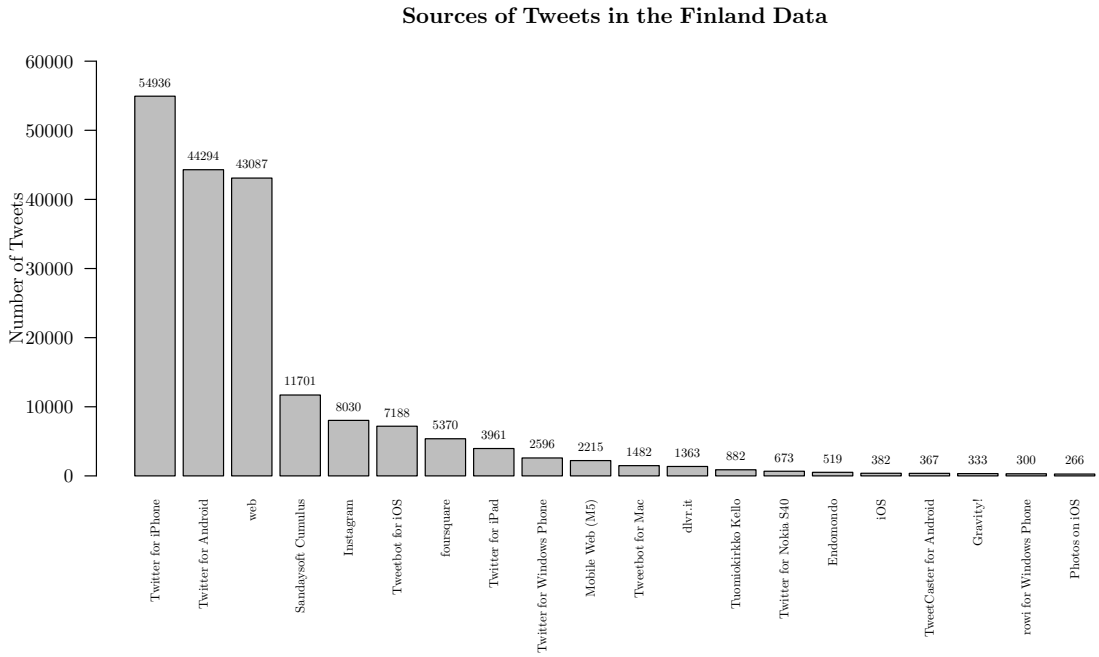

FIGURE 4.1: 20 Most Frequent Sources of Tweets

4.2 Geographical Distribution of Tweets in the Finland Corpus

Data for the Finland maps was downloaded from open source GIS files provided by GADM, the database of Global Administrative Areas.[2] The map data was incorporated into the *R* framework primarily by utilizing the tools provided in the *rgdal* package (Bivand, Keitt and Rowlinson 2014). Visualization was achieved by creating the appropriate code and utilizing the plotting facilities in the *ggplot2* package (Wickham 2009).

[2] http://www.gadm.org/

Of the tweets initially collected by the Python script for the Finland dataset, only 141,253 proved to have latitude/longitude coordinates, despite the initial filtering parameter having been defined to only return tweets from the Twitter Streaming API with geo-coordinates. Of these ∼141k messages, 138,875 had latitude and longitude coordinates that corresponded to the parameters encoded in our Python collection script. 101,612 of those 138,875 originated within the borders of Finland; the other 37,263 tweets originated from the adjacent parts of Russia, Sweden and Norway that fall within the latitude/longitude box of 60–70 degrees North and 21–30 degrees East (see Figure 4.2).

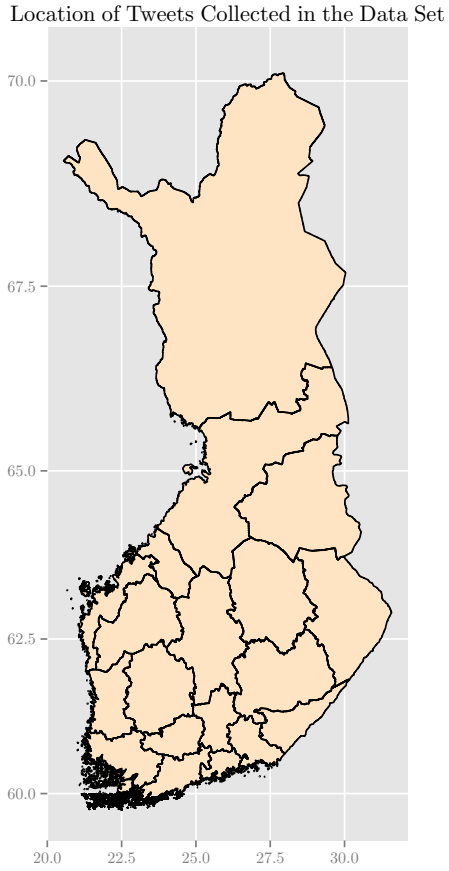

FIGURE 4.2: Location of Tweets Collected by Python Script

Why were so many tweets with no geo-coordinates returned by the Twitter Streaming API? Of 62,521 tweets that did not originate from the pre-defined geographical area (ac-

cording to the data provided by Twitter), the majority of these (51,922 tweets) had the character string "None" as the entry for the latitude/longitude field instead of a pair of latitude/longitude coordinates.

These "None"–source tweets were not generated by the default settings of common mobile apps such as Twitter for iPhone or Twitter for Android that automatically generate geo-coordinates using triangulation of mobile phone signals; rather, the source of the majority of these tweets (83%), according to the data, was listed as "Web". "Tweetbot for iOS", and "Tweetbot for Mac"-sourced tweets comprised a further 16% of these messages.

User profiles in Social Media often contain inaccurate information about location when users are prompted to provide the data themselves (Hecht, Hong, Suh and Chi 2011). In this case, the uniformity of the datum in the geo-coordinates field ("None") points to a faulty configuration in the settings of the Twitter web client or in the settings or code of the third–party app "Tweetbot" as the likely source of the faulty geo-coding information. It may be the case that the API simply delivered all tweets for which any value at all was entered in the latitude/longitude field and that correctly configured Web clients do not provide data for this field.

The other 11,437 tweets that did not satisfy the latitude/longitude parameters defined in the collection script listed geolocations that were clearly errors. These tweets had geo-coordinates corresponding to a wide variety of places, ranging from Bulgaria to Mongolia. A cursory examination of this subset showed that there was frequently no correspondence between the geo–encoded location and the language of the tweet; i.e. Indonesian–language tweets were frequently tagged with geographical coordinates corresponding to Bulgaria. All of these tweets listed the "source" as "Mobile Web M5" (i.e. an HTML5–compliant software application developed for using websites on mobile devices). Without further knowledge of the individual user settings and server–side parameters in use when the tweets were generated, it is impossible to determine the source of these errors, although it seems likely

that they also resulted from faulty settings in client–side code or user settings. In any case, these wayward tweets were not considered at further stages in the analysis.

Figure 4.2 provides an overview of the geographical distribution of the 138,875 Finland Corpus tweets with legitimate geographical coordinates within the designated latitude and longitude range, prior to additional filtering and processing steps. Unsurprisingly, the tweets are clustered in regions of relatively high population density, such as the capital region of Helsinki and the urbanized St. Petersburg–Vyborg corridor. Sparsely populated areas such Lapland and the eastern half of Northern Karelia show few or no tweets.

4.3 Distribution of Tweets in 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 the GADM GIS files.[3]

It should be noted that the Finnish government sometimes re–organizes the subnational units of territorial and administrative organization; major reorganizations were undertaken in 1999 and 2007, and there are continual adjustments of the boundaries of municipal and regional administrative districts. A consequence of this is that the units of description of open–source GIS datasets are sometimes not equivalent when sub–national level Finland data is acquired. In our data set, the GIS files break down the mainland Finnish territory (excluding Åland, which is administratively and politically independent) into 20 subnational units which correspond to the 19 Finnish *maakunnat*, or regions, with the addition

[3]The corresponding code can be found in Appendix A; it makes use of functions developed in the *R* packages *rgdal* (Bivand, Keitt and Rowlingson 2014), *sp* (Bivand, Pebesma and Gomez-Rubio 2013), *maps* (Becker, Wilks, Brownrigg, and Minka 2014), and *maptools* (Bivand and Lewin–Koh 2014).

of an additional *maakunta*. Our GIS files include the region "Eastern Uusimaa", which was incorporated into Uusimaa by a decision of the Finnish government in January 2011.

As the GIS files (and the corresponding tweet distribution) are based on a geography that includes Eastern Uusimaa, and some of the analysis below relies on statistical data for the regions of Finland published after the dissolution of Eastern Uusimaa, when necessary, demographic statistics such as GDP growth or educational level derived from the neighboring region of Uusimaa have been used for Eastern Uusimaa. The demographic characteristics of the former Eastern Uusimaa area are unlikely to be radically different from those of Uusimaa; the area is near the capital and many people commute from its towns to Helsinki in Uusimaa for employment and access to services.

Additionally, the GIS files encode three distinct territories for the region of "Päijänne Tavastia", but corresponding polygon vectors (i.e. borders) for the three territories are not included in the files provided by GADM. This oversight presumably has to do with the changing status of the region; in recent years some municipalities in the region have been re–assigned to different administrative units.

Slightly more than 100,000 tweets originated from within the borders of Finland; filtering to remove automated tweets removed about 6000 tweets from the data. Figure 4.3 shows how the 93,451 tweets that originated within the borders of Finland were distributed in the mainland regions of the country. The region with the most tweets is overwhelmingly Uusimaa, the area centered on Helsinki with a significant proportion of the country's population. The least Twitter activity is in North Karelia, with only 38 tweets.[4] The neighboring region of Kainuu also exhibits relatively limited Twitter activity, with 373 tweets within the data collection time period.

[4]It should be recalled that the Twitter Streaming API provides default access to only 1% of tweets. It is fair to assume that the actual number of tweets originating in the specified regions during the data collection period is larger.

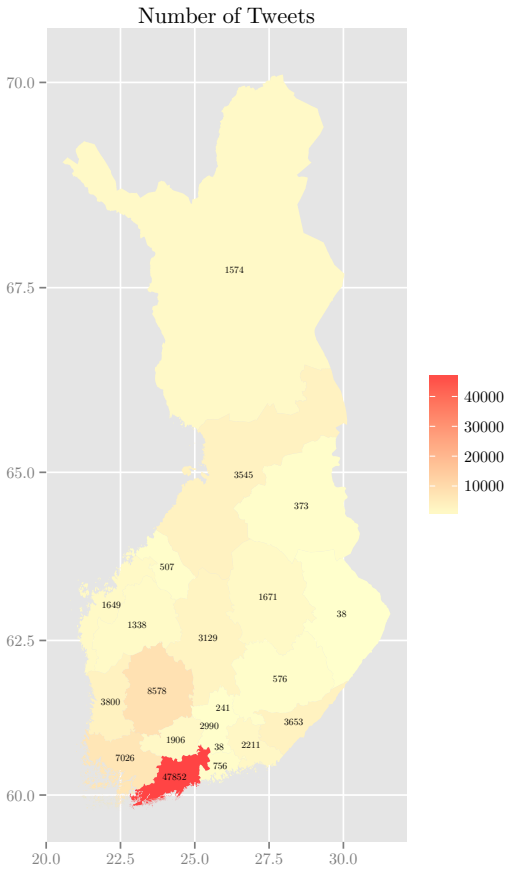

FIGURE 4.3: Number of Tweets per Province

Normalized for population as number of tweets per thousand inhabitants per region, the data show a similar trend. Twitter use is more common in the more densely populated areas of Finland along the Gulf of Finland, as seen in Figure 4.4. For our data, the lowest per capita use of Twitter was in North Karelia, followed by Southern Savonia and Kainuu. Uusimaa had the highest rate of use, at more than 30 tweets per one thousand inhabitants in our data, followed by South Karelia, then the southwestern regions of Pirkanmaa, Satakunta, and Finland Proper.

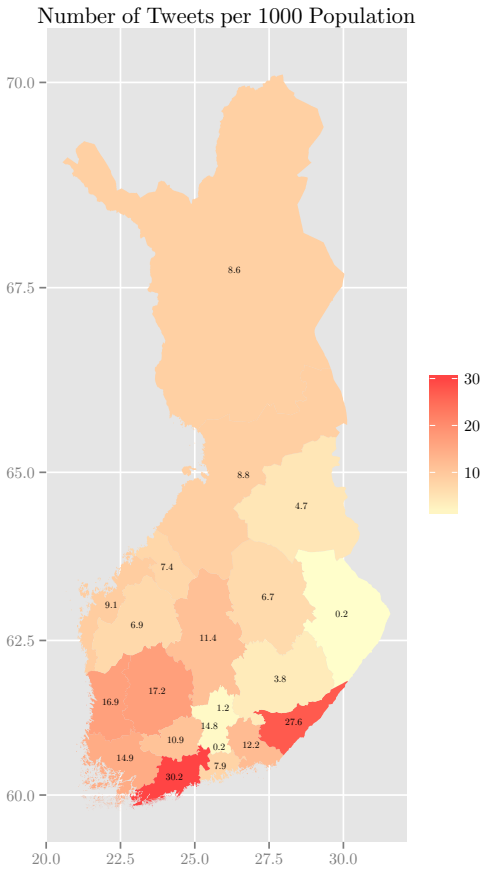

FIGURE 4.4: Number of Tweets per 1000 Population

Prior to the filtering of automated tweets described above in Section 3.4.5, an initial mapping showed an extremely high per capita tweet count for Eastern Uusimaa. More perplexing was the fact that an astounding 90.1% of 5338 tweets from Eastern Uusimaa were in English. A closer examination of the data, however, revealed that the vast majority of these user messages were automated tweets sent out periodically by a script associated with a weather reporting and meteorology app. Systematically removing obviously automated tweets resulted in the language breakdown described above in Section 3.4.3 for the tweets

originating from within the borders of Finland and resulted in a proportionate tweet count for Eastern Uusimaa comared to the rest of the country (Figure 4.4).

4.4 Finnish Language Tweets

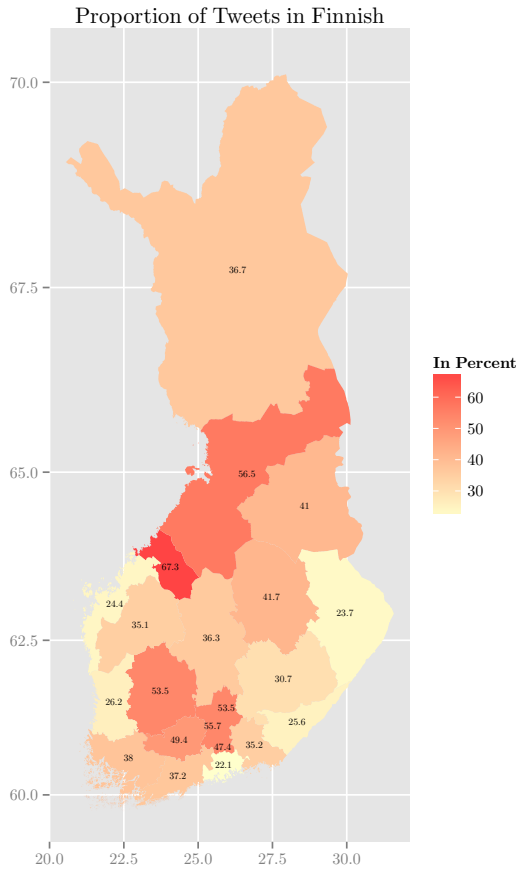

FIGURE 4.5: Proportion of Tweets in Finnish

After filtering to consider only tweets which originate from within Finland and whose language can be identified with greater than 60% accuracy, Finnish emerges as the most–tweeted language in Finland: 44.8% of tweets in the data are Finnish-language.[5] Figure 4.5 shows

[5]Not considering the tweets originating from neighboring Sweden, Norway or Russia naturally increases the relative proportion of Finnish tweets in the data.

the percentage of tweets in Finnish per region. Of the 5,664 unique Twitter users from this data, 2,665 (47.1%) authored at least one tweet in Finnish.

4.5 English Language Tweets

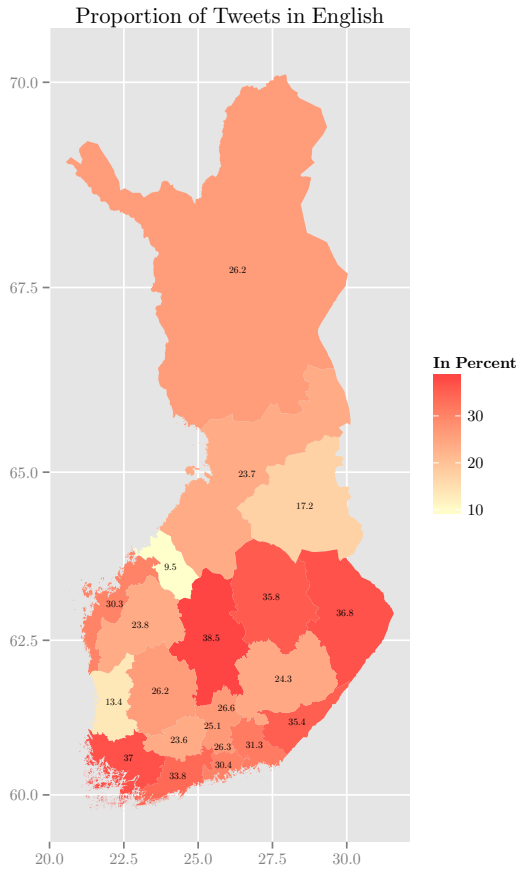

FIGURE 4.6: Proportion of Tweets in English

Figure 4.6 shows the percentage of tweets in English per Finnish province. 35.7% of the tweets in the data are in English. 54.2% of users in the data authored at least one tweet in English.

4.6 Swedish Language Tweets

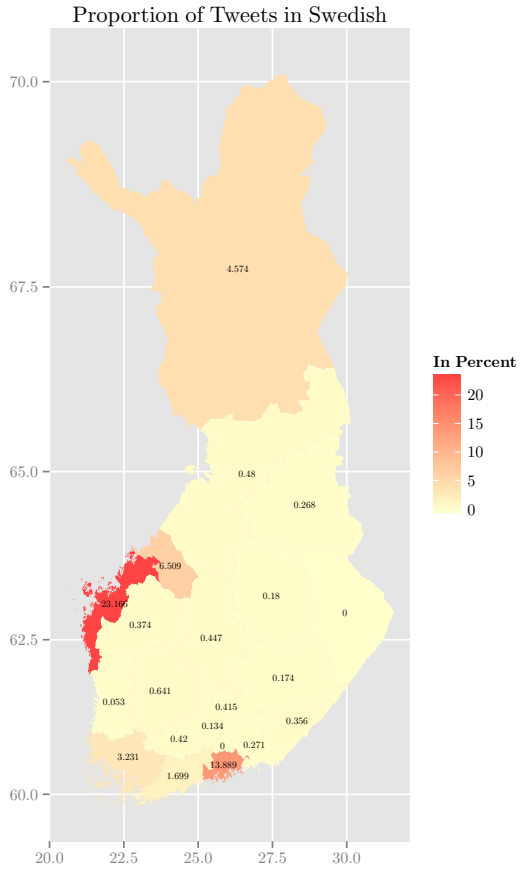

FIGURE 4.7: Proportion of Tweets in Swedish

Figure 4.7 shows the percentage of tweets in Swedish per Finnish province. After filtering the geo–tagged data to remove tweets originating from outside of Finland (e.g. Northern Sweden), the relative proportion of tweets in Swedish drops dramatically: Only 2.2% of tweets are in Swedish. The highest proportion of Swedish-language tweets occurs, unsurprisingly, in Ostrobothnia, historically a region with a high population of Finland Swedes. Schools and municipal services are in Swedish in many towns of Ostrobothnia. Eastern Uusimaa, (*Östra Nyland*), which also has a number of historically Swedish-speaking small municipalities, also exhibits a higher rate of Swedish tweets. Although Swedish is

not well represented in terms of overall volume, it is better represented in terms of the proportion of users who sent at least one message in Swedish. 6.89% of Finland's Twitter users posted at least one tweet in Swedish. This may be an indication of the continuing multilingualism of a certain proportion of the Finnish population, despite the relative decrease in the prominence of Swedish over the last century. According to census data, 5.4% of the Finnish population reports Swedish as their main language.

4.7 Russian Language Tweets

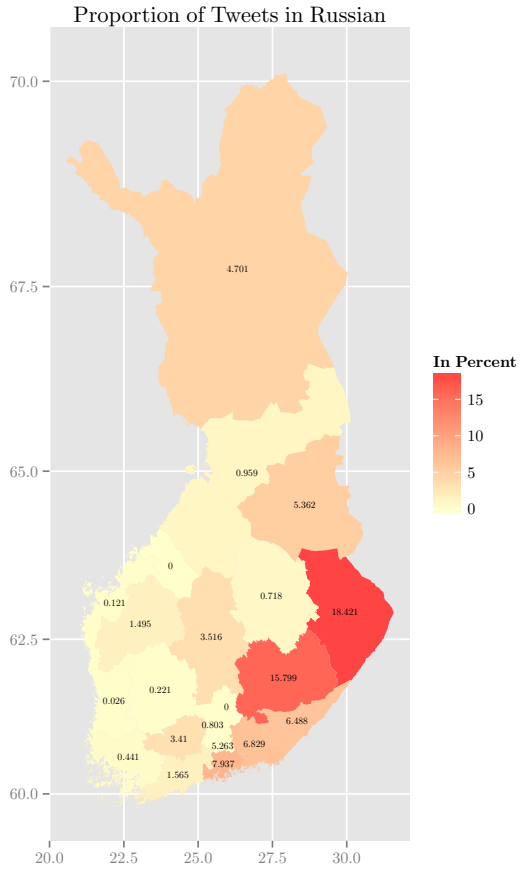

FIGURE 4.8: Proportion of Tweets in Russian

Figure 4.8 shows the percentage of tweets in Russian per Finnish province. 2.1% of tweets originating from within the Finnish borders are in Russian. The highest proportion of Russian tweets occurs in North Karelia with 18.4% (although the number of tweets in this region, 38, is extremely low), followed closely by South Savonia with 15.8%. Table 4.1 shows ten of the Russian tweet messages from South Savonia. Finnish regions that border Russia show higher proportions of Russian language tweets. Western Finnish regions exhibit very low proportions of Russian tweets. Interestingly, more Finland-based Twitter users sent at least one tweet in Russian than sent at least one tweet in Swedish: 7.89% of users sent a Russian tweet. This may represent activity by short–term visitors to Finland from the large St. Petersburg metropolitan area: the southeastern part of Finland sees a large number of Russian visitors who typically visit Finland in order to take advantage of lower prices on certain consumer goods such as electronics. A cursory scan of the content of some Russian language tweets originating from Eastern Finland suggests that the Twitter users who posted the message may reside in the neighboring Russian provinces of Vyborg or St. Petersburg.[6]

Table 4.1: Examples of Russian Tweets from South Savonia

	Text
1	Грязно, пыльно, в тени все еще ”минус”, но, однозначно - #весна @ В Городе http://t.co/plxeI5HxEi
2	может я буду белочкой,а ты будешь зайчиком
3	Кто нибудь, дайте пищу для сквернословия.. а то в голову ничего не лезет.
4	Просто потому что все давно не так, как должно быть.
5	О! К нам вернулась обновленная белочка)) Ну, #весна, как ни как;)) @ ООО ”Верные Решения” http://t.co/b1XmbdUL96
6	продолжают любить их дальше.” (С)
7	За окном. Оделась соответствующе. Теперь мне жарко. @ Östra Strandgatan (В) http://t.co/HH91tTzfPr
8	сука, не звонить···(С)
9	какой-то суке я его уже видела!:-)
10	@JaredLeto ты спишь хотя бы иногда?)

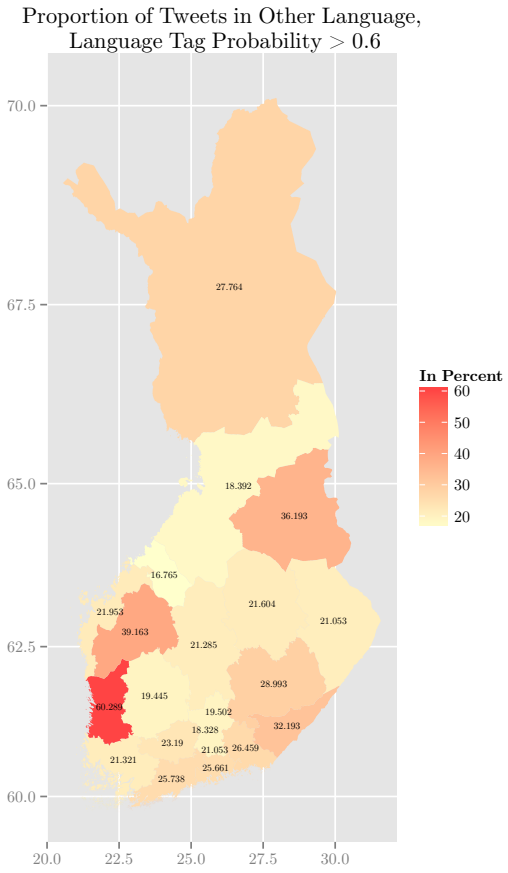

FIGURE 4.9: Proportion of Tweets in Other Language

4.8 Tweets in Other Languages

Many other languages were represented in the data to varying extent (Figure 4.9). French, Spanish and Estonian were well represented, with 2.60%, 2.04%, and 1.41%, respectively of the total Finland tweets. German tweets constituted 1.21% of the data.[7] Other commonly

[6]Russia–Finland border crossings were at their highest in 2013 and early 2014; they declined dramatically in late 2014 due to political and economic developments in the EU and in Russia.

[7]The relatively small percentage of tweets in German perhaps reflects the recent steep decline in the popularity of German as a school subject; the language was for much of the 20th century the most widely-taught foreign language in Finland.

tweeted languages in Finland include Italian, Dutch, Norwegian, Polish, Turkish, and Lithuanian. Non-Western languages such as Mandarin and Arabic were also significantly represented. In total, 67 languages were identified in the Finland data, and 15.3% of the total tweets originating from within Finland were not in Finnish, English, Swedish, or Russian.

The value represents tweets that are actually in other languages as well as some cases in which orthographic variation, symbols such as emoticons, proper names, urls, and language mixing phenomena cause misclassification; as described above, although a probabilistic assignation value of greater than 0.6 tends to indicate correct language identification, some misclassification is still possible. Table 4.2 shows an example of eleven tweets from the Finland data classified as not Finnish, English, Swedish, or Russian.

TABLE 4.2: Tweets Classified as Other Languages

	Text	Lang
1	Bentar lagi dirumah ;) (at @HelsinkiAirport (HEL) w/ 22 others) [pic]: http://t.co/MABKQ0917j	id
2	@mustardmon “no hotlinking”	it
3	Äiti sano et krisus looks laik shisus and i b like naaa suho’s da mini shisus	la
4	Shoot ’Em up no he visto película más surrealista en mi vida jajaja, eso sí, menuda tele se gastan aquí los amigos http://t.co/DLAqwYONZj	es
5	Jatkot! Brrrrrädi! (@ Pormestari Bar & Kitchen w/ @bradi03) http://t.co/JwmVU3iRkU	de
6	Iha ok?? http://t.co/bCSJXSKRlO	es
7	@jazzytheELF lol sul varmaa o hauska xD	lb
8	@iidulainen i am!	no
9	Creep desu	nl
10	@sanew__1 hehehe good ;D POKEMON OMAKSENI SAAN! USKO VAAN! (Joo ja e tyst nu xD)	sq
11	@jazzytheELF haha lol nii varmaa xD plondi mikä plåndi.	sw

Two of the ten are clearly not in English, Finnish, Swedish, or Russian (numbers 1 and 4). No. 2 is English but is misclassified as Italian due to the short length of the text and the presence of the username *@mustardmon*. No. 3 is a combination of truncated Finnish forms and non–standard English written in phonetic representation(*Äiti sanoo, että* [‘Mother says that’]*Krisus looks like Jesus and I be like no, Suho’s the mini Jesus*), misclassified as Latin. Nos. 5, 6, and 7 are very short Finnish messages with non-standard orthography and proper nouns or urls, misclassified as German, Spanish and Luxembourgish. No.

8 is a username and two English words, misclassified as Norwegian, and No. 9 is an English word and a transliterated Japanese word (meaning 'you are a creep'), misclassified as Dutch. Number 10 is an English–Finnish–Swedish language mixture with a username, non–standard Swedish orthography, and an ASCII emoticon, misclassified as Albanian, and No. 11 is Finnish with a username, an emoticon, and non–standard Finnish orthography, misclassified as Swahili.

For the most part, however, the probabilistic tagger accuracy estimation provides a reasonable indication of the correctness of the language tag. An inspection of tweets assigned a language tag probability of greater than 0.6 found that most were indeed correctly classified. From the perspective of a characterization of the English–language discourse of the data, language misclassification seems to result primarily in type II errors, where intelligible English or English–mixed tweets are classified as non–English. Much more problematic for the ensuing analysis would be type I errors, where non–English tweets were consistently classified as English.

The case of Satakunta, where a high proportion of tweets are not in Finnish, English, Swedish or Russian, represents a different situation: It is the result of extremely high Twitter activity of a single user tweeting in French; the content of the tweets suggest this may have been an exchange student residing in Pori during the data collection time period in early 2013.

In summary, a short analysis of the geographical distribution of tweets posted from within the borders of Finland suggests a substantial degree of linguistic diversity, with English–language tweets almost as prevalent as those in the most widely used language of the country, Finnish. This fact reflects the global orientation of many Twitter users, particularly those in Finland, as the analysis of the lexical content of Finland English Corpus tweets in Section 5.2.1 shows.

Chapter 5

Analysis of Tweet Length, Lexical and Grammatical Features

In this chapter, some of the principal properties of the data in the corpora used in the study (Finland English Corpus and Comparison English Corpus) are examined. First, some non–language–code surface features of the data, such as tweet and token length, are examined. Then, some basic theoretical and background considerations pertaining to the distribution of lexical items (token types) are presented, after which the relative type frequencies are examined in the Finland English and Comparison English data and in the gendered Finland English data. A derived lexical feature, non-standard orthography, is considered, as is the frequency of items in the lexical class profanity/taboo words. The frequencies of the non-standard lexical feature of emoticons are considered.

In a similar manner, the distributional profiles of grammatical items as determined by the CMU Twitter Tagger are compared after having been introduced with some examples from the data sets. Closer consideration of some grammatical word classes shows characteristic differences in frequencies between the two principal corpora. A non-canonical grammatical feature, expressive lengthening, is introduced and examined.

The ensuing discussion situates the findings of the analysis within the context of previous research on lexical and grammatical variation in the English of non-L1 users, language variation in Social Media, and gender differences in language.

5.1 Tweet Length

The length of Twitter user messages is constrained by the Twitter parameter limiting user message length to 140 characters, but within the range of three to 140 characters, there is wide variation in tweet length. As a variable, tweet length differs between the Finland English and the Comparison English corpus in a way that suggests Finland Twitter English is less information-oriented and more interactive. Tweet length also varies according to gender or sex within the gendered subsection of the Finland English Corpus.

5.1.1 Tweet Length by Geography

Tweet lengths for the Finland English and Comparison English corpora, as measured by number of characters and number of words per tweet, are shown in Figures 5.1 and 5.2.

The spike at n=140 is due to the automatic shortening of longer tweets by Twitter; tweet messages longer than 140 characters are truncated to 120 characters and a 20-character url linking to the longer text is added. A few tweets in the data are longer than 140 characters; these are user messages that contain Unicode symbols that can't be rendered in *R* and are therefore automatically converted to an eight-character sequence representing their position in the Unicode code scheme <U+XXXX> (i.e. from four to seven characters longer than the corresponding Unicode text, depending on the byte length of the non-displayable Unicode character).[1] Disregarding the spike due to addition of a url, the mode

[1]As discussed above in Section 3.4.2, filtering steps were applied to remove most Unicode characters that are not renderable in *R*. Most of these characters are found in a specific Unicode block range and are therefore relatively easy to filter using regular expressions, but others are in code blocks for which all other characters have corresponding CP-1252 equivalents.

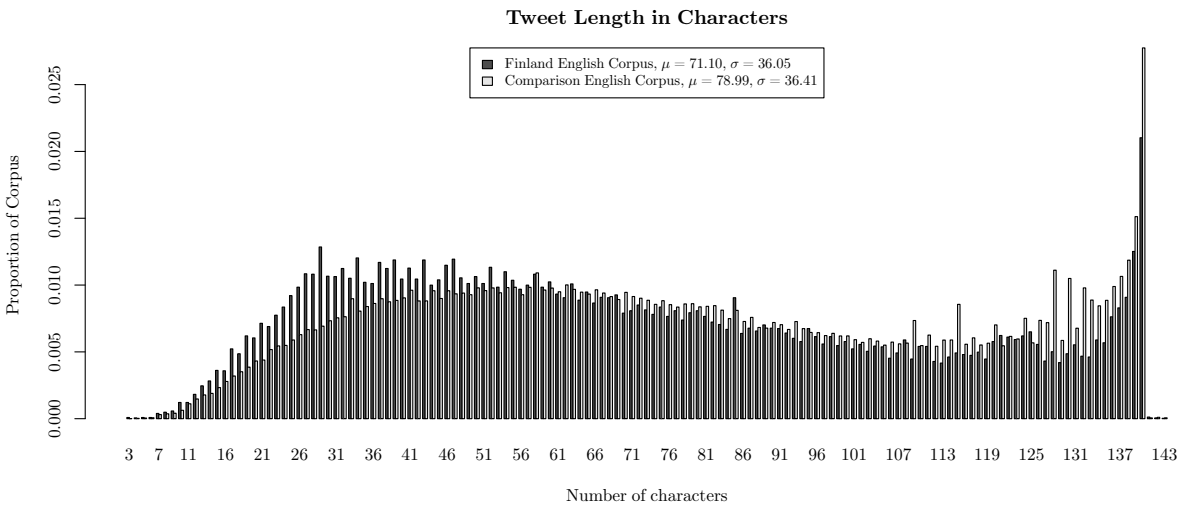

FIGURE 5.1: Tweet length in Characters, Finland English and Comparison English Corpora

for the Finland English data is 28 characters, whereas the mode for the Comparison English data is 58 characters. English tweets from Finland are shorter than English tweets from everywhere. In our data, the difference in mean tweet length is 7.89 characters between the two data sets.[2] The variance is similar for the data sets.

Tweet length in number of tokens shows a similar pattern, with tweet lengths distributed over a range of 1 token to 42 tokens. Again, Comparison English messages are longer than Finland English messages, by an average of 2.46 tokens.[3] The variance is similar for the data sets.

Some research has suggested that average utterance lengths in Twitter have been gradually decreasing since the service was initiated in 2007. Alis and Lim (2013) show that mean tweet length for a large Twitter English dataset compiled between 2009 and 2012 has decreased by approximately 8 characters, from ~85 to ~75 characters per tweet. This

[2]Highly significant according to a two–sample t–test (Welch's); t = -36.84, p-value $< 2.2e^{-16}$.
[3]Highly significant according to a two-sample t-test (Welch); t = -55.72, p-value $< 2.2e^{-16}$.

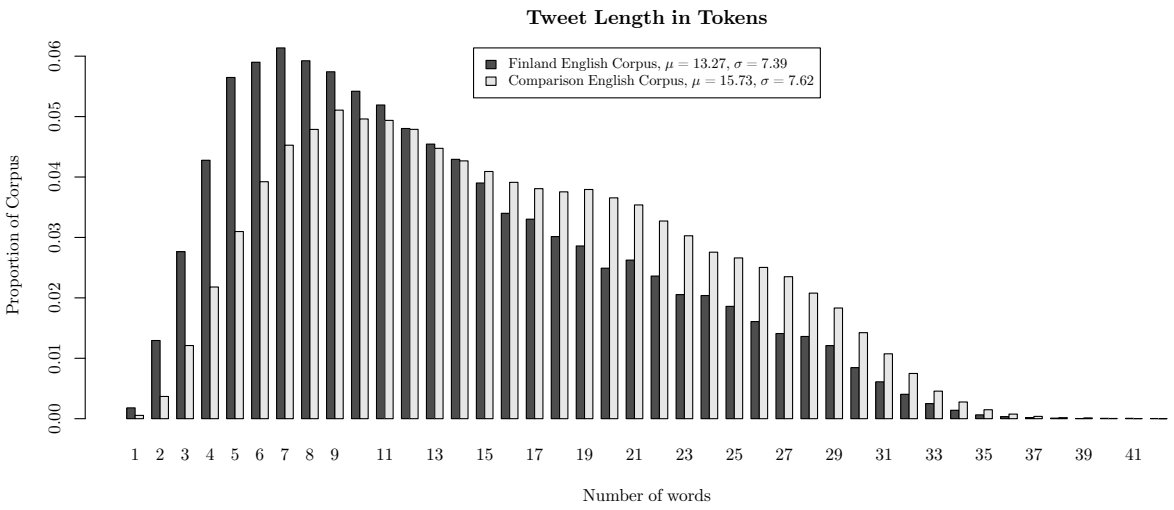

FIGURE 5.2: Tweet length in Tokens, Finland English and Comparison English Corpora

value of 75 characters per tweet is comparable to the mean tweet lengths in our corpora, compiled in 2009 and 2013. Alis and Lim find that tweets are shortest in US states and metropolitan statistical areas with a high population proportion of African–Americans; this is attributed to "increased use of jargon" in the African–American community (7).

Average token length also differs between the Finland English and Comparison English corpora, but in this case the Finland English tokens are somewhat longer: on average 4.54 characters, whereas Comparison English tokens are on average 4.21 characters long.[4] As discussed below in Section 5.3.4.2.1, this is possibly due to much lower rates of article use in the Finland English data.

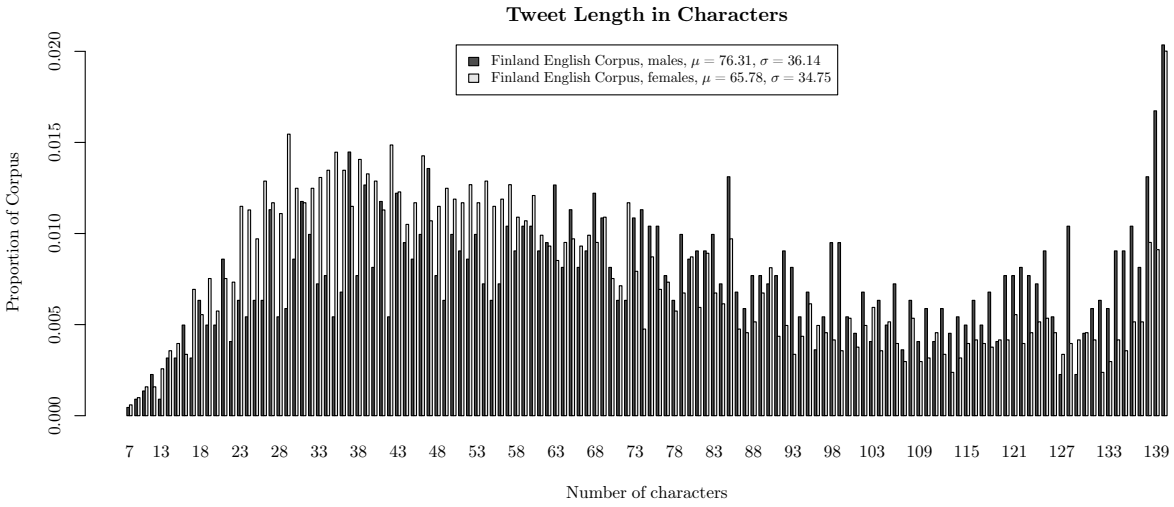

FIGURE 5.3: Tweet length in Characters by Gender, Finland English Corpus

5.1.2 Tweet Length by Gender (Finland English Corpus)

Male-authored tweets are longer than female-authored tweets in the Finland English data by an average of 10.53 characters (Figure 5.3).[5] Male tokens are longer than female tokens, on average (4.83 characters to 4.62 characters).[6] The variance between the two groups is similar. Newman, Groom, Handelman and Pennebaker (2008) found that males tend to use more longer words than do females. This is also true for the gendered portion of the Finland English Corpus: The proportion of tokens longer than 6 characters is 0.21 for males and 0.19 for females.[7] Gendered difference in the tweet length feature is discussed below in Section 5.3.3.2.

[4]Highly significant according to a two–sample t–test; t = 54.43, p-value $< 2.2e^{-16}$.
[5]Highly significant according to a two–sample t–test; t = 11.55, p-value $< 2.2e^{-16}$.
[6]Highly significant according to a two–sample t–test; t = 7.03, p-value $= 2.12e^{-12}$.
[7]Highly significant according to a χ^2 goodness–of–fit test; $\chi^2 = 100.28$, p-value $< 2.2e^{-16}$.

5.2 Lexical Features

In this section, the most frequent lexical items in the Finland English and Comparison English corpora will be considered; the most frequent lexical items in large composite corpora will serve as a point of reference. The distinctiveness of Finland Twitter English begins to emerge in the relative rankings of the most frequent lexical types compared to those of other Twitter English discourse (in the Comparison English Corpus) and those of reference corpora. By using a contingency table of expected and observed type frequencies and calculating the odds ratio θ for types that occur in both the Finland English and the Comparison English corpora, it is possible to characterize Finland Twitter English in more detail, including any differentiation as it relates to the sociolinguistic variable of gender/sex. The section concludes with an investigation of two categories of lexical variables that have been shown to exhibit correlation with sex/gender: the use of taboo words or profanity and the use of emoticons. It can be shown that these variables are distinctive for Finland Twitter English, compared to non–specific Twitter English, and that while some earlier findings pertaining to sex/gender and language variation can be confirmed, others are not.

5.2.1 Lexical Features by Geography

Frequency–based studies typically distinguish between a word type, or the unique words in a data set, and a word token, a single instantiation of a type. The distinction can be illustrated by considering an example tweet from the Finland English Corpus, shown in Table 5.1.

Table 5.1: Finland Corpus Message with 26 Tokens and 24 Types

Had a nice day out #shopping and chatting with neighbor/friend. :) Got new shoes, couple dresses and bunch new hair material.

In the example, there are 21 total word tokens in the sentence. In addition, there are 5 non-word tokens [*/ . :) , .*], for 26 total tokens.[8] There are only 24 types, however, as the words *and* and *new* are each used twice in the sentence.

Frequencies of all token types were investigated in the Finland English and Comparison English corpora. As a basis for comparison, the twenty most frequent non–punctuation types of the Finland English and Comparison English corpora are shown in Table 5.2 with those of the 440 million–word Corpus of Contemporary American English (*COCA*, Davies and Gardner 2010) and 100 million–word British National Corpus (Leech, Rayson and Wilson 2001);[9] frequencies are reported per thousand word tokens.[10]

While the most frequent types in the corpora are similar (Table 5.2), both the relative frequencies of the types per thousand words and the rankings differ somewhat. Some of the most frequent types reported by Davies and Gardner (2010) represent verbal lemmas (*be, have, do, say*); the reported frequencies correspond to the sum of the constituent verbal forms such as *am, is, are, was, were, being, be*, etc. As the Finland English and Comparison English corpora and the BNC spoken and written data frequencies do not represent lemmatized forms, these entities are not directly comparable. Davies and Gardner (2010) and Leech, Rayson and Wilson (2001) consider *to* as a preposition to be a different word than *to* as an infinitive marker; they also treat the possessive genitive morpheme *'s* as a distinct type, whereas the tokenization procedure used to create the Finland English and Comparison English Corpora does not make these distinctions. *Er*,

[8]The tokenization procedure used to compile the corpora does not treat the # symbol as a distinct token when it occurs in user messages immediately before a word form.

[9]The frequency of punctuation types is not commonly reported in corpus–based lexicology. For comparison purposes, the frequencies reported for the Finland English Corpus and Comparison English Corpus in Table 5.2 represent values after the removal of punctuation.

[10]Frequencies in lexicology are often reported per million words, but the decision to report per million words or per thousand words with two decimal places is somewhat arbitrary: For example, Biber et al. (1999) report frequencies in terms of thousand occurrences per million words. For the Finland English and Comparison English corpora, the total number of tokens is in the range 10^5 to 10^6; for this reason (and for simplicity's sake), the figures here are reported per thousand words with two decimal places.

TABLE 5.2: 20 Most Frequent Types in the Finland English Corpus, Comparison English Corpus, COCA, and the Written and Spoken Sections of the BNC with Frequency per 1000 Word Tokens (COCA Data Derived from Davies and Gardner 2010, pp. 8–9; BNC Data from Leech, Rayson and Wilson 2001, p. 144 and p. 181)

	Finland		Comparison		COCA		BNC Written		BNC Spoken	
1	i	31.61	the	32.64	the	46.44	the	64.42	the	39.60
2	to	20.98	to	26.45	be	32.59	of	31.11	i	29.45
3	the	20.11	i	23.29	and	22.49	and	27.00	you	25.96
4	and	17.47	a	21.29	of	21.79	a	21.97	and	25.21
5	you	14.67	and	15.23	a	18.54	in	18.98	it	24.51
6	a	14.57	is	13.78	in	14.72	to	16.44	a	18.64
7	NUM	14.11	of	12.57	to	13.28	is	9.96	's	17.68
8	my	12.02	in	12.28	have	10.36	to	9.62	to	14.91
9	is	11.07	for	12.14	to	8.09	was	9.37	of	14.55
10	in	10.55	NUM	11.68	it	8.15	it	9.30	that	14.25
11	me	9.13	my	10.87	i	8.31	for	8.66	n't	12.21
12	of	8.75	you	9.95	that	7.21	that	7.32	in	11.61
13	it	8.54	on	9.87	for	6.87	with	6.82	we	10.45
14	so	8.18	it	9.62	you	6.45	he	6.76	is	9.59
15	for	7.71	that	7.56	he	6.08	be	6.74	do	9.33
16	that	7.71	with	6.46	with	5.61	on	6.57	they	8.54
17	im	7.07	at	6.31	on	5.20	i	6.49	er	8.10
18	this	6.43	just	5.55	do	5.41	by	5.49	was	7.89
19	just	6.02	have	5.49	's	4.52	's	4.95	yeah	7.49
20	but	6.00	this	5.03	say	4.02	at	4.87	have	7.31

in the BNC spoken data, is a hesitation/filler word or interjection. NUM represents all numbers in the Finland English and Comparison English corpora, whether as numerals or word forms. Davies and Gardner consider word forms, but not numerals; Leech, Rayson and Wilson consider numerals and transcribed word forms for the written data and word forms for the spoken data.

In the Finland English Corpus, the personal pronoun *i* is the most frequent word, followed by the preposition *to*, the article *the*, the conjunction *and*, and the personal pronoun *you*. In the Comparison English data, the article *the* is the most frequent type, followed by *to*, *i*, the article *a*, and *and*. *The* is also the most frequent word type in COCA and the BNC written and spoken sections. If we do not consider the lemma *be* (due to the inconsistency in the frequency counting procedures of the corpora), *and*, the preposition

of, *a* and the preposition *in* occupy the following four ranks in both COCA and the written section of the BNC .

Personal and possessive pronouns, word classes that are generally much more common in spoken than written language, are more frequent in the Finland English and spoken BNC data, followed by the Comparison English data and COCA: *i*, *you*, *my*, and *me* have the ranks 2, 8, 10, and 14 in the Finland English Corpus and frequencies in the range of approximately 32 to 9 occurrences per thousand words. The types occupy the ranks 3, 12, and 11 in the Comparison English data (*me* is not among the 20 most frequent word types), with frequencies in the range 23 to 10 per thousand words. In COCA, *i* and *you* occupy ranks 11 and 14; *me* and *my* are not represented among the 20 most frequent types, but the personal pronoun *he* occupies the 15th rank. The three pronouns have frequencies in the range of approximately 8 to 6 per thousand words. In the BNC written data, only *i* is among the twenty most frequent types. In the BNC spoken data, *i* and *you* are the second and third most frequent types; the personal pronouns *we* and *they* are also represented among the top twenty types in the ranks 13 and 16.

Articles are more common in COCA and the written BNC: *the* and *a* comprise 65 occurrences per thousands words in COCA and 85 per thousand words in the written BNC, compared to 54, 48, and 35 per thousand words in the Comparison English Corpus, the spoken BNC, and the Finland English Corpus, respectively. This finding, again, corresponds to known distributional profiles for word classes in written versus spoken language: Determiners and nouns are more frequent in written than in spoken language.

Prepositions are used more often in the Comparison English Corpus. *To*, used both as a preposition and a marker of infinitives, is slightly more frequent in the BNC written and in the Comparison Corpus, at 26 occurrences per thousand words in both corpora, compared to 22 occurrences per thousand words in COCA and 21 in both the Finland English Corpus and the BNC spoken data.

Other prepositions (*of, in, for, with, on, by, at*) are more frequent in the written BNC, accounting for 82 occurrences per thousands words. In the Comparison English Corpus, (*of, in, for, on, with, at*) account for 60 occurrences per thousands words; in COCA the types *of, in, for, with* and *on* are represented among the 20 most frequent types and collectively account for 54 occurrences per thousand words. Prepositions are much less frequent in the Finland English Corpus and the spoken portion of the BNC: the types *in, of* and *for* are represented among the 20 most frequent types in the Finland English Corpus, but only account for 27 occurrences per thousand words. *Of* and *in* comprise 26 occurrences per thousand words in the BNC spoken data.

The frequencies of the pronoun *it* are quite similar in four of the corpora, at approximately 8–10 occurrences per thousand words; *it* is much more common in spoken British English (24.5 per thousand words). *That* shows a similar pattern, occurring at a rate of approximately 7–8 per thousand words in the Finland English and Comparison English corpora, COCA, and the written BNC, but much more often (14.3 per thousands words) in the spoken BNC. The other word types among the 20 most frequent types vary, in part due to the slightly different counting and tokenization procedures. Four verbal lemmas (*be, have, do say*) are amongst the 20 most frequent COCA types, as is the possessive morpheme *'s*. The rates of occurence of numerals (NUM), of *is*, and of *just* are similar in the Finland English and Comparison English corpora.[11] *So* is represented among the most frequent types in the Finland English data, as are the proximal demonstrative *this*, the coordinating conjunction *but*, and the contracted form *im* (from *I'm*), but not in the Comparison English Corpus. An interpretation of this patterning of word frequencies is offered below in section 5.4.

[11]NUM was substituted for tokens consisting only of numerical digits, such as *7*, *100*, or *54*, not for tokens consisting of a combination of numerical digits and letter characters, such as *1st* or *7-inch*. See Section 5.3.4.2.4.

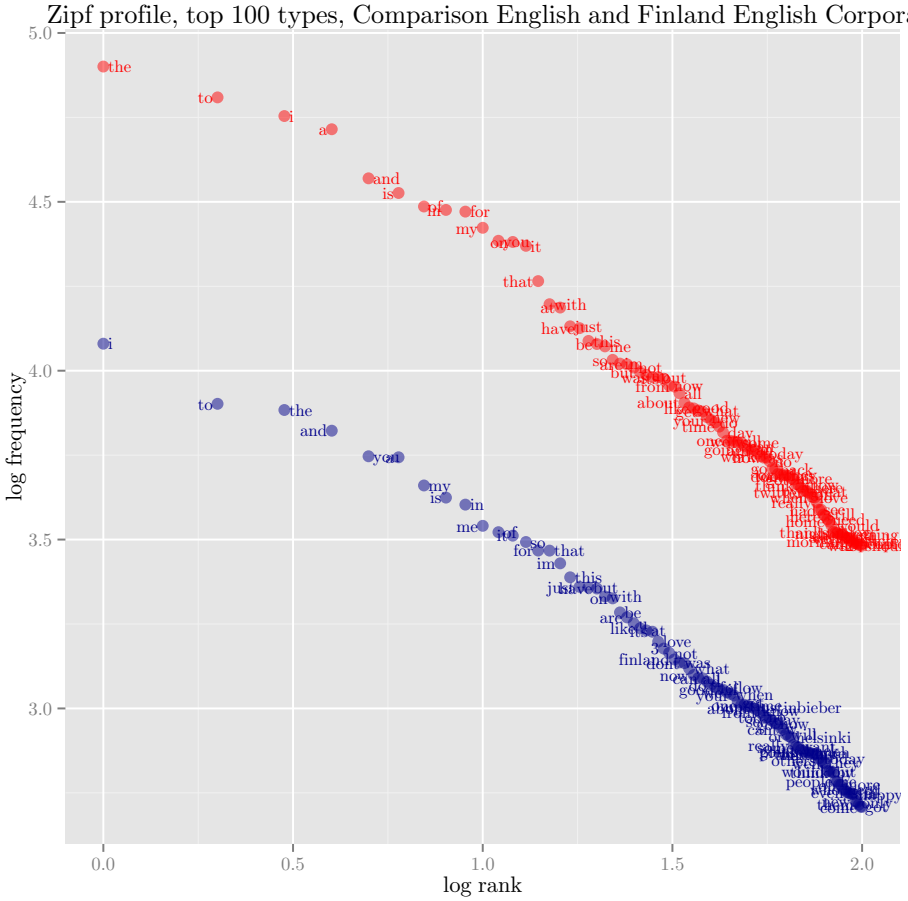

FIGURE 5.4: Rank–frequency Profile for Top 100 Word Types, Comparison English (red) and Finland English (blue) Corpora

Ranking the types that occur most frequently in the Finland English and Comparison English corpora and plotting the results in double–logarithmic space results in two approximately straight lines (Figure 5.4 shows the 100 most frequent types) corresponding to the well–known Zipf distribution (Zipf 1935; 1949). Overall, it can be remarked that the most frequent types in the two Twitter corpora as well as COCA and the BNC exhibit the expected profile: They are mainly short function words containing relatively little semantic content and are used in the establishment and maintenance of discourse,

for example as discourse coordinators, definiteness markers, or deixis indicators. The relative frequencies of some of these types in the data reveals something about differences in underlying communicative and discourse strategies: Twitter English is pronoun–heavy and article–light, compared to COCA; in this way it is more similar to the spoken BNC. The texts in COCA, which comprises mainly written text types, and in the BNC written section have very different communicative functions than the functions most typical of Twitter messages. Twitter English is not primarily informational: It is interactive and situational, consisting of linguistic expressions that allow users to engage in dialogue with other users and situate their propositions in relation to those of other users. Finland Twitter English is especially pronoun heavy and especially article–light, and makes much less use of prepositions than does global Twitter English.[12] Some of the implications of these findings are discussed below in Section 5.4.

As the Finland English and Comparison English corpora are both random samples of English–language Twitter user messages, for which the most relevant difference is the place of origin of the messages (Finland or the entire world), a discrepancy in the rank order of the most frequent types may shed light on functional differences in the use of English for discourse organization reflecting slightly differing communicative intent between the groups. Another approach is to examine those word types for which the differences in frequency are most pronounced. In the following section, the most "Finnish" and least "Finnish" types are considered in the Finland English and Comparison English corpora. Here, punctuation is once again taken into account, as punctuation symbols play an important role in much Twitter language.

[12]The role of personal pronouns and other word classes is considered in more detail in Section 5.3.3.

5.2.2 Quantifying Lexical Similarity

The similarity or dissimilarity of a type distribution in two corpora can be quantified on the basis of a contingency table that takes into account observed and expected frequencies. There are many association measures for quantifying similarity based on mathematical expressions that consider the principal parameters of observed frequency, expected frequency (under the assumption of independence), and corpus size (Evert 2004: 28ff., Evert 2008, Bouma 2009). In this section three association measures used to compare relative type frequencies are reviewed: the Chi–square test statistic, the log–likelihood test statistic G_2, and the odds ratio θ.

TABLE 5.3: Two Example Corpora Containing the Lexical Type *remembered*

Corpus 1	You guise... i just **remembered** all the posters i got for xmas... nearly cried bc happinesas-dfhcb
Corpus 2	i was so shocked when our teacher said we come to school on Friday. until i **remembered** the week when we had Friday off was last week.

TABLE 5.4: Contingency Table for Comparison of Lexical Type Frequencies

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

	$corpus_1$	$corpus_2$	
$remembered$	1	1	$= 2$
$\sim remembered$	17	28	$= 45$
	$= 18$	$= 29$	$= 47$

A contingency table for word types serves as the basis for statistical modeling of frequency distributions (Evert 2004: 75ff.). It is created by counting observed frequencies for a type in two samples and calculating expected frequencies under an assumption of independence. An example for a sample of two tweets from the Finland English Corpus can be seen in Table 5.3: There are 47 total tokens in the two small "corpora". The lexical type of interest, *remembered*, occurs once in each corpus.

The expected frequencies under the assumption of independence can be derived from the observed frequencies; they are the marginal frequencies of the occurrence and the

TABLE 5.5: Expected Frequencies from Three Example Tweets

	$corpus_1$	$corpus_2$
$remembered$	$E_{11} = \frac{R_1 C_1}{N} = \frac{36}{47}$	$E_{12} = \frac{R_1 C_2}{N} = \frac{58}{47}$
$\sim remembered$	$E_{21} = \frac{R_2 C_1}{N} = \frac{810}{47}$	$E_{22} = \frac{R_2 C_2}{N} = \frac{1305}{47}$

non–occurrence of each constituent. The cells in Table 5.4 add up to the number of tokens in the two corpora: 47.

Using these observed and expected frequencies, one can then proceed to quantification of association strength using various measures. Fisher's exact test is considered the best test statistic for contingency tables from skewed distributions, but as it involves calculating hypergeometric distributions, it is too computationally intensive to be used as an association statistic for large numbers of pairs (Evert 2004b). Instead, less computationally intensive association measures are used in lexical statistics.

Pearson's Chi–square test statistic χ^2 (Equation 5.1), well known as a test of independence for two samples of count data arranged in a contingency table, can be used as an association measure to gauge the relative frequency for a type in two texts.

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} \quad (5.1)$$

The equation in 5.1 can be considered similar to the mean squared error for a contingency table. However, the statistic achieves best accuracy for distributions that approximate the normal distribution; it thus can give an accurate measure of collocational strength for common types, but provides too high values for type pairs for which there are only few tokens in the texts being compared. Dunning's log–likelihood test statistic G_2 (Equation

5.2) is more sensitive to skewed samples.

$$G^2 = 2\sum_{ij} O \log \frac{O}{E} = 2(O_{11} \log \frac{O_{11}}{E_{11}} + O_{12} \log \frac{O_{12}}{E_{12}} + O_{21} \log \frac{O_{21}}{E_{21}} + O_{22} \log \frac{O_{22}}{E_{22}}) \qquad (5.2)$$

The odds ratio, θ, represents another measure that is relatively insensitive to low–frequency counts (Equation 5.3).

$$odds\ ratio\ \theta = \log \frac{O_{11}O_{22}}{O_{12}O_{21}} \qquad (5.3)$$

A short consideration of the lexical types that are most characteristic of the Finland English Corpus and the Comparison English Corpus, as calculated on the basis of the odds ratio θ, will allow interpretation of content–related aspects of the English discourse of tweets originating from Finland and how it compares with the discourse of tweets with no geographical specification.[13]

Table 5.6 shows the 20 types for which the ratio of use Finland English Corpus: Comparison English Corpus is the highest and the 20 types for which the ratio Comparison English Corpus: Finland English Corpus is the highest.

The types overrepresented in the Finland English Corpus represent a mixture of lexical entities. Many of the types are Finland– or Northern–Europe specific, but several of the most Finnish types are characteristic to the domain of CMC or Social Media. One of the most "Finnish" types is the Unicode heart symbol ❤, which is 1606 times more frequent in the Finland English Corpus than in the Comparison English Corpus; an ASCII representation of a double heart (*<3<3*) is also highly overrepresented in the Finland data. The non–pronounceable form *xx*, a non–facial emoticon typically held to represent kisses; the common CMC initialisms *ikr*, ('I know, right?') and *fml* 'fuck my life'); and the lengthened emoticon *:DD*, referred to as "bigsmile" in Schnoebelen's (2012) proposed

[13]As the odds ratio values correspond to large discrepancies between observed and expected frequencies and extremely low p–values (< 0.0001) in a χ^2 test, the difference in use between the two corpora is significant for all of the items in Table 5.6.

TABLE 5.6: Finland English–Comparison English Frequency Ratio for 20 Most and Least "Finnish" Types

	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

taxonomy (see Section 5.2.5) comprise the other CMC–domain types that are represented among the most Finnish types compared to the Comparison English Corpus. The type *niall* is the name of an English singer popular with a younger, mainly female demographic. *#Party* and *#food* are hashtags, used as topic indicators in Twitter (but see Section 5.4), and *youu* may be a result of widespread smartphone typing. Finally, *apparatus* seems to have high prominence in the Finnish data as a result of some technical journal article titles being sent as tweets, most likely by an app or other automated source (and despite efforts to filter tweets of automatically sent content).

The types overrepresented in the Comparison English Corpus correspond to mainly American and British named entities. American culture is represented by the types *steelers, palin, obama* and *nfl*; British culture is represented by the types *m4* and *a5* (traffic

conditions on British roads and motorways are frequently tweeted by British transportation authorities). Other "less–Finnish" types include American entertainment, technology and internet–related entities such as *flickr, xbox, firefox, ebay* and *myspace*. The type >> represents a topic–content transition marker in tweets that was apparently widespread in certain tweets prior to the adoption of the hashtag for this purpose.

Some of the differences are clearly due to the time difference in compilation of the two corpora. Types such as *niall* or *gaga* were uncommon in 2008–2009 simply because the named entities referred to had not yet achieved widespread recognition.[14] Similarly, words such as *palin* and *myspace* refer to entities that enjoyed far greater media prominence in 2008 compared to 2013; the use of the sequence >> is also no longer common on Twitter.

Other differences, however, such as the prominence of the emoticon types ❤, *xx, <3<3*, and *:DD* seem to represent a salient difference in the communicative functions typical of Finland Twitter users. This is discussed in more detail in Section 5.4.

5.2.3 Lexical Features by Gender (Finland English Corpus)

Many lexical features have similar distributional profiles for male and female Twitter users in Finland tweeting in English. There are, however, some differences in the relative frequencies with which items are used according to gender. In order to shed light on the dynamics of gendered English language usage on Finland Twitter, lexical items were analyzed by gender. In an initial step, all tokens from tweets that had been disambiguated for gender were sorted according to their type frequency rank.

Table 5.7 shows the 20 most frequent types for Finland English Corpus tweets disambiguated for gender; these are (unsurprisingly) quite similar and reflect the well–known phenomenon that short, closed–class function words are the most frequently used lexical items. The frequencies per thousand words are somewhat different for the individual types.

[14]These refer to the entertainers Niall Ferguson and Lady Gaga.

TABLE 5.7: Most Frequent Types in the Finland English Corpus by Gender per 1000 Words

	Males	Frequency	Females	Frequency
1	i	22.01	i	33.27
2	and	18.86	to	20.36
3	to	18.74	you	19.54
4	the	17.38	and	17.79
5	NUM	14.54	the	17.20
6	a	12.29	NUM	12.28
7	you	10.42	a	12.25
8	is	10.23	my	12.14
9	in	10.15	me	11.73
10	my	9.68	is	11.45
11	of	9.25	in	9.67
12	for	9.06	so	9.30
13	it	8.09	it	7.77
14	that	7.78	im	7.70
15	me	7.35	of	7.42
16	with	6.88	this	7.29
17	so	6.18	that	7.09
18	w	6.07	just	6.85
19	at	5.44	for	6.68
20	just	5.29	but	6.38

Females use personal pronouns at a higher rate than do males, whereas males use articles at a slightly higher rate. Males use the prepositions *to, in, of, for, with, w*, and *at* approximately 66 times per thousand words; females use the prepositions *to, in, of*, and *for* at a lower rate of 44 times per thousand words.

The male–female ratio for the 20 types that are most frequent overall in the gendered subsection of the Finland English Corpus are shown in Table 5.14. Types that differ most according to gender among the most frequent types are the personal pronouns *you* and *me* (used 1.9 and 1.6 times more often by females); the contraction *im* (from *I'm*), used 1.72 times more often by females; *so*, *this*, and *just*, used 1.52 times, 1.43 times, and 1.3 times more often by females; *i*, used 1.52 times more often by females; and the possessive pronoun *my*, used 1.25 times more often by females. Used at equal rates (±20%) are the

Table 5.8: Gender Ratio, Most Frequent Types

	Word	Male–Female ratio		Word	Male–Female ratio
1	i	0.66	11	in	1.05
2	to	0.92	12	so	0.66
3	and	1.06	13	of	1.25
4	the	1.01	14	it	1.04
5	you	0.53	15	for	1.36
6	NUM	1.18	16	that	1.10
7	a	1.00	17	im	0.58
8	my	0.80	18	this	0.70
9	is	0.89	19	just	0.77
10	me	0.63	20	with	1.29

types *and*, *the*, *a*, *is*, *in*, *it*, *NUM*, and *that*. Used more often by males are the type *of*, *for*, and *with*, which are 1.25, 1.36, and 1.29 times more frequent.

Notable in the data is that relatively few of the words have similar frequencies for males and females. Of the 2,348 word types used at least once by both males and females, only 380 have corrected male-female rates of use within the range $0.833 \leq x \leq 1.2$.

Differences in frequency for content words by gender are perhaps more informative in terms of characterizing the most salient semantic or discourse functional differences between gendered groups using English on Twitter in Finland. Using the procedure described above, the lexical items for which the gender difference was most prominent were identified. Here, again, punctuation is considered.

Table 5.9 shows the 20 types in the gender–identified portion of the Finland English Corpus for which the male–female odds ratio is most pronounced. The most "male" type is a "Japanese"–style emoticon representing arched eyebrows, frequently taken to indicate satisfaction, happiness, or contentment.[15]

[15]See Section 5.2.5.

TABLE 5.9: M–F and F–M Ratios for the 20 Most "Male" and Most "Female" Types in the Gendered Subsection of the Finland English Corpus

	Word	θ	Word	θ
1	^^	4.583	@austinmahone	-4.362
2	american	3.277	@niallofficial	-3.882
3	public	3.110	x	-3.094
4	w	2.946	#happybirthdayaustin	-2.654
5	#wNUMfijb	2.909	everybody	-2.619
6	buss	2.909	@harry_styles	-2.583
7	design	2.909	him	-2.545
8	oulu	2.909	harry	-2.506
9	@jklanacNUM	2.791	niall	-2.465
10	event	2.791	song	-2.465
11	#sky	2.658	ily	-2.423
12	heh	2.658	aw	-2.283
13	helsingin	2.658	babe	-2.232
14	style	2.658	account	-2.178
15	@alexstubb	2.503	he	-2.086
16	currently	2.503	bc	-2.080
17	#my	2.503	april	-2.060
18	cleaning	2.503	bieber	-2.060
19	college	2.503	idk	-2.060
20	exhibition	2.503	tour	-1.996

For the other items in the list, two are usernames (one of which being that of Alexander Stubb, a minister in the Finnish Government in 2013 and an well–known Twitter enthusiast). Three are hashtags, serving as overt topic indicators or discourse markers. Three of the types are adjectival descriptors or nouns referencing places (*american, oulu, helsingin*). One is an interjection (*heh*), one is a truncated preposition (*w*, for *with*), one is an adverb (*currently*), one is a participle or gerund (*cleaning*), and several are adjectival or nominal noun phrase elements (*public*, *buss*, *design*, *event*, *style*, *college*, and *exhibition*).

Of the types with the highest female-male ratio (Table 5.9), three are usernames, including the two most "female" types. The usernames are associated with anglophone male popular entertainers who appeal to a mainly younger female demographic segment. One type is a hashtag (the topical content refers to the birthday of Austin Mahone, whose

username is the most "female" type in the gendered subsection of the Finland English Corpus). The type *x* is a common emotional indicator that represents positive affective orientation. The other types most strongly associated with female users in the data are proper nouns (the personal names *harry*, *niall*, *bieber* and the proper noun *april*, personal pronouns (*him*, *he*), the indefinite pronoun *everybody*, two initialisms (*ily* 'I love you' and *idk* 'I don't know), an abbreviation (*bc*, 'because'), the pronounceable non–dictionary word *aw*, and the nouns *song*, *babe*, *account*, and *family*.

The content of the the items with the highest female-male ratio provides some insight into the communicative functions favored by female Twitter users in Finland writing in English. The prominence of items associated with young male Anglophone popular entertainers, despite filtering steps having removed large numbers of automated or multiple tweets containing these types, suggests that affective involvement with foreign celebrities is a prominent topical interest of this group.

5.2.4 Profanity and Taboo Words

Several analyses have found that males exhibit higher rates of use of profanity, taboo words and vulgar expressions in spoken language and some written genres such as CMC (McEnery 2005; Herring 2006; Bamann, Eisenstein and Schnoebelen 2014). Others have found that females exhibit a slightly higher rate of use of swear words in online blogs (Argamon et al. 2007). In order to investigate the use of taboo language in the data, a list containing 694 English word forms considered to be taboo, profane, or vulgar was compiled, based on material available in an online compendium of offensive English language (Noswearing.com 2015). The aggregate frequencies of the words were then determined for the Finland English Corpus, the Comparison English Corpus, and the gendered portion of the Finland English Corpus. The Finland English Corpus shows a much higher rate of use of taboo or swear words than does the Comparison English Corpus: 4.49 vs. 1.65 per thousand words.

In the gendered portion of the Finland English data, we find that females use profanity significantly more than do males: 4.06 vs. 2.80 per thousand words.[16] An interpretation of the role of profanity and taboo words in the discourse of Finland Twitter English can be found in Section 5.4.

5.2.5 Emoticons

Frequent use of glyphs indicating emotion is characteristic for CMC and particularly for Twitter (Hentschel 1998; Crystal 2006; Schnoebelen 2012; Bamann, Eisenstein and Schnoebelen 2014). The data collected from Finland exhibits frequent use of emotion indicators, both as symbolic elements and lexicalized interjections or particles. "Emoticons", in our usage, consist of (typically ASCII) character combinations meant to represent emotional content, often through the representation of facial expressions, as well some Unicode symbols with byte values less than U+1F300 (such as U+2764 ❤ or U+2665 ♥) that can be shown in almost all text processing software.

In the Finland English and Comparison English data, emoticons were detected in the corpora by filtering the output of the CMU Twitter Tagger using the Penn Treebank model, then selecting only those tokens that had been assigned the interjection tag *UH* tag.[17] In order to filter out only the emoticons, regular expressions were used to select the subset of those tokens that contained primarily non-letter symbols. In a further step, emoticon types were examined, and types whose emoticon–ness seemed questionable were removed, leaving a total of the 240 most common emoticon symbols for the ensuing analysis.[18]

[16]The results are significant according to Chi–square goodness–of–fit tests, p–value $< 2.2e^{-16}$.

[17]The Penn Treebank model uses this tag for interjections such as politeness forms, affective particles, and similar word types. The CMU Twitter Tagger applies the tag to emoticons as well.

[18]For example, the glyph combination *(@* was sometimes incorrectly tagged as an interjection instead of a preposition. The sequence occurs most frequently as an automatically generated addendum to user input that indicates location based on geo-coordinates, for example in tweets such as *I am at the cinema (@ Kinopalasti Pasila)*. See Appendix A for the code used to filter emoticons from all interjection–tagged tokens.

TABLE 5.10: 20 Most Frequent Emoticon Types in the Finland English Corpus, as Percent of All Emoticon Types

	Type	Percent		Type	Percent
1	:)	27.2	11	xD	1.1
2	<3	11.3	12	:P	1.0
3	:D	10.9	13	;))	1.0
4	;)	8.0	14	(:	0.9
5	:(	6.1	15	:))	0.9
6	:-)	4.5	16	:'(	0.7
7	❤	3.0	17	;-)	0.6
8	♥	2.7	18	:')	0.6
9	XD	1.6	19	(;	0.6
10	^^	1.3	20	:-D	0.5

The twenty most frequent emoticon types in the Finland English data (shown Table 5.10) comprise 84.5% of all emoticon tokens, and include 16 "smiley" emoticons (representations of facial expressions oriented along the x–axis), two Unicode symbols for hearts, and an ASCII character representation of a heart (<3).[19] One sequence (^^) is a so–called "Japanese" or "Asian"–style emoticon, in which the representation of the face (in this case only arched eyebrows) occurs along the y–axis rather than the x–axis, as is the case for Western-style "smiley" emoticons.[20]

The data from the Finland English and Comparison English corpora show a large range of variation in the use and distribution of emoticons. Of the 8,789 unique authors of the tweets originating from within Finland, 3,417 (38.8%) made use of an emoticon at least once. For all the tweets originating from Finland, (i.e. including non-English tweets), 18.9% of the messages contained at least one emoticon. The maximum number

[19] Schnoebelen (2012) suggests that emoticons were first invented on Usenet bulletin boards in 1982, but the typographical representation of faces was noted already in 1881. See http://www.cs.cmu.edu/~sef/Orig-Smiley.htm and Ptasyznski et al. (2011).

[20] There has been little research into variation between "Western" and "Japanese"–style emoticons. Takagi (1999) provides an overview these emoticons, known in Japanese as *kaomoji* (顔文字, lit. 'face–character'), and suggests that their origins lie in the 2–byte encoding scheme used for Japanese characters. Kawakami (2008) has test subjects rate the expressiveness of a set of Japanese–style emoticons in four affective categories. Arched–eyebrow emoticons, for example, are rated as strongly expressive (p. 70).

of emoticons in one tweet was 18. For the Finland English Corpus, 24.9% of all tweets contained at least one emoticon, and 56.1% of users represented in the Finland English Corpus used at least one emoticon. This suggests that Finland–based Twitter users who tweet in English are more likely to use emoticon symbols than Finland–based Twitter users who tweet in other languages. Overall, Finland-based Twitter users include emoticons in their messages significantly more often than do non–Finland based Twitter users. See the discussion below in Section 5.4.

TABLE 5.11: 20 most frequent emoticon types in the Comparison English Corpus, as percent of all emoticon types

	Type	Percent		Type	Percent
1	:)	32.3	11	:-(	1.8
2	:D	9.3	12	(:	1.4
3	:(	9.1	13	XD	1.1
4	;)	8.1	14	:p	1.0
5	:-)	7.7	15	=(	1.0
6	;-)	4.4	16	xD	1.0
7	=)	3.6	17	):	1.0
8	:P	2.7	18	>.<	0.5
9	:-D	2.4	19	♫	0.5
10	<3	1.9	20	(;	0.4

For the comparison data, the 20 most frequently used emoticons include the common smiley types as well as one Unicode figure (musical notes) and one Japanese–style emoticon (>.<) (Table 5.11). The total number of emoticon types in the Comparison English Corpus, as well as the distribution of types, was more limited than the Finland English Corpus: The twenty most frequent types in the Comparison English Corpus comprise 91.3% of all emoticon types in the corpus.

The prevalence of emoticons in the Comparison Corpus English was also more limited than in the Finland English data. Only 9.8% of the tweets in the Comparison English Corpus included at least one emoticon. 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. It may be the case that this large difference results from an increase in use of emoticons overall in Twitter in 2013 compared to 2008-9: There seems to be no research into the relative prevalence of emoticon type use in Twitter over time. Vandergriff (2014) finds much higher rates of emoticon use by Swedes in a corpus of online interaction comprising messages by Swedish and American university students. The analysis below and in Chapter 6 suggests that high rates of use of emoticons in Finland Twitter English is not a chronological artifact, but rather a characteristic feature of the variety.

Schnoebelen (2012) reports that the "smile" emoticon :) is the most frequent emoticon in a large dataset of American English tweets, accounting for almost 40% of all emoticon use. This is followed by the "wink" ;), the "frown" :(, the "bigsmile", :D, the "smilenose" :-), the "tongue" :P, the "slant" :/, the "xeyesbigsmile" XD, and the "eq eyessmile" =) emoticons.

Schnoebelen's analysis was replicated in part using the Finland English and Comparison English data; the findings are summarized in Table 5.12 and 5.5. For the most part, the Finland English data show a similar rank/frequency profile for the most widely used emoticons.[21] The 28 most frequent emoticons in the Finland English data and the Comparison Corpus English data were somewhat different than the reported most frequent 28 emoticons in Schnoebelen 2012. For comparison purposes, the relative frequency of the 28 emoticon types in Schnoebelen's set were investigated in the Finland English and Comparison English data. The percent figures give an idea of the relative likelihood of a particular emoticon being used compared to the other 27 emoticons in the set.[22]

[21]Unlike in this study, Schnoebelen seems to have considered only those emoticons that can be considered representations of faces, i.e. not Unicode symbols or other graphemes, although it is possible that no Unicode symbols or non-face emoticons were among the top 28 types tagged as emoticons in his dataset. Our extensive testing of the CMU Twitter Tagger suggests this is unlikely for a social media dataset. See Schnoebelen (2012: 118).

[22]The figures in Table 5.12 don't, however, (at least for our data), accurately reflect the proportion of emoticon tokens in the each corpus comprised by the particular type. That information (for the top 20 types) is summarized above in Tables 5.10 and 5.11.

TABLE 5.12: Relative Frequency of 28 Emoticon Types in Schnoebelen 2012, Finland English Corpus and Comparison English Corpus

Name of Emoticon	Symbol	Representation in					
		Schnoebelen 2012		FE Corpus		CE Corpus	
		Rank	%	Rank	%	Rank	%
smile	:)	1	39.6	1	40.2	1	36.3
wink	;)	2	10.5	3	11.8	4	9.1
frown	:(	3	8.3	4	9.1	3	10.2
bigsmile	:D	4	7.5	2	16.0	2	10.4
smilenose	:-)	5	4.9	5	6.6	5	8.7
tongue	:P	6	4.5	8	1.5	8	3.0
rsmile	(:	7	4.1	9	1.4	11	1.6
slant	:/	8	3.4	6	2.3	14	0.8
xeyesbigsmile	XD	9	3.0	7	2.3	12	1.3
eqeyessmile	=)	10	2.1	19	0.3	7	4.0
winknose	;-)	11	1.9	12	0.9	6	5.0
omouth	:O	12	1.6	20	0.3	26	0.1
winkbigsmile	;D	13	0.9	15	0.7	24	0.2
doublesmile	:))	14	0.8	10	1.4	21	0.3
frownnose	:-(	15	0.7	16	0.7	10	2.0
smileapose	:')	16	0.6	13	0.9	28	<0.1
dworry	D:	17	0.6	17	0.7	15	0.5
smilebrac	:]	18	0.6	23	0.1	27	0.1
eqeyesbig smile	=D	19	0.6	22	0.2	17	0.4
slantnose	:-/	20	0.5	24	0.1	16	0.5
eqeyesbrac	=]	21	0.5	27	0.0	20	0.3
winktongue	;P	22	0.5	18	0.3	25	0.2
tonguenose	:-P	23	0.4	21	0.2	18	0.4
frownapos	:'(	24	0.4	11	1.0	22	0.2
bigsmilenose	:-D	25	0.4	14	0.8	9	2.7
eqeyesslant	=/	26	0.4	25	0.1	19	0.3
eqeyestongue	=P	27	0.4	28	<0.1	23	0.2
eqeyesfrown	=(	28	0.4	26	<0.1	13	1.1

Overall, the distributions for the three data sets for the relative proportions of this specific set of emoticons are rather similar.[23]

[23]A paired Wilcoxon ranked sums test shows no significant difference between the median ranks of the relative frequencies of the 28 emoticons for the Finland English Corpus and the Comparison Corpus English data, the Finland English Corpus and the Schnoebelen data, or the Comparison Corpus English data and the Schnoebelen data. For Finnish and Comparison data V = 179, p–value = 0.60; for Finnish and Schnoebelen data V = 160, p–value = 0.34, and for Comparison and Schnoebelen data V = 172, p–value = 0.49.

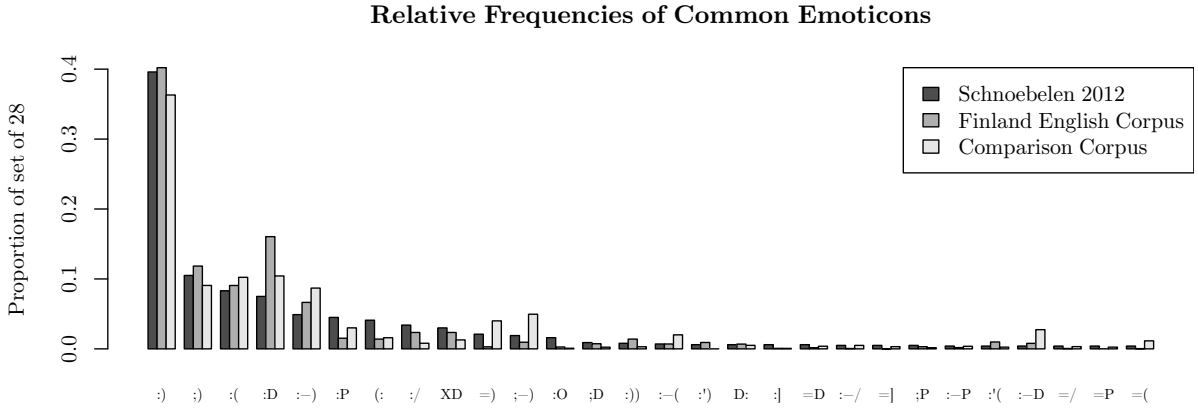

FIGURE 5.5: Relative Frequencies of Common Emoticons in Schnoebelen 2012, Finland English Corpus, and Comparison English Corpus

Figure 5.5 shows the relative proportions of the 28 Schnoebelen emoticons in the three data sets as a barplot. Despite the overall similarity in the distribution shape, some differnces are apparent. The "bigsmile" smiley (:D) is much more common in the Finland English data. The emoticons with tongue representations :P, :-P, ;P, =P are less common in Finland than in the other data sets, as are the emoticons with dash nose representations :-), :-D, :-P, ;-), :-(.[24] Comparison English Twitter users are 2.54 times more likely to use "dash nose" emoticons compared to Finland English Twitter users.[25]

Finland English users are, however, 3.02 times more likely to use emoticons with apostrophe noses such as :') and ;'(. "Japanese"–emoticons viewed along the y–axis are not considered in the Schnoebelen data. For the Finland English and Comparison English

[24]It should also be remarked that the CMU Twitter tagger is not good at recognizing multiple–dash-nose emoticons such as :--). Sequences such as this are tagged as two, three or four punctuation symbols. Emoticons of this type, however, are very infrequent in the data.

[25]For the set of dash nose emoticons comprising those which occur in both the Finland English and the Comparison English corpora, a Chi–square goodness–of–fit test with Yates' continuity correction test gives a value of $\chi^2 = 417.05$, p–value $< 2.2e^{-16}$.

corpora, Finland English tweeters are 1.51 times more likely to use any emoticon from the Japanese set. They are about half as likely to use the most common type ^_^, but much more likely to use other types such as ^^, o_O, O.o, O_o, or -_-.[26] Although this set of emoticons is said to have originated in Japanese online communication, the relative rates of use of Japanese/Asian emoticons and Western emoticons by different demographic groups has not been extensively investigated. The national flagship (and majority state–owned) airline Finnair began to offer direct flights to a number of East Asian cities from Helsinki in the late 2000s; it may be conceivable that Asian–style emoticons are used in part by Chinese or Japanese tourists or university students in Finland. However, the language distribution data from Chapter 4 does not suggest high rates of use of e.g. Mandarin or Japanese in Finland Twitter.

5.2.5.0.1 Gender Distribution of Emoticons

An analysis of gender correlation with emoticon use was conducted on the relatively small subset of the Finland English Corpus marked up for gender. In the section above, we found that Finland–based Twitter users who write in English are less likely to use dash–nose emoticons, and more likely to use apostrophe–nose emoticons and Japanese–style emoticons, compared to non–Finland users. These three categories are distributed differently within the set of Finland English users according to gender.

Emoticon use was also shown to vary according to gender. The gender–tagged portion of the Finland English corpus shows that females exhibit a mean rate of use of all emoticon types of 23.92 per thousand words, whereas males use emoticons 15.35 times per thousand words. On average, females used 0.290 emoticons per tweet, whereas males used 0.208 emoticons per tweet.[27] Females are 1.41 times more likely to use dash–nose emoticons and 1.31 times more likely to use apostrophe–nose emoticons. Interestingly, males in the

[26] Apostrophe emoticons: $\chi^2 = 96.05$, p–value $= 2.2e^{-16}$; Asian emoticons: $\chi^2 = 108.26$, p–value $< 2.2e^{-16}$.
[27] Highly significant according to a paired t–test; t $= 5.93$, p–value $= 3.19e^{-9}$.

Finland English corpus are 4.50 times more likely to use a Japanese–style emoticon than are females. The first two effects are, however, not significant due to the low counts in some cells of the contingency tables.[28] The third effect, while highly significant, seems to be due to the extensive use of an emoticon from this set by a single user.[29]

TABLE 5.13: Most Frequent Emoticons, Gendered Portion of Finland English Corpus

	Type	odds ratio θ	significance		Type	odds ratio θ	significance
1	:)	2.317	***	11	XD	2.727	
2	:D	2.136	***	12	♥	1.629	
3	<3	2.493	***	13	;-)	0.239	
4	;)	1.677		14	✌	3.110	
5	:(	2.151	**	15	:-D	3.015	
6	❤	2.189	**	16	:))	2.791	
7	:-)	2.043	*	17	☺	1.965	
8	]:	-0.030	***	18	:')	2.658	
9	^^	-3.218	***	19	:P	0.712	
10	(:	2.370		20	;3	-0.385	*
*** = $p < .001$; ** = $p < .01$; * = $p < .05$							

Table 5.13 shows the most frequent emoticons in the Finland English Corpus portion that is disambiguated for gender, along with the results of a χ^2 test.[30] Of the twenty most frequent emoticons used at least once by both males and females, seventeen are more likely to be used by females; some much more so. Discounting the effect of the ^^ emoticon (which has a very high male–female use ratio due to its frequent use by a single user), the emoticons that are slightly more "male" are the the :[and the ;3 emoticons.

Table 5.14 shows emoticons in the Finland English data that have the lowest odds ratio θ, i.e. the most disproportionate use by males. Only five of the types have a negative θ value. In most of these cases, low (< 5) cell counts result in χ^2 values that are not enough to achieve significance for the effect.

[28] For dash noses, $\chi^2 = 0.87$, p–value = 0.35; for apostrophe noses, $\chi^2 = 0.04$, p–value = 0.83.

[29] $\chi^2 = 60.14$, p–value = $78.84e^{-15}$.

[30] Low cell counts for some of the emoticons mean that the difference in gendered use does not achieve significance at $p \leq 0.05$ for 11 of the 20 most frequent emoticon types.

TABLE 5.14: Emoticon Types with the Lowest Odds Ratios, Gendered Subsection of the Finland English Corpus

	Type	odds ratio θ	significance
1	^^	-3.218	***
2	x3	-0.891	*
3	x_x	-0.673	
4	;3	-0.385	*
5	]:	-0.030	***
6	0_0	0.019	
7	D:	0.019	
8	;-)	0.239	
9	:P	0.712	
10	\o/	0.712	
*** = $p < .001$; ** = $p < .01$; * = $p < .05$			

Prior research has suggested that emoticons associated with the communication of sarcasm, flirting behavior, or negative affect are more likely to be used by males (Wolf 2000, Herring 2013). Discounting the effect of the ^^ emoticon due to its heavy use by a single user, many of the most "male" types in the Finland English Corpus seem to be associated with the expression of non–positive emotions: x_x and 0_0 may signal flat affect or bewilderment, whereas "winking" emoticons such as ;3, and ;-) can be associated with sarcasm or expressions of romantic or sexual interest (Herring 2013). The types]: and D: may mark negative affect such as disappointment or sadness. While the limited extent of the data does not provide conclusive evidence for gendered differences in certain types of emotional expression as demonstrated by emoticon use, it seems that prior findings suggesting that the use of emoticons as discourse markers of negative affect or sarcastic/romantic communicative function may also hold true for Finland–based males who compose messages in English.

Almost all the emoticons in the Finland English data are more used by females than by males. Table 5.15 shows those emoticons that are most heavily "female", i.e. the ones with the highest ratio of female to male use of those emoticons used at least once by both a male

TABLE 5.15: Emoticon Types with the Highest Odds Ratios, Gendered Subsection of the Finland English Corpus

	Type	odds ratio θ	significance		Type	odds ratio θ	significance
1	✌	1.686		11	:)	0.891	***
2	:-D	1.589		12	❤	0.765	**
3	:))	1.366		13	:(	0.727	**
4	XD	1.302		14	:D	0.712	***
5	:')	1.232		15	:DD	0.672	
6	<3	1.068	***	16	:-)	0.620	*
7	(:	0.947		17	☺	0.542	
8	-.-	0.896		18	;(	0.385	
9	:'(	0.896		19	xD	0.385	
10	o.o	0.896		20	;)	0.254	

**** = $p < .001$; ** = $p < .01$; * = $p < .05$*

and a female. Two of the emoticons in the set are Japanese–style; both are associated with neutral to negative affective states such as surprise, bewilderment or frustration. Of the remaining emoticons, four are Unicode or ASCII graphemes depicting positive affective states: the "peace sign", a Unicode heart symbol, a sideways heart composed of ASCII characters, and a Unicode smile symbol. The remaining 14 emoticons are faces with the familiar sideways orientation. Three of these twelve represent negative affective states: :(, ;(, and :'(, with the other nine representing positive affective states such as smiling or laughter. Herring (2013) reports that emoticons associated with negative affect are more used by males than by females. Our data suggests that for females in Finland tweeting in English, this is not always the case: Females in our data are willing to use the expressive resources of emoticons to represent negative or indifferent emotions as well as positive affect. On the whole, however, the most frequent emoticons utilized by females in Finland writing in English on Twitter tend to expressive positive affective states.

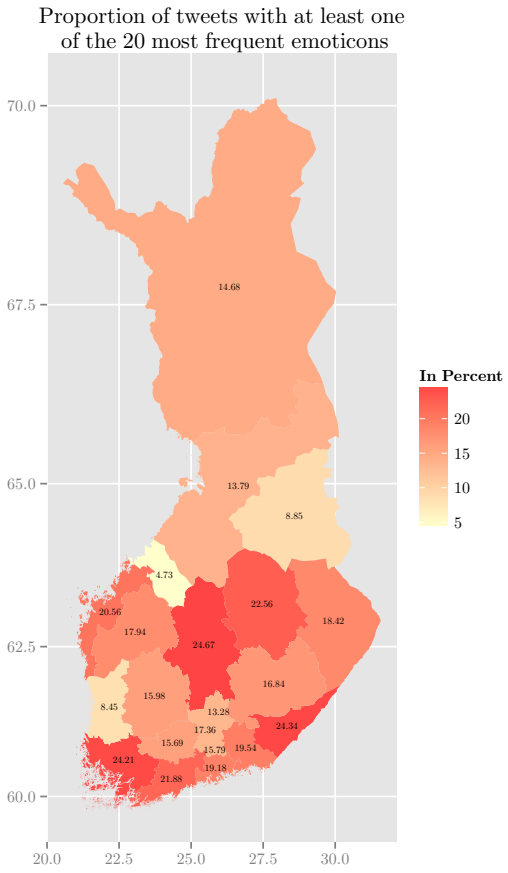

FIGURE 5.6: Emoticon Density, Finland English Corpus

5.2.5.0.2 Geographical Distribution of Emoticons

Figure 5.6 shows the geographical distribution of emoticon use in Finland tweets. The highest percentage of emoticon usage occurs in Central Finland, followed closely by Southern Karelia and Finland Proper.

The geographical distribution of the emotions occurring most frequently in the Finland data is shown in Figure 5.7 (to be read from top to bottom and left to right; note the gradient difference between the first five types, the second three types, and the remaining types). Users from Northern Finland, from the regions of Lapland, Northern Ostrobothnia

and Kainuu, seem to use graphical resources to indicate emotional or affective content somewhat less often than do other Finland Twitter users: 14.68% of Lapland tweets, 13.79% of Northern Ostrobothnia tweets, and 8.85% of Kainuu tweets contain at least one emoticon. With the exception of the regions of Ostrobothina (4.73%) and Pirkanmaa (8.45%), central and southern regions of Finland have higher rates of emoticon use, ranging from 15.48% in Päijänne Tavastia[31] to over 20% in Southern Ostrobothnia (a region with a large Swedish L1 population), Uusimaa, and North Savo, as well as Southern Karelia, Finland Proper, and Central Finland.

TABLE 5.16: Correlation between Selected Variables and Emoticon Density per Region

	Spearman's ρ	signif.
Russian tweets per capita	0.92	*
English tweets per capita	0.85	*
All tweets per capita	0.82	*
Other language tweets per capita	0.78	*
Finnish tweets per capita	0.64	*
University degrees 2012 per capita	0.61	*
GDP 2012 per capita	0.49	*
Number of Pupils per capita	0.43	
Polytechnic degrees 2012 per capita	0.38	
High school graduates 2012 per capita	0.38	
Swedish tweets per capita	0.37	
Vocational qualifications 2012 per capita	0.33	
Land area	-0.11	

A correlation test utilizing Spearman's ρ was conducted to quantify the relationship between the variables described in Section 3.1.4 and emoticon density in the 19 regions of Finland under consideration. The results are shown in Table 5.16. The results for seven of the demographic variables are significant at $p < 0.05$. Of the demographic statistics described in Section 3.1.4, emoticon density per region correlates most strongly with the per capita number of tweets in Russian. The correlation with tweets per capita is strong for English and other language tweets, but slightly less strong for Finnish language tweets;

[31] As described in Section 4.3, Päijänne Tavastia is represented as three distinct sections in the GIS polygons used to create the maps; the values for the three sections in Figure 5.6 are 13.28%, 17.36%, and 15.79%.

it is only moderate for Swedish language tweets. There are also fairly strong positive correlations between regional GDP per capita and regional number of university degrees granted per capita in 2012. The correlations between emoticon use and polytechnic degrees, pupils, vocational qualifications, or school leavers per capita are moderate. There is a slight negative correlation between emoticon density and region size. The patterning between the geographical distribution of specific emoticons and other emotional indicators within Finland is discussed below in Section 5.4.1.

Schnoebelen suggests that emoticons are "not simply representations of emotional states", but interactive units of communication which position authors and audiences with respect to the affective stance suggested in the proposition (2012).

For Finland, the data suggest that emoticons are an interactive resource used by relatively educated persons with higher-than-average mobility and access to resources.

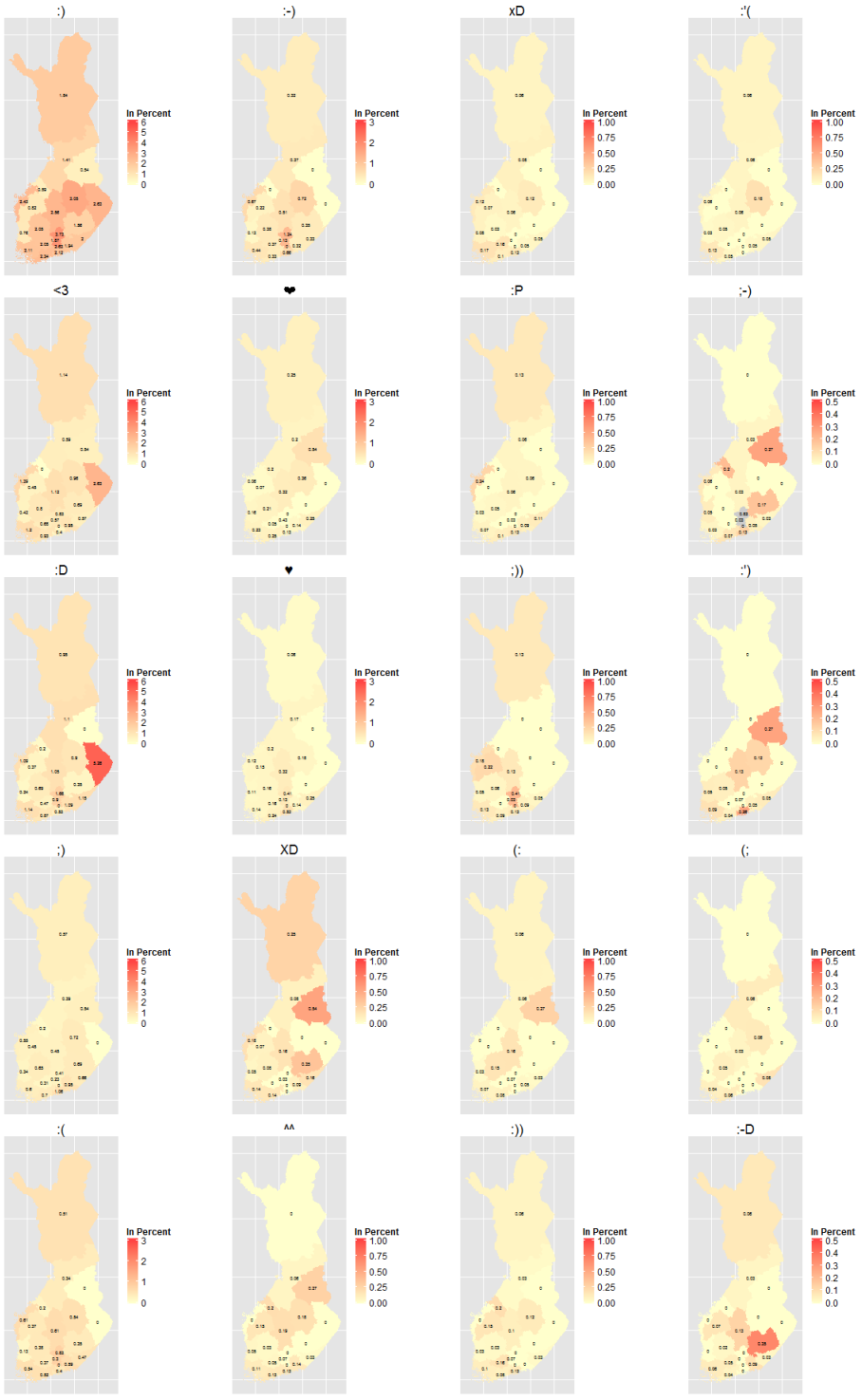

FIGURE 5.7: Geographical Distribution of the 20 Most Frequent Emoticons in the Finland English Corpus (Top to Bottom, Left to Right)

5.3 Grammatical Features

In this section, standard and non–standard grammatical features in the Finland English and Comparison English corpora will be considered. Non–standard orthography, whether the result of error or used as an expressive resource, is prevalent in Twitter and other CMC genres (Paolillo 2001; Tagliamonte and Denis 2008). The use of non–standard orthography can be shown to further disambiguate Finland Twitter English as a distinct variety. Frequencies of grammatical types (as instantiated by part-of-speech assignations) can be used to compare Twitter discourse with the discourse of well–studied genres/registers of spoken and written English. Based on frequency data, Twitter English, like spoken language and some types of CMC, can be shown to reflect an interactive, rather than an informational communicative orientation. Again, Finland Twitter English can be identified on the basis of part-of-speech frequencies, particularly for those grammatical tags that are most indexical of the platform Twitter as an interactive technology. It can be shown that the Finland Twitter English data confirms some prior results pertaining to the sociolinguistic correlation of feature frequency and author gender. Although linguistic interference phenomena between L1 Finnish and L2 (or L3) English may play a role in some of the observed frequencies for the Finland data, the investigation of grammatical features shows that for Finland Twitter English, users embrace the use of those features that are most characteristic of the genre.

5.3.1 Non–standard Orthography

Non–standard orthography is a characteristic feature of English messages on Twitter. In order to gauge the extent to which Finland English tweet messages compare with global English tweet messages in terms of non–standard orthography, the relative frequencies of some common non–standard spellings were calculated.

A list of 4,259 common orthographical errors in English was used in order to compare the frequency of non–standard orthography in the Finland English and Comparison English Corpora. This list is compiled and maintained at Wikimedia, and consists of misspellings that occur in Wikipedia articles at least once per year.[32]

The Finland English Corpus showed a rate of 1.90 occurrences of items from this set per thousand tokens, while the Comparison Corpus English data showed a much lower rate of 1.22 per thousand tokens.[33] Figure 5.8 shows the frequency of the twenty most frequent non–standard orthography word types from the set in the Finland English Corpus, with the Comparison English Corpus rate of occurrence alongside. For all types, the Finland data has substantially higher rates of use.

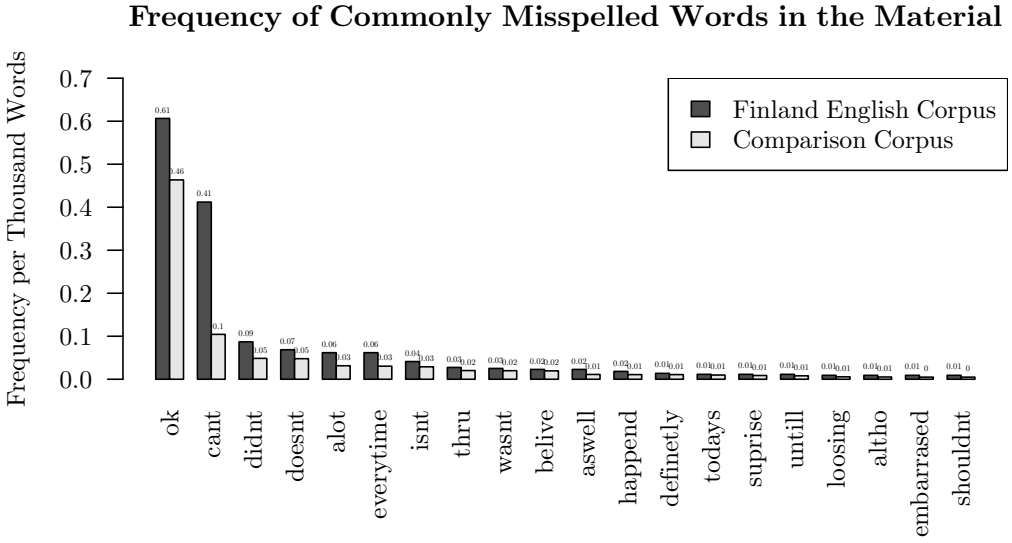

FIGURE 5.8: Most Frequent non-standard Orthography Types, Finland English and Comparison English Corpora

[32]Available at http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines.
[33]Significant according to a Chi-square goodness of fit test: $\chi^2 = 133.78$, p–value $< 2.2e^{-16}$.

Gender was also significantly correlated with non–standard orthography. Female users tweeting in English from Finland used non–standard lexical items from the set at a rate of 2.06 per thousand tokens, while males exhibited 1.23 instances per thousand tokens.[34]

Although geographical location for the tweets of the Comparison English Corpus is not available, the English tweets may be comprised primarily of messages by users for whom the first language is English. Topical content in the tweets suggests, in many cases, an American or British context for many of the Comparison English Corpus user messages. To that extent, differences in rates of non-standard orthography may simply reflect differential exposure to the orthographical norms of English – many Finland Twitter users writing in English may be using it as a second language. The set of non–standard orthographical items used for comparison certainly does not comprise the total set of non–standard forms in the Finland English Corpus data. Most of the items from the Wikimedia list represent lexical items that are used relatively infrequently (and are therefore presumably more subject to misspelling due to lack of familiarity with the standard orthographical variant). Non–standard orthography on Twitter, on the other hand, may represent functional differentiation. This is discussed in more detail in Section 5.4.

5.3.2 Expressive Lengthening

Expressive lengthening refers to a particular type of non–standard orthography: the repetition of individual characters in a word string (e.g. *cooooooool, yesssss, dumbbbb*). In previous research, the feature has often been interpreted as an affective discourse marker. Rao et al. (2010) and Bamann, Eisenstein and Schnoebelen (2014) find that this expressive resource is correlated with gender in Twitter, with female users more likely to employ the feature. The extent to which expressive lengthening use varies in geographically or socially distinct language varieties has not yet been the subject of research. To investigate the

[34]Significant according to a Chi-square goodness of fit test: $\chi^2 = 72.23$, p-value = 0.0007.

comparative use of this feature in the data, all cases in which a token contained at least three characters repeated in sequence were identified. Tokens containing the string "www." were filtered to remove urls. The most frequent types exhibiting expressive lengthening in the Finland English and Comparison English corpora are shown in Figure 5.9.

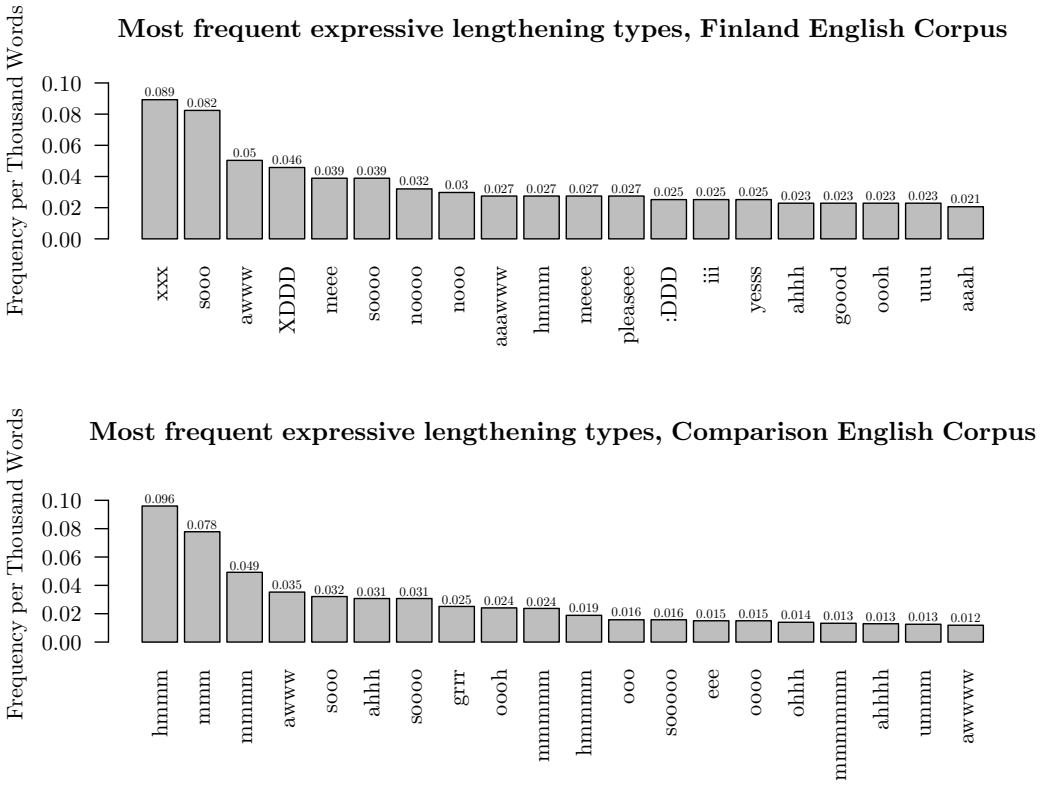

FIGURE 5.9: Most Frequent Expressive Lengthening Types

Many of the most frequent types, such as *awww, sooo, sooo, iii, ahhh* and *oooh* are shared by both corpora. The most frequent type in the Finland data is the non-pronounceable non-dictionary word *xxx*, commonly interpreted as representing kisses. Two emoticon types (*XDDD* and *:DDD*) are among the most frequent expressive lengthenings in the Finland data; the other types consist of lengthened dictionary words (*meee, sooo, nooo, noooo, pleaseee, iii, yesss, goood*) and lengthened pronounceable non-dictionary words (*awww, aaawww, hmmm, 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. The most frequent types in the Comparison English Corpus, on the other hand, are almost all pronounceable non-dictionary words (*hmmm, mmm, mmmm, awww, ahhh, grrr, oooh, mmmmm, hmmmm, ooo, eee, oooo, ohhh, mmmmmm, ahhhh, ummm*). Only *sooo*, *soooo*, and *sooooo* correspond to dictionary words. Thirteen of the twenty most frequent lengthening types contain 3–character sequences, four four–character sequences, two five–character sequences, and one a six–character sequence. A comparison of expressive lengthenings by letter lengthened and number of characters lengthened between the Finland English and Comparison English Corpora is summarized in Figures 5.10 and 5.11; a preliminary analysis is presented below, with more discussion in section 5.4.

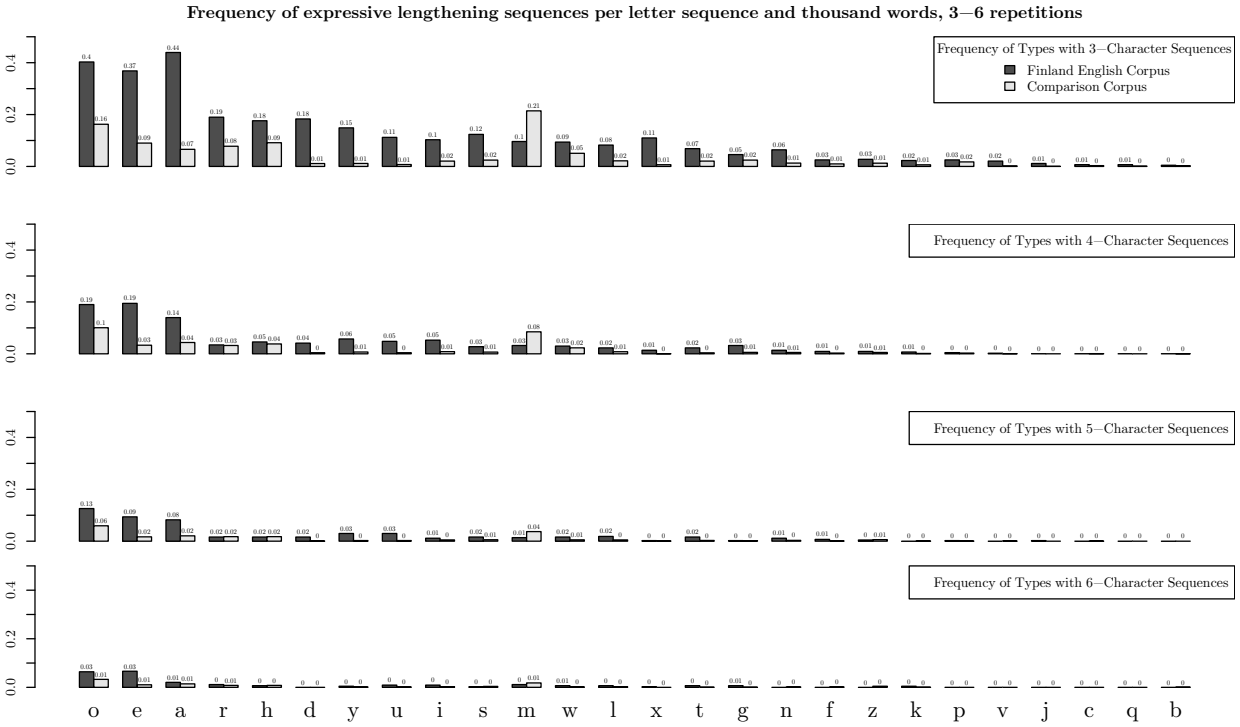

FIGURE 5.10: Expressive Lengthening by Letter, 3-6 Character Repetitions (per 1000 tokens)

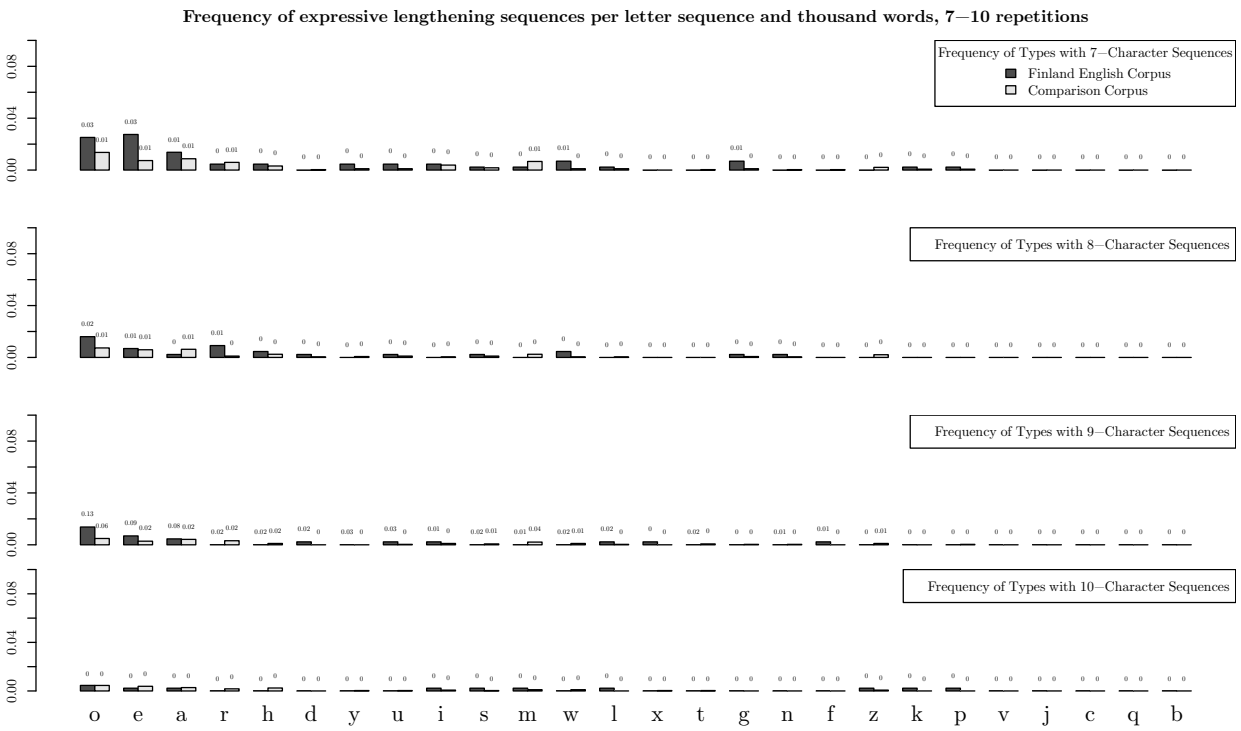

FIGURE 5.11: Expressive Lengthening by Letter, 7-10 Character Repetitions (per 1000 tokens)

In our data, Finland English users of Twitter are much more likely to employ expressive lengthening than are non-Finland English users: the total rate of tokens containing expressive lengthenings is 5.00 per thousand for the Finland English data and 1.86 per thousand for the Comparison English data.[35]

The characters that are most frequently lengthened differ somewhat for the Finland and non-Finland corpora.[36]

As Figures 5.10 and 5.11 indicate, for the Finland data, the letter *o* seems most susceptible to lengthening: the frequency of the sequence *ooo* is 0.40 per 1,000 words, *oooo*

[35] χ^2=1645.02, p-value $< 2.2e^{-16}$.

[36] A rank correlation using Spearman's Rank-Order Correlation Test gives a rather low value of $\rho = 0.015$, p-value = 0.94.

0.19, and *ooooo* 0.13 per thousand words. The overall frequency for lengthened sequences containing *o* is 0.84 per thousand words. Although the frequency of *aaa* is higher for 3–character sequences, *o* is more likely to be lengthened into longer sequences. 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 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. Lengthening itself seems to be subject to a length constraint: Among the most frequent 100 lengthening types in the Finland English data, 81 contain 3–character sequences, ten 4–character sequences, five 5–character sequences, three 6–character sequences, and one an 8–character sequence. This rank–frequency profile corresponds to the Zipfian distribution commonly exhibited when feature types are counted and sorted into frequency ranks; it may represents a balance between communicative economy considerations, individual stylistic preferences, and locally negotiated meaning; these are all manifest in phenomena of aggregate language use (Piantadosi et al. 2011; Kretzschmar 2009).

The Comparison Corpus English data shows lower lengthening frequencies overall, but also a somewhat different distribution of lengthening types and frequencies (letter rankings are given in Figure 5.12). The character most susceptible to lengthening is again *o*: The sequence *ooo* appears 0.16 times per thousand words, *oooo* 0.10, and *ooooo* 0.06 times. The overall frequency for lengthened sequences containing *o* is 0.37 per thousand words, less than half that of the Finland English Corpus. 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*). As in the Finnish data, the characters *h* and *r* are among

the most frequent targets for consonant lengthening; they are among the only characters for which the Comparison English Corpus exhibits some higher lengthening rates (for longer lengthenings ≥ 6 characters; this may be, however, an effect of corpus size). The relative status of *m* as a lengthening target is already apparent from the consideration of most lengthened word types: It is the most-lengthened consonant in the Comparison English Corpus and the second–most–lengthened character after *o*. The consonants *w*, *s*, *l*, *g*, *z*, *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.

Characters with higher ranks in the Comparison English Corpus relative to the Finland English Corpus include, in addition to *m*, the characters *w*, *i*, *s*, *l*, *g*, *z*, *t*, *n*, *p*, *c*, and *b* (ranks 7, 8, 9, 10, 11, 12, 13, 15, 16, 22 and 23 versus 12, 9, 10, 13, 16, 19, 15, 17, 21, 24 and 26). Characters with higher ranks in the Finland corpus are *e*, *a*, *r*, *d*, *y*, *u*, *x*, *f*, *v*, and *j* (ranks 2, 3, 4, 6, 7, 8, 14, 18, 22, and 23 versus 3, 4, 6, 18, 14, 17, 21, 19, 24 and 26). Relatively, it appears that consonants are more attractive targets for lengthening in the Comparison English data.

The distributions of lengthening sequences according to character suggest that 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 (Fry 1955, Klett 1976). Vowel graphemes 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, which merits further study.

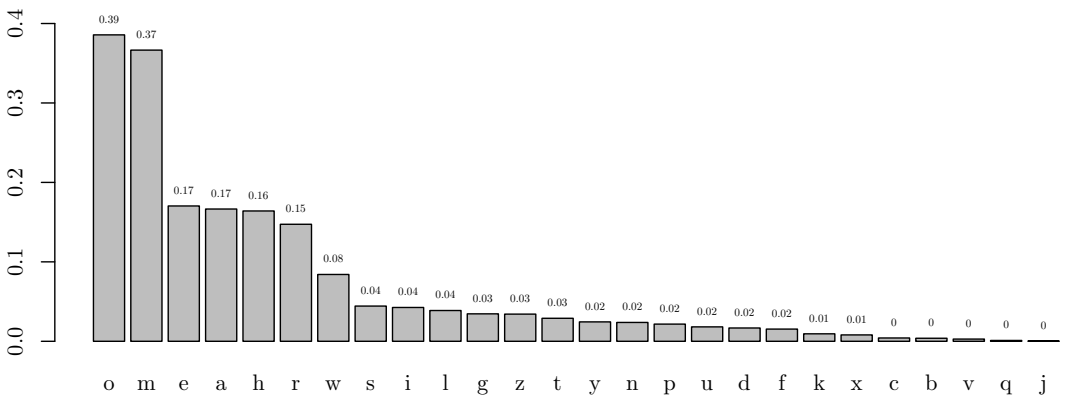

FIGURE 5.12: Expressive Lengthening, Comparison English Corpus

5.3.2.0.1 Expressive Lengthening: Gender Analysis

Males in the Finland English Corpus filtered for gender exhibited a rate of expressive lengthening of 1.88 per 1,000 words; females a rate of 4.05 per 1,000 words.[37]

Only a small number of lengthened types were used by both males and females in the Finland English Corpus. Of these types (*sooo, meee, seee, hmmmm, looooong, oooooh, pleaseeee*, and *sooooo*), none were used more frequently by males than by females. However, there were relatively few tokens in this group.

As far as is known, Schnoebelen (2012) is the only study that treats the phenomenon of expressive lengthening in some detail; he finds that the feature correlates with the use of particular emoticons (the "noseless" emoticons such as :) or ;)). The correlation between expressive lengthening and gender in the Finland English data has been briefly in-

[37]Significant according to a Chi-square goodness of fit test: $\chi^2 = 28.25$, p-value $= 1.01e^{-7}$.

troduced. The breakdown of lengthening types according to word type (dictionary words, non-dictionary pronounceable words, non–dictionary non–pronounceable words, or emoticons) and the interaction of this category with other linguistic variables, in both L1 and non–L1 Twitter English, would merit further study.

5.3.3 Grammatical and Part-of-Speech Features

In this section the relative prevalence of part-of-speech features in the Finland English and Comparison English corpora are examined. For each feature, examples from the corpora are provided, followed by a brief discussion of the feature frequency in the two corpora. Again, tests of statistical significance for the feature distributions in the two corpora assume equivalent distributions; significance is calculated by applying Pearson's χ^2 test for goodness of fit.[38]

5.3.3.1 Grammatical and Part-of-Speech Features by Geography

The Finland English and Comparison English data exhibit different distributional profiles for the relative frequencies for 37 grammatical feature tags as assigned by the CMU Twitter Tagger. The relative frequency of grammatical features in the Finland English and Comparison English corpora is shown in Table 5.17 as the logarithmic odds ratio θ of frequency in the Finland English Corpus to frequency in the Comparison English corpus. Items in boldface are substantially (20%) more or less frequent than in the other corpus. The differences in frequencies for each item are significant according to the results of a Chi–squared goodness–of–fit test based on a contingency table constructed from the number of occurrences and non–occurrences of each feature at $p < 0.001$, except for those features marked with an asterisk, which fail to achieve significance.

[38]For the calculation of the test statistic, see 5.1.

TABLE 5.17: FI–CC odds ratios for 37 grammatical features

	Feature	odds ratio θ		Feature	odds ratio θ
1	**Hashtag**	**3.360**	19	Cardinal number (*)	-0.020
2	**Retweet**	**1.678**	20	Adjective, comparative (*)	-0.072
3	**Username (preceded by @)**	**0.770**	21	Adjective	-0.083
4	**Interjection**	**0.751**	22	Preposition or subordinating conjunction	-0.105
5	**Personal pronoun**	**0.350**	23	Noun, plural	-0.150
6	**Wh-adverb**	**0.350**	24	Adverb, comparative	-0.150
7	**Verb, non-3rd person singular present**	**0.336**	25	Verb, past tense	-0.162
8	**Universal Resource Locator**	**0.223**	26	Comma	-0.174
9	**Adjective, superlative**	**0.215**	**27**	**Verb, 3rd person singular present**	**-0.198**
10	**Coordinating conjunction**	**0.207**	**28**	**Period (*. ? !*)**	**-0.210**
11	Adverb	0.173	**29**	**Noun, singular or mass**	**-0.235**
12	Modal	0.165	**30**	**Determiner**	**-0.235**
13	Wh-pronoun	0.157	**31**	**Proper noun, singular**	**-0.301**
14	Verb, base form	0.104	**32**	***to***	**-0.301**
15	Existential *there* (*)	0.086	**33**	**Wh-determiner**	**-0.314**
16	Adverb, superlative (*)	0.076	**34**	**Verb, past participle**	**-0.385**
17	Possessive pronoun	0.067	**35**	**Verb, gerund or present participle**	**-0.494**
18	Quotation mark (*)	0.019	**36**	**Other punctuation (*: ; ... + - = <> [ ]*)**	**-0.494**
			37	**Particle**	**-0.562**

Of the 37 grammatical features, 18 are more likely to be used in the Finland English data. Ten features are used substantially more frequently (i.e. difference in use more than 20/

5.3.3.1.1 Features More Common in the Finland English Corpus

The most Finnish feature by far is the hashtag, used in the Finland English Corpus at a rate almost 29 times that of the Comparison English Corpus. Examples can be seen in Table 5.18.

TABLE 5.18: Examples of Hashtags from the Finland English Corpus

	Text	Hashtag
1	Having a good time! #iSmoothRun	#iSmoothRun
2	First leg done, continuing on the road. #duathlon http://t.co/qf3cc6B38P	#duathlon
3	#SongsIWillAlwaysLove You Know I'm No Good, Minority, Creep .. IDK too many!!	#SongsIWillAlwaysLove
4	ice cream in spring #awesome you know it #refreashing	#awesome #refreashing
5	Easter finally #hangin #with #gals #siltamäki @Janskun Terde ? http://t.co/FubvIwcxlw	#hangin #with #gals #siltamäki

Hashtags are used primarily as overt indicators of topic, but are sometimes also used with multiple words in a longer phrasal or clausal sequence, as in (5) in Table 5.18. The relative lack of use of hashtags in the Comparison English Corpus is likely due to the fact that the 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 Twitter user homepages (Bowman 2010); this interface change by Twitter prompted a rapid increase in the prevalence of hashtags. The communicative function of hashtags in Twitter has been interpreted variously (Zappavinga 2011, Wikström 2014). The high rate of hashtag use in the Twitter English Corpus may exemplify the technological and social factors that have led to Twitter adoption in Finland (see below, Sections 5.4 and 5.4.2).

Retweets, or the re-posting of an already posted tweet, are more than five times more common in the Finland English Corpus than in the Comparison English Corpus. Some examples are shown in Table 5.19.

Table 5.19: Examples of Retweets from the Finland English Corpus

	Text	Retweet
1	@themonkeyOG i love u. Rt if u see this	Rt
2	@themonkeyOG Rt if u see this okay??	Rt
3	@themonkeyOG make me happy and rt this!! I loove u!	rt
4	@themonkeyOG I think im just dreaming and i got rt. I have to wake up.	rt
5	I rt bc i need to check tyem out later	rt

Retweets are the re-posting of an already posted tweet; these are appended the sequence *RT* by the Twitter interface. The sequence *rt* is also employed by users referring to retweets or requesting that others re-tweet their own messages. The frequencies determined for the data reflect the number of times the sequence *rt* appears as a word token in the corpora, but the CMU Twitter Tagger does not verify that the content of the message which includes the sequence *rt* corresponds to the content of other user messages. *RT* is more than five times more common in the Finland English Corpus than in the Comparison English Corpus. In the first four examples in Table 5.19, *rt* is an abbreviation for retweet, but

in the fifth example, the sequence *rt* may represent an abbreviation of *right*, and not an abbreviation of *retweet*. In some cases, such phenomena may lead to overestimation of the actual frequency of retweets. However, there is no reason to assume that this would affect the ratio of use between the two data sets.

Finland-based Twitter users tweeting in English are more than twice as likely as Comparison English users to utilize usernames preceded by the @ symbol. They are also more than twice as likely to utilize interjections and other types indicating affective or emotional content, such as emoticons. Some examples of this category from the Finland English Corpus are seen in Table 5.20.

TABLE 5.20: Examples of Interjections from the Finland English Corpus

	Text	Interjection
1	@plasticcupp Also when you really want one TT_____TT	TT_____TT
2	60 year old grandmothers on the bus, Skype isn't a new thing. It's not magic. Keep your pants on please.	please
3	Vanilla yoghurt, bananas and grapes. OMFG	OMFG
4	Absolutely stunning photos ;-)	;-)
5	Everybody are like "Omg yiss holidays yaay a trip yaay booze!" And i'm like "I'm ... i saw a new pic of kris."	Omg yiss yaay yaay

The first example in Table 5.20 has a Japanese–style emoticon, possibly representing consternation. In the second example, the politeness marker *please* takes the interjection tag. Initialisms from this category are seen in texts 3 and 5. Text 4 features a classic sideways emoticon with a dash nose.

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. Examples are seen in Tables 5.21 and 5.22.

Non-3rd-person present singular verb forms are more common in the Finland English Corpus than in the Comparison English Corpus by a factor of 1.4. In the examples in Table 5.23, most of the forms are seen to be first-person usages. Second-person forms and the non-inflected form *be* are also present.

TABLE 5.21: Examples of Personal Pronouns from the Finland English Corpus

	Text	Personal Pronoun
1	if money dont change u.... u need to get more	u u
2	Yo are the stores open bc i was gonna go to herushinki tomorrow	i
3	Ohh wait a sec it's thursday. I'm going on saturday lol	it
4	I need to phone anga. Nhhh.	I
5	I should have bought you flowers and held your hand, should have gave you all my hours when I had the chance.	I you you I

TABLE 5.22: Examples of Wh-adverbs from the Finland English Corpus

	Text	Wh-adverb
1	"@XSTROLOGY: #Leo's are always right even when they are wrong."	when
2	@steve the mx is a pretty decent switch when it wants to be one, lots of deploys before ex came	when
3	Why do kyut left liverpool fc ? Commeant and @LFC folleo me ;) YNWA	why
4	@moyashi_78: people tweet when they go to the toilet? that's kinda weird to me made me laugh	when
5	@Austin_Support_ when I read your #imagines it feels like that's really going to happen	when

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. Examples of the latter two can be seen in Tables 5.24 and 5.25.

5.3.3.1.2 Features More Common in the Comparison English Corpus

Fourteen features are used in the Comparison English Corpus at a rate at least 1.2 times that of the Finland English Corpus. Comparative adjectives are 1.21 as common in the Comparison data. Examples are shown in Table 5.26.

Past tense verb forms are more common in the Comparison English data by a factor of 1.21. Some examples are shown in Table 5.27.

Commas are also 1.21 times more common in the Comparison English data. 3rd-person-present singular verb forms are 1.22 times as common in the Comparison English Corpus as in the Finland English Corpus. Examples are seen in Table 5.28.

TABLE 5.23: Examples of Non-3rd-person Present Singular Verbs from the Finland English Corpus

	Text	Non-3rd-person present singular verb
1	@justinbieber I know that u never notice me!! mut i still love u allways!!<3 u r my inspiration!!!;)<3	know notice love r
2	Time to fly south again, this time not in suit though. But in Batman underwear!! No, not really...but I wish I did. #TDK	wish
3	How much more bored could I be...	be
4	@AustinMahone ohhh that's cool but you wasted a watermelon i could've eat!!!!! i love you haha	wasted love
5	@acharya2 thanks but personally sonr believe in luck but I desperately need it :-)	need

TABLE 5.24: Examples of Superlative Adjectives from the Finland English Corpus

	Text	Superlative adjective
1	Worst day of my life.	worst
2	The best part of my day has to be @mustardmon 's reactions to Luhan's manliness.	best
3	May Allah take my parents to highest jannah?	highest
4	yeah the white/orange one is the best he's like my cat repe„ that's why he's the coolest	best coolest
5	I often search for "cat" tweets those are the best	best

The punctuation items period, exclamation point and question mark are more common in the Comparison English Corpus by a factor of 1.24. Singular noun forms are more common by a factor of 1.26. Determiners such as articles and demonstrative pronouns are more common in the Comparison English data by a factor of 1.29. Some examples are shown in Table 5.29.

The Comparison English Corpus exhibits a rate of use of the word *to* and of proper nouns of 1.29 times that of the Finland English Corpus. Some examples of singular proper noun use in the Comparison English Corpus are shown in Table 5.30.

Wh-determiners,[39] past participles, and present participles or gerunds are more common in the Comparison English data, by factors of 1.42, 1.48, and 1.49, respectively. Examples of these categories are shown in Tables 5.31, 5.32 and 5.33.

[39] The Penn conventions state that "What and Which are tagged as Wh-determiner (WDT) when NOT acting as the head of a Wh-noun phrase, otherwise (at the head) they are tagged as Wh-pronoun (WP)" (http://www.ling.upenn.edu/histcorpora/annotation/pos-wh.htm; Marcus, Santorini, and Marcinkiewicz 1993). *That* is also tagged as a Wh-determiner if not an NP head (Santorini 1995: 6).

Table 5.25: Examples of Coordinating Conjunctions from the Finland English Corpus

	Text	Coordinating conjunction
1	@ddlovato FINLAND IS RIGHT NEXT TO IT PLEASE COME BY TRAIN AND LEAVE THE TRAIN IN VAINIKKALA thank u. :D x	and
2	@steve srx5800 is a mx960 in reality although target audience and features are different :)	and
3	I'm at Ilves Bar & Night (Tampere) w/ 5 others http://t.co/VM2SG1ezup	&
4	Watching Lesbian Vampire Killers with my dad and two uncles. Awesome family moment enabled by #Netflix	and
5	@steve to be honest the mx is a much better switch than any of the ex ones in terms of features and what it can do :-)	and

Table 5.26: Examples of Comparative Adjectives from the Comparison English Corpus

	Text	Comparative adjective
1	@jasonarredondo we would love to have you! The more the merrier, or scarrier!	merrier scarrier
2	Better turn in. I'm heading to Dublin tomorrow for the Irish Web Awards.	better
3	I went nuts in the bookstores in London, so thrilled was I to be surrounded by English language books. I'm set for a mont or more.	more
4	For all the crazed hype about one candidate or the other, I think America has never had a poorer selection of candidates in my lifetime.	poorer
5	PicoWiki has collected twice as much in donations in two weeks than a much bigger site for a client earned from adsense in a year.	bigger

Items receiving the *other punctuation* tag, such as the colon, the semi-colon, the ellipsis and symbols for basic mathematical operations are 1.52 times more common in the Comparison English Corpus than in the Finland English Corpus. Finally, the least "Finnish" grammatical feature is the particle, the prepositional component of phrasal verbs, which is 1.82 times more common in the Comparison English Corpus. Examples are shown in Table 5.34.

5.3.3.2 Grammatical and Part-of-Speech Features by Gender

Male and female Twitter users from Finland tweeting in English differ significantly in terms of their relative use of some grammatical features and lexical items. The relative frequency of grammatical features for males and females is shown in Table 5.35, along with the results of a Chi-square goodness-of-fit test based on a contingency table constructed

TABLE 5.27: Examples of Past Tense Verb Forms from the Comparison English Corpus

	Text	Past tense verb
1	@Rockyblizzard haha.we were out in cyberjaya on Monday..got scolded by MCMC! anyway...goodies? do let us know what shows u'd like to see ya	were got
2	@pandamerv yeah, MCMC suspended our late nite chat programs.:(Got some kickass late nite show being developd in place. hope test in Dec	suspended
3	@Rockyblizzard yeah those shows wld b good. I sent your thoughts to the Brand team. thnx	sent
4	M'sia MCMC Q1: Internet Reach grew frm 18% to 21% - 60% of Malaysians connected -12million dialup - 4.4m Broadband (or wat passes for it!)	grew
5	8TV's Latte@8 tonite! live from Starbucks , Sunway Lagoon. Come along, join the audience..free coffee i heard! on 8TV at 11pm	heard

TABLE 5.28: Examples of 3rd-person Present Singular Verb Forms from the Comparison English Corpus

	Text	3rd-person present singular verb forms
1	McCaint seems angry, combative and deluded.	seems
2	Who is writing McCaing is very intellectual on the issues? We must be watching a different debate	is
3	why is the media not talking about the serious ethics problem that she has? Abuse of power for personal gain is very bad.	is has is
4	watching Obama in Toledo Ohio as he lays out his new Econ rescue plan	lays
5	McCain has been running a negative campaign almost from the beginning despite his lie that he wouldn't.	has

from the number of occurrences and non-occurrences of the feature by gender in all tweets filtered for male and female.

In the Finland English data, of 37 grammatical features, males are more likely to use 17 features. Nine features are used substantially more frequently (i.e. have a male-female ratio ≥ 1.2); for eight of these the difference in male/female use is significant at p–value < 0.001. 19 features are more likely to be used by females than by males; of these eight are substantially more likely to be used by females (female-male ratio ≥ 1.2). The difference in use by females versus males is highly significant (at p-value < 0.001) for six of these eight features and significant (at p-value < 0.01) for one other.

TABLE 5.29: Examples of Determiners from the Comparison English Corpus

	Text	Determiner
1	anyone hear weird audio on 8TV during the weekend? like tape slowing down momentarily? if so were u watching on astro? wat prgrm affectd?	the
2	@pandamerv Thanks for the headsup....will review back the 'onair capture' and see.	the the
3	8TV had a strong month online in Sept. 13.2m pageview, 2.7m vid views. Not quite record, tho Media Prima did - 30.02mil vidviews! Amazing	a
4	@pandamerv yeah, MCMC suspended our late nite chat programs.:(Got some kickass late nite show being developd in place. hope test in Dec	some
5	@Rockyblizzard yeah those shows wld b good. I sent your thoughts to the Brand team. thnx	those the

TABLE 5.30: Examples of Singular Proper Nouns from the Comparison English Corpus

	Text	Singular Proper Noun
1	4e Ventures is looking forward to attending Innovision Awards tonight in Greenville	4e Ventures Greenville
2	8TV birthday today! we're 5yrs old. seem like yestrdy that Izam was pushing the button with Chairman, worrying tat he shlda worn a suit!	8TV Izam Chairman
3	Great Screenings last nite. Our CEO Ahmad Izham is a real comedian! should be up on website v soon. oh, rocky...Fringe on ntv7 nxt yr	Ahmad Izham ntv7
4	BJ's apparently carries over 20 Belgian Beers! Mmmm, Chimay Rouge.	BJ's Chimay Rouge
5	tomrw is MPB Screenings. A presentation to market of new shows, plans and strategies for 09. 8TV CEO alwys gets a starring role - yay!	MBP 8TV

5.3.3.2.1 Features More Commonly Used by Males

In the following section, some user message examples of the features that are substantially more "male" than female and vice-versa in the Finland data will be considered.

The most "male" feature is the existential *there* construction, used three times as often by males as by females in our data.[40] The effect is significant ($p < 0.001$) using a Chi-square test of proportions and applying the Yates continuity correction. Examples from the data are shown in Table 5.36.

Universal resource locators (web addresses) and hashtags are used 1.8 times more often by males than by females in the Finland English data. Singular proper nouns, such as names of places, internet services, apps or video games, are used 1.75 times more often by males than by females.

[40] This is a stative construction consisting of the demonstrative *there* and a form of the copula verb *to be*; Biber (1988) associates the use of this feature with a "static, informational style common in writing" (228).

TABLE 5.31: Examples of Wh-determiners from the Comparison English Corpus

	Text	Wh-determiner
1	finishing up with the work that supports me - on to the website development side of life.	that
2	@onlineteacher thx for rec on Nova Mind. How will you use it? Which U of CA holds your time? What online teaching do you do? =curious day	Which What
3	People are always busy AND interested in new ideas - hard to know at the beginning of a project which will win out as it moves forward	which
4	Finishing signs for Transition Towns Kinsale Tidy Towns Spring Fair, which is coming up on Saturday. More blogs to come.	which
5	I'm looking for websites/communitites that specialize in presentations. I know about slideshare but looking for one that will take flash	that

TABLE 5.32: Examples of Past Participles from the Comparison English Corpus

	Text	Past participle
1	Headed to take Xavier to school.	
2	@danieltyack Do you have any specific scotch that you recommend? I'm headed out in a bit to pick up a bottle for us.	headed
3	The ultrasound is done. Now we wait again.	done
4	The CT scan is done. They estimated 20 minutes to get the results, then we'll see a doctor again.	done
5	Looking for Web Developers/Designers that are interested in writing articles for http://webdevnews.net Contact me: http://tinyurl.com/6qr5an	interested

Male Twitter users in Finland are much more likely to use some types of punctuation, such as the colon, the semi-colon, basic mathematical symbols, the ellipsis, or square brackets than are females. These punctuation symbols are used 1.75 times more by males than by females; the effect is highly significant. Males also use punctuation symbols tagged with the "Period" tag (comprising the period, question mark, and exclamation mark) slightly more often than do females (1.07:1, close to significance at p-value = 0.112). Quotation marks are used slightly more often by males than females, and commas less often by males than by females; the effects, however, are not significant for these two features. The potential discursive function of the variation in these types of punctuation is discussed below.

Males are more likely to use hashtags than are females. Typically, hashtags are explicit topic indicators or keywords that Twitter users utilize in order to organize content and communicate topical concerns to other users. They can also, however, be used in other

Table 5.33: Examples of Present Participles from the Comparison English Corpus

	Text	Present participle
1	Is attracting and retaining staff a problem across all industries?	attracting retaining
2	Still looking for good writing talent in NZ	looking writing
3	Distributor SPI NZ has gone into receivership. No comment yet, except that staff are trying to absorb the news.	trying
4	so happy we just had a 3-day weekend. Keep thinking it's Monday & then get a nice surprise.	thinking
5	Time to head home for a nice glass of wine. Looking forward to a new British show premiering in NZ tonight - Confetti.	premiering

Table 5.34: Examples of Particles from the Comparison English Corpus

	Text	Particle
1	at 531 defense at 80, not getting hit enough for my base defense skill to go up those last 2 points >:(	up
2	@renatawc You play a rogue, do you have any idea what AToL would be referring to? I cannot for the life of me figure it out, thanks :)	out
3	Just touched down in Phoenix.	down
4	Getting all packed up. We fly out of Boston back to Phoenix in a couple hours.	up
5	On the "T" headed out to enjoy the day.	out

ways, for example as discourse organizational particles with functions such as "hedging, disclaiming, and managing face" (Wikström 2014: 149).

Cardinal number tags were applied to tokens containing numerals as well as the word forms of cardinal numbers such as one, two, etc. Males are more likely to use these types (for a consideration of number tokens, including ordinal and cardinal forms, in a Finland/non–Finland comparison, see below, Section 5.3.4.2.4).

The relativizers given the Wh–determiner tag, such as *which, what,* and *that*, are used 1.3 times more often by males than by females in the Finland English data. The effect, however, does not achieve significance according to a Chi-squared test, due to the low frequency count in one cell of the contingency table. Males are more likely by a ratio of 1.24:1 to use prepositions than are females in the Finland English data. Some examples are shown in Table 5.37.

TABLE 5.35: Male–female odds ratio θ for 37 features in the Finland English Corpus

	Feature	odds ratio θ	signif.		odds ratio θ	signif.	
1	**Existential *there***	**1.101**	***	19	Comma	-0.020	
2	**Universal Resource Locator**	**0.587**	***	20	Verb, gerund or present participle	-0.061	
3	**Proper noun, singular**	**0.559**	***	21	Adverb, comparative	-0.072	
4	**Other punctuation (*: ; ... + - = <> [ ]*)**	**0.559**	***	22	Determiner	-0.083	**
5	**Hashtag**	**0.500**	***	23	Verb, 3rd person singular present	-0.094	
6	**Cardinal number**	**0.314**	***	24	*to*	-0.116	*
7	**Wh-determiner**	**0.262**		25	Adjective, superlative	-0.139	
8	**Preposition or subordinating conjunction**	**0.215**	***	26	Verb, base form	-0.150	***
9	Noun, plural	0.182	***	27	Adverb	-0.150	***
10	Adjective, comparative	0.131		28	Verb, past tense	-0.186	***
11	Period (*. ? !*)	0.067	*	**29**	**Modal**	**-0.235**	***
12	Verb, past participle	0.067		**30**	**Username (preceded by @)**	**-0.274**	***
13	Noun, singular or mass	0.058	**	**31**	**Wh-pronoun**	**-0.371**	**
14	Quotation mark	0.058		**32**	**Wh-adverb**	**-0.385**	***
15	Coordinating conjunction	0.029		**33**	**Interjection**	**-0.385**	***
16	Adverb, superlative	0.019		**34**	**Verb, non-3rd person singular present**	**-0.400**	***
17	Adjective	0.009		**35**	**Possessive pronoun**	**-0.415**	***
18	Particle	0.000		**36**	**Retweet**	**-0.430**	
				37	**Personal pronoun**	**-0.462**	***

**** = $p < .001$; ** = $p < .01$; * = $p < .05$*

Finally, males are 1.2 times more likely to use plural noun forms than are females; examples are shown in Table 5.38. The male-female difference in use of prepositions and plural nouns are both highly significant.

For a number of grammatical features, the rate of use for males is slightly higher than for females. This is the case for, for example, major word classes such as comparative adjectives (1.14:1 but the effect is statistically insignificant); the punctuation marks ., *?*, and *!* (1.07:1, p-value = 0.112); past participles (1.07:1, insignificant); singular or mass nouns (1.07:1, p-value = 0.0003); quotation marks and coordinating conjunctions such as *but* or *and* (1.06:1 and 1.03:1, both insignificant); and superlative adjectives and adjectives (1.03:1 and 1.01:1, both insignificant).

In the Penn Treebank part-of-speech conventions, particles are the prepositional component of phrasal verbs such as *to go out*, *to come across*, *to wake up*, or *to cross off*.

TABLE 5.36: Examples of Existential *there* from the Finland English Corpus

	Text	Existential *there*
1	G.I. Joe 2 was entertaining but kind of shallow on the story. It must have sucked really bad before the rewrite ;-) Hope there isn't a 3rd.	there isn't
2	Oh btw i'm just playing with FCPX and GoPro footage˜ There will be slow/fast motion footage and also something to see 4K if i get it right.	there will be
3	There is nothing new and interesting going on in the world of music Completely wrong. Listening new #theknife album.	there is
4	Nordically Approaching Kenya samples @NordicApproach Hunkute. @NikiLeskinen yes there is some left for you http://t.co/K3TLp6BwjJ	there is
5	Let me reveal that there will be timelapses at least.	there will be

TABLE 5.37: Examples of Prepositions from the Finland English Corpus

	Text	Preposition
1	Sukumizu ninja costume could be nice ;3 with proper foot wear and socks and that red scarf and mask. Oh and i want katars or dual tonfas!	with
2	I think i will see how it is when i have it and if nothing intresting i will or may get rid of it finally. Then it's just twitter and g+	of
3	What happened to idea of our twitter RO group?	of
4	Well it's not even 100% that i go... maybe i know when there is only few weeks before it :3 damn and my date would been month before it ;_;	before, before
5	Humm i really wonder what should i do when con comes... i think i'm quite ok by then but i need somewan there to be with and hotel room	with

These particles (and hence phrasal verbs in general) are used at the same rate by males and females in the Finland English Corpus.

5.3.3.2.2 Features More Commonly Used by Females

Other grammatical features which can be easily disambiguated and classified by automatic taggers according to word form or word order are used more often by females. Rates of use by females versus males are slightly higher for commas (1.03:1, insignificant); present participles and gerunds (1.06:1, insignificant); comparative adverbs such as *better* or *more* (1.08:1, insignificant); determiners (1.09:1, significant at p-value = 0.0008); third-person singular present verb forms (1.10:1, insignificant); the word form *to* (1.12:1, significant at p-value = 0.037), superlative adjectives (1.15:1, insignificant), verb base forms (1.16:1, significant at p-value = $1.27e^{-5}$), and adverbs (1.17:1, significant at p-value = $2.11e^{-7}$).

TABLE 5.38: Examples of Plural Nouns from the Finland English Corpus

	Text	Plural Noun
1	Once again i feel vlog things just doesn't work for me. I want to move back for more professional style i had long time ago.	things
2	@eve_19931001 Feeling hatred inside and do not care much about things annymore and she hates her own child ;_; and has evil but sad eyes :3	things, eyes
3	@eve_19931001 at least i feel from inside =D but damn oh... nothing i can't spoil things =D u need to watch it! Only 12h episodes hey!	things
4	@Viconno @Rindesu_ have you planned anything yet like where u stay over nights? Hotel or outside or none of them 0w0	nights
5	Humm now need to plan my next moves :3 and watch anime so yeah.	moves

Ten features are used substantially more often by females compared to males (female-male ratio $\geq$ 1.2:1). Significance testing shows that the difference in use is significant at p-value < 0.001 for eight of these features and at p-value < 0.01 for one other feature.

Past tense verbal forms are used 1.21 times more often by females than by males in the Twitter English data.[41] Examples are shown in Table 5.39.

TABLE 5.39: Examples of Past Tense Verbal Forms from the Finland English Corpus

	Text	Past tense verb
1	I would be the happiest girl in the world if you noticed me!!? Please make my dream come true! @Austin-Mahone http://t.co/Yu3AJah0gX	noticed
2	Such a long day today,worked and then strait to game vs jazz. Great start 2:0 but we lost in the end...so pissed! Hope new guy can help	worked, lost
3	Bus got here as fast as Kimi Räikkönen. He clearly knew what he was doing and I decided not to give advice. http://t.co/leUfzvOwyO	got, knew, decided
4	I just became the mayor of Ylevän Tietoisuuden Kehto on @foursquare! http://t.co/4r9dfbAMLd	became
5	Hahaha I made it on time. :> And now I'm getting some food from Hesburger, hehehe.	made

Females are more likely to use modal verbs by a ratio of 1.27:1. Examples are shown in Table 5.40.[42]

Usernames, preceded by the @ symbol, are more likely to be used by females by a 1.31:1 ratio.[43]. Some examples from the data are shown in Table 5.41.

[41] p–value = 0.00073.
[42] p–value = 0.00023.
[43] p–value $< 2.2e^{-16}$

TABLE 5.40: Examples of Modal Verb Forms from the Finland English Corpus

	Text	Modal verb
1	@Harry_Styles i would give anything for hug. Ily. X	would
2	I believe I can fly. We will fly together with Laku! <3	can
3	i can't wait to see my cats i'm sure they have missed me	will
4	@Real_Liam_Payne Woukd you please follow me it would mean a world to me! </3 :) Xx	can't
5	@JaiBrooks1 yes u should cos im here too honey	woukd

TABLE 5.41: Examples of Usernames from the Finland English Corpus

	Text	Username
1	@AustinMahone my dream is for you to follow me and RT me?	@austinmahone
2	@dephunkt aw thank you!!?? ;__; made my day!<3	@dephunkt
3	@McWill99 @alan_elsworth hahah that' s perfect!	@mcwill99
4	@Psych_USA loveing shwan and gues dance and great choice for z #100thepisode awesome show ever #PsychWhoDDit	@psych_usa
5	@bikemadcarl ahh yes, those too!	@bikemadcarl

Females are more likely than males to use Wh-pronouns such as *who* or *what*, by a ratio of 1.44:1.[44] Some examples from female tweets in the Finland English Corpus are shown in Table 5.42.

TABLE 5.42: Examples of Wh-pronouns from the Finland English Corpus

	Text	Wh-pronoun
1	Idk why I'm laughing, but there's a button in the bus, and on the button it reads "retarder" ?? Idk what that even means.	what
2	@selandemi totally agree! That what happened in Greg's wedding = way too far!	what
3	Here's some of what I baked today. Fun times! @ Keijumäki http://t.co/a38xuDuBcu	what
4	@justinbieber Hey u my bro! What's up? ??	what's
5	@eepings I certainly have :) and from what ive seen from twitter and fb, you are enjoying yours as well :) see you soon again!! :)	what

Females are also more likely to employ Wh–adverbs such as how, when, or why than are males, by a ratio of 1.47:1.[45] Corpus examples are shown in Table 5.43.

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

[44] p–value = 0.00127.
[45] p–value = $1.12e^{-5}$.

TABLE 5.43: Examples of Wh-adverbs from the Finland English Corpus

	Text	Wh-adverb
1	...and some messages have not been sent by me. Apologies. If anyone has a clue how to dispose of this problem, I'm listening.	how
2	Having a "kill myself" -mood. Not yet, but someday. When I stop being a little pussy ugh.	when
3	@justinbieber How many eggs u got in Easter? ??	how
4	@justinbieber i just freezed when u came online and idk why	when
5	@justinbieber Why u call us Belieber? ????	why

laughter indicator *haha*, as well as lexical items such as profanity and politeness markers (Table 5.44). Females are more likely to use tokens from this class by a ratio of 1.47:1.[46]

TABLE 5.44: Examples of Interjections from the Finland English Corpus

	Text	Interjection
1	Vanilla yoghurt, bananas and grapes. OMFG	omfg
2	Ehdin ^^ (@ Alko w/ 4 others) http://t.co/ZkXVpygA1c	^^
3	@dephunkt aw thank you!!?? ;__; made my day!<3	;__;, <3
4	What a brilliant idea. Instead of the Oxford Cambridge boat race...An Oxford to Cambridge Goat Race!...might be illegal though:(	:(
5	@McWill99 @alan_elsworth hahah that' s perfect!	hahah

Females are much more likely to use non-3rd-person present singular verb forms than are males, by a ratio of 1.49:1.[47] Some examples from the data are shown in Table 5.45.

TABLE 5.45: Examples of non-3rd-person Singular Verb forms from the Finland English Corpus

	Text	Non-3rd-person singular verb form
1	#LittlemonstersMissGagaOnHerBirthday we all love you <3 xoxo	love
2	there are 3 cats but the one without a leg is my bff	are
3	@AustinMahone all you have to Do is take a picture of yourself and post it here ;)	have
4	@ItsAlyssaShouse I don't know but in the summer at the beach in sunset ooh!?	don't
5	@AustinMahone i just listened to it :D wanna follow me? :)	wanna

Patterns of use for pronouns are among the most "female" features in the data. Females use possessive pronouns at a rate 1.52 times that of males.[48] Examples from the data are shown in Table 5.46.

[46] Male–female difference highly significant (p–value $< 2.2e^{-16}$.
[47] p–value $< 2.2e^{-16}$.
[48] p–value $= 1.88e^{-11}$.

TABLE 5.46: Examples of Possessive Pronouns from the Finland English Corpus

	Text	Possessive pronoun
1	I just love my parents so much ?	my
2	there are 3 cats but the one without a leg is my bff	my
3	yeah the white/orange one is the best he's like my cat repe„ that's why he's the coolest	my
4	My birthday dinner with hubby :) (at Hard Rock) −http://t.co/bXZ3KIRKps	my
5	Having a movie night with my friends!? What about you? @AustinMahone and pleasepleasepleaseeew follow meee!!? http://t.co/24YFyo0LrF	my

The female-male ratio of utilizing tokens with the retweet tag is 1.55:1, although the difference doesn't reach statistical significance (p–value = 0.111).

Finally, personal pronouns are used by females more than by males by a ratio of 1.59:1.[49] Examples are shown in Table 5.47.

TABLE 5.47: Examples of Personal pronouns from the Finland English Corpus

	Text	Personal pronoun
1	@NiallOfficial u lost phone connection again or r u mad to us?	u, u, us
2	Fader! I need money, cus I'm really about to BUY AN APARTMENT.. In my mind	i, i'm
3	lol, we arent sorry to selena, we hate her. #BeliebersArentSorry	we, we
4	@DrakeBell yak, u r ugly	u
5	@DrakeBell everyone hate u drake	u

5.3.4 Selected Grammatical Word Class Frequencies

Primary grammatical word class frequencies for the Finland English and Comparison English data were determined by tags assigned by the CMU Twitter Tagger. Additional details, such as the frequencies of different subtypes of determiners, were calculated with processing scripts in *R*. The corresponding code can be found in the Appendix, Section A.

As discussed above in Section 5.3, the frequencies of grammatical word classes in the two main corpora, while exhibiting the same general distributional pattern, differ significantly for most individual grammatical word classes. Figure 5.13 shows the distribution of

[49] p–value $< 2.2e^{-16}$.

grammatical word classes in the two corpora per thousand words; this corresponds to the information presented in Table 5.17.

Most corpus–based approaches to the frequencies of word and grammatical types are based on corpora compiled from a mixture of spoken and print media. For example, Biber et al. (1999) provide a detailed account of the relative frequencies of different word classes in the genres of conversation, fiction, news, and academic writing, based on frequency statistics derived from the British National Corpus. Word class frequencies in CMC registers such as Twitter have been somewhat less well-studied. In the following section, selected grammatical word classes will be discussed in more detail in the context of their differing distributional profiles in the Finland English and Comparison English corpora.

Differences in the distributional profiles are provisionally interpreted as a consequence of the similar, but non–equivalent communicative orientations of Twitter English users in Finland and in locations that are not specified. However, linguistic interference phenomena may also play a role for persons using English as a second language. For example, Nation notes that linguistic interference, as reflected in the process of language acquisition, is most acute in situations when "the grammar and collocations of a word are not similar to those in the first language or to known second language synonyms" (2006: 449). Some frequency phenomena in the Finland English Corpus may reflect discongruence of the grammatical paradigms of Finnish and English and non–equivalence of their collocational patterning. The following short section provides a closer look at a select few word classes: the major word classes nouns, verbs, and pronouns, and the function word classes articles, demonstrative determiners, and numerals. Frequencies for the word classes are introduced for the Finland English and Comparison English corpora and compared with those found by Biber et al. (1999). Some L2 interference phenomena that may contribute towards a discrepancy in frequencies between the Finland English and Comparison English data are remarked upon.

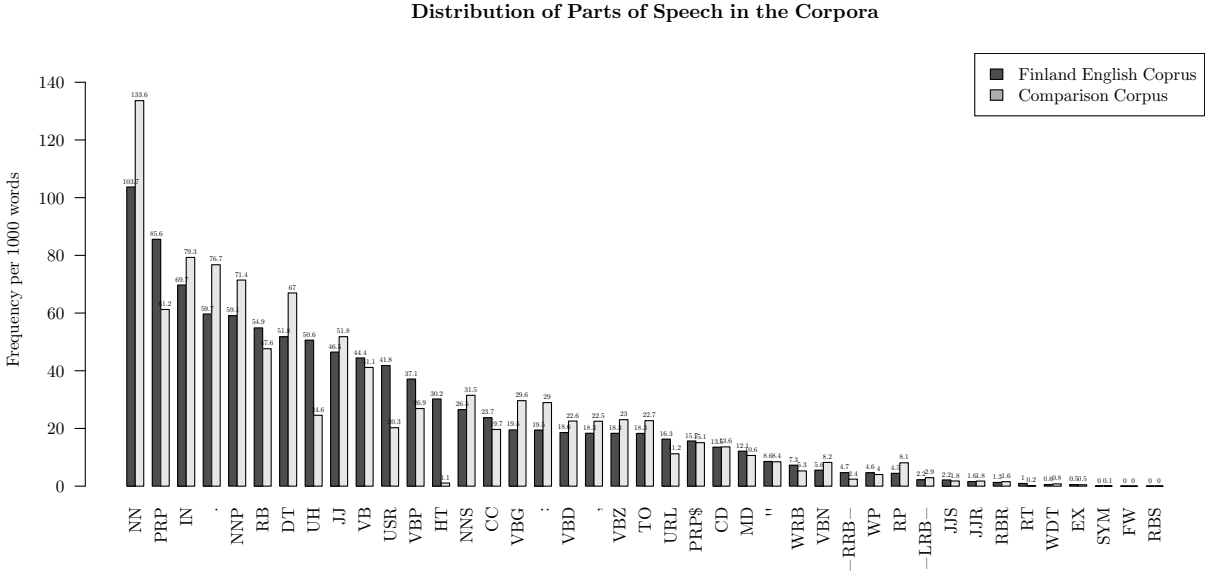

FIGURE 5.13: Frequency of Grammatical Word Classes

5.3.4.1 Main Word Classes (Nouns, Verbs, Pronouns)

Previous research has established the fact that register and genre effects strongly influence the relative frequencies of grammatical word types in language (Biber 1988, Biber 1995, Biber et al. 1999). More formal written registers such as academic journal articles or textbooks exhibit much higher rates of noun use than of verb or personal pronoun use, whereas spoken language typically exhibits approximately balanced frequencies for the three classes (Biber, Conrad and Cortes 2004: 378).

Figure 5.14 presents the frequencies for the major word classes for the Finland English and Comparison English corpora alongside the frequencies found by Biber, Conrad and Cortes (2004) for the spoken genres of conversation and classroom teaching and the written genres of academic textbooks and academic journal articles. It is apparent from the

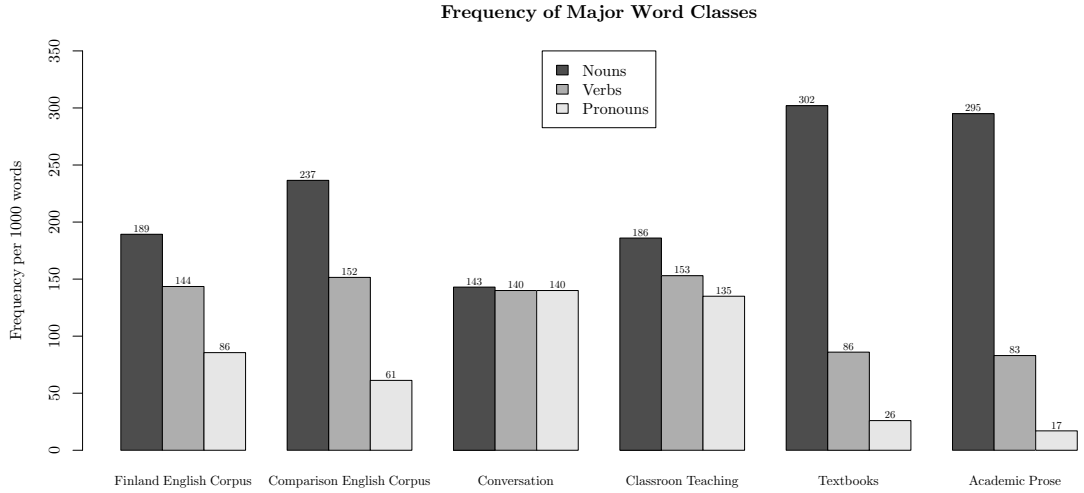

FIGURE 5.14: Frequency of Major Word Classes (Frequencies for Conversation, Classroom Teaching, Textbooks, and Academic Prose are from Biber, Conrad and Cortes 2004: 378)

figure that the Finland English and Comparison English data represent an intermediate position. The Twitter data show significantly higher rates of pronoun use than the written registers, although still much lower than those of the spoken registers. Nouns are used more frequently in the Twitter data than in the spoken registers, but not as frequently as in the two formal written registers. Verbs are used in the Twitter data at approximately the same rate as in the spoken registers and much more often than in the two written registers. The Finland English data exhibit fewer nouns and slightly fewer verbs than do the Comparison English data, but more pronouns.

5.3.4.2 Determiners

Determiners occur in the Finland English Corpus at a lower rate than in the Comparison English Corpus (Figure 5.15). The frequencies per thousand words, approximately 52 and 67 for the two data sets, fall between values calculated for the registers of "conversation" and

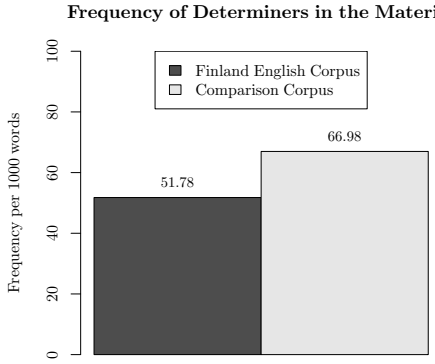

FIGURE 5.15: Frequency of Determiners

"fiction" by Biber et al. (1999: 273) from the British National Corpus, and are significantly lower than the rates for academic writing. Linguistic interference between Finnish and English may play a role in the lower frequencies for the Finnish data. For example, the grammatical category of definiteness is handled quite differently in Finnish compared to English, typically being marked in Finnish by declinational morphemes on the head noun of a noun phrase. Thus, some functional properties of the determination of the Finnish noun phrase can occur without the use of overt lexical determiners as a distinct word class.[50] Definiteness, for example, interacts closely with the grammatical category of verbal aspect in Finnish.

Verbal aspect in Finnish is either imperfective or perfective, depending on the morphological case of the verbal object. Partitive object case "imperfectivizes" Finnish verbs, whereas genitive implies perfectivity (Leiss 2007, Chesterman 1991, Larsson 1983). The close relationship in Finnish between definiteness and perfectivity, and between indefinite-

[50]It has been suggested that for Finnish, definiteness and the verbal phrase grammatical category of perfective/imperfective aspect are "two instantiations of the same grammatical function" (Leiss 2007: 73).

ness and imperfectivity, has been remarked upon from a comparative typological perspective (Wexler 1976).

5.3.4.2.1 Articles

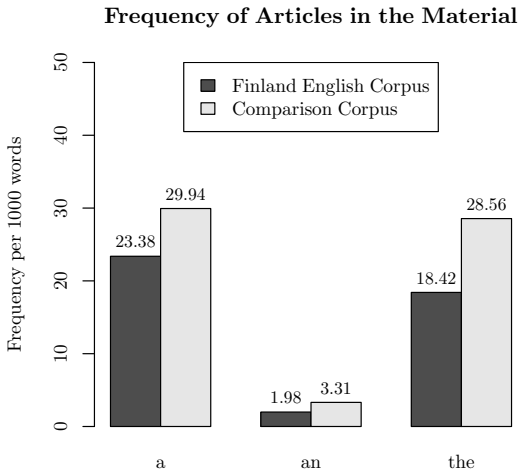

FIGURE 5.16: Frequency of Articles

There are no articles in Finnish. 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 (Figure 5.16). 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. 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.

5.3.4.2.2 Demonstrative Determiners

The Finnish English Corpus exhibits slightly higher use of demonstrative determiners overall. Biber, Conrad and Leech note that "in general, *this/these* are about twice as common as *that/those*" (1999: 274). This does not hold true for the Finland English and Comparison English data, as the ratio of proximal to distal determiners is 1.01 in the Finland English Corpus. For the Comparison English Corpus, the proximal/distal ratio is lower, at 0.82. In terms of characterizing text genre differences, higher proximal/distal demonstrative ratios are more typical of academic writing than other registers, but *that* is more common in conversation (Biber et al. 1999: 274). The Finland English and Comparison English data suggest that Twitter discourse exhibits a distributional pattern for the frequencies of proximal and distal determiners that is more typical of spoken conversation. The higher distal/proximal ratio for the Finnish users might be indicative of a slight preference among L1 Finnish users to use demonstratives (which have a corresponding word class in Finnish) when writing in English in phrases where L1 English users would set articles (which do not have a corresponding word class in Finnish).

5.3.4.2.3 Quantifiers

The quantifying determiners comprise a class consisting of words that provide quantification information to nouns or noun phrases. The frequencies of invariant single word quantifiers *all, both, another, every, many, some, any, either,* and *neither* in the Finland English and Comparison English data were considered.

The Finnish material has a slightly lower rate of quantifying determiner use, with the exception of the quantifier *every* (Figure 5.17). This may reflect the slightly different norms of use for the Finnish indefinite quantifiers *kaikki* ('every', 'all') and *joka* ('every'), both of which have a broader combinatorial range as constituents in Finnish than do the corresponding items *all* and *every* in English. A quantitative study of frequency distribu-

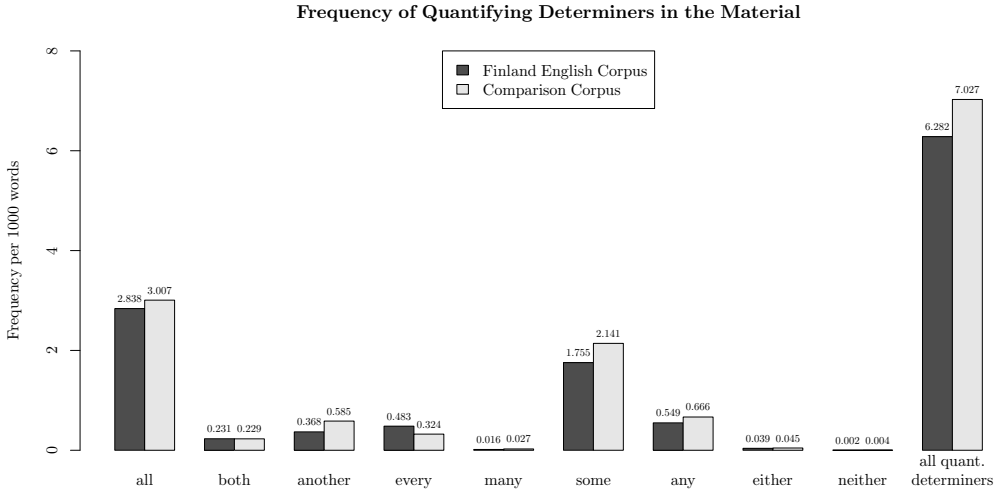

FIGURE 5.17: Frequency of Quantifiers

tions for Finnish quantifiers in online social media in the Finnish language would provide a useful point of comparison.

5.3.4.2.4 Numerals

In Section 5.2 it was found that numerical digits are used more often in the Comparison English data than in the Finland English data, and that numerals are used more often within the Finland English data by males than by females. A closer examination of some types of numerical expression follows.

Numerical Digit Graphemes

The rate of occurrence of cardinal numerical digit graphemes such as *0*, *1*, *2*, etc. is comparable in the Finland English and Comparison English data, at just over seven occurrences per thousand words (Figure 5.18).

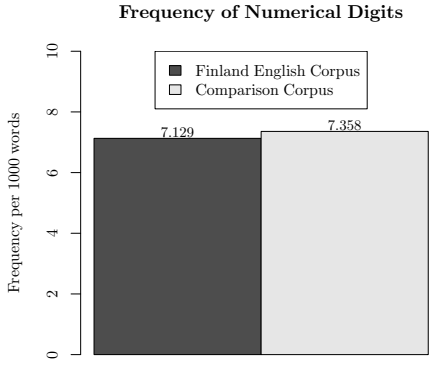

FIGURE 5.18: Frequency of Numerical Digits

The twenty most frequent cardinal numerical digits in the Finland English Corpus are shown in Figure 5.19 arranged in decreasing order of their frequency per thousand words, alongside the same types from the Comparison English Corpus. In general, the relative frequency of digits corresponds to the expected distribution of numerical digits in data as proposed by Benford's Law (see Hill 1998). The digit *1*, ranked fifth in the data, is more often expressed as a word type. Of the twenty most frequent types in the Finland English Corpus, the types *3, 1, 6, 7, 10, 8, 9, 30, 12, 20, 15, 18, 11*, and *0* are more frequent in the Finland English than the Comparison English Corpus, whereas the types *2, 4, 5, 100, 16*, and *14* are more common in the Comparison English Corpus.

Ordinal Digit Graphemes

Numerical digits with ordinal suffixes, such as *1st*, *2nd*, etc. are infrequent in both datasets, but somewhat less frequent in the Finnish data (Figure 5.20). This may also have to do with morphosyntax of numerals and numeral indicators in Finnish, where numerals are marked for ordinality, number and case (Karttunen 2006: 410), and a resulting disinclination on the

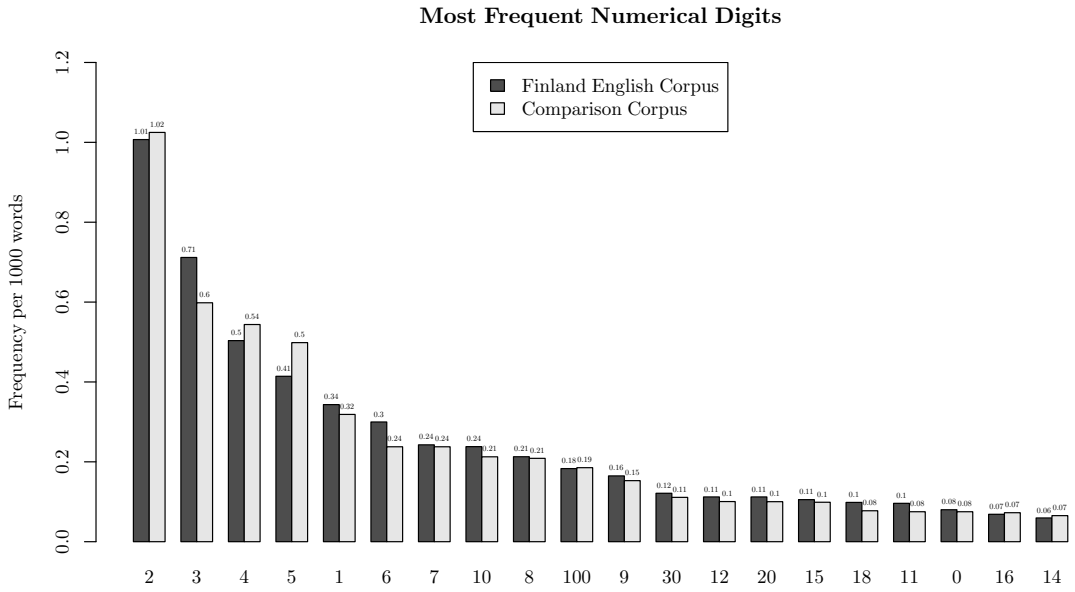

FIGURE 5.19: Most Frequent Numerical Digits

part of L1 Finnish speakers to utilize the equivalents in English, which are morphologically fixed only for ordinality.

Numeral Word Types

Overall, the Comparison English Corpus exhibits with a frequency of 2.90 per thousand words a slightly higher rate of use of cardinal numeral word types (such as *one*, *two*, *three*, etc.) than does the Finland English Corpus at 2.81 per thousand words. The twenty most frequent numeral word types in the Finland English Corpus are shown in Figure 5.21 with their frequency per thousand words alongside the frequencies of the same types in the Comparison English Corpus.[51] With the exception of *one*, these words are quite infrequent

[51] All types have a positive value; small values are rounded to zero in the figure.

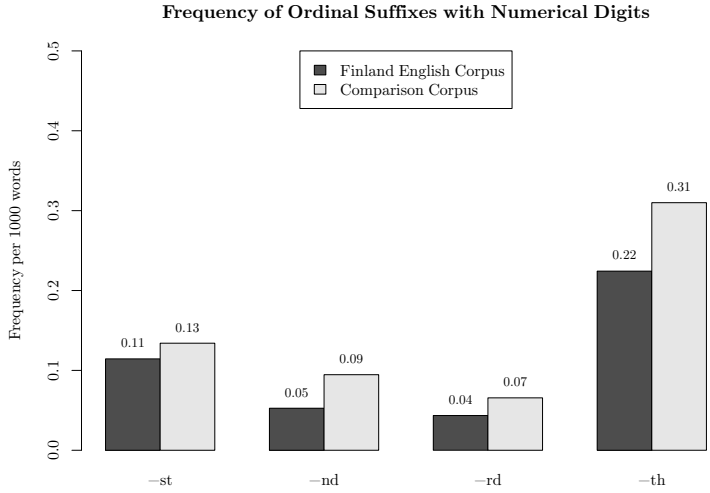

FIGURE 5.20: Frequency of Ordinal Suffixes with Numerical Digits

in both corpora. Once again, the rank distribution corresponds approximately to Benford's law, with the exception of the types *million*, *billion*, and *hundred*, which as single words may be more frequent as word types than other large numbers. The Finland English data shows higher frequencies of *one*, *ten*, *fifteen*, *fourteen*, *hundred*, *nine*, *seventeen*, and *thirteen*. All other types are more frequent in the Comparison English data.

Ordinal Words

Ordinals written as word forms, such as *first*, *second*, *third*, etc., are less frequent than the corresponding cardinal word forms *one*, *two*, *three*, etc., occurring at a rate of 1.12 per thousand word in the Finland English Corpus and 1.17 per thousand words in the Comparison English Corpus. The type frequency according to suffix is shown in Figure 5.22; *–st* corresponds to the type *first* (and hyphenated types such as *twenty–first*), *–nd* corresponds to *second* and related hyphenated types, *–rd* to *third* and related hyphenated types, and *–th*

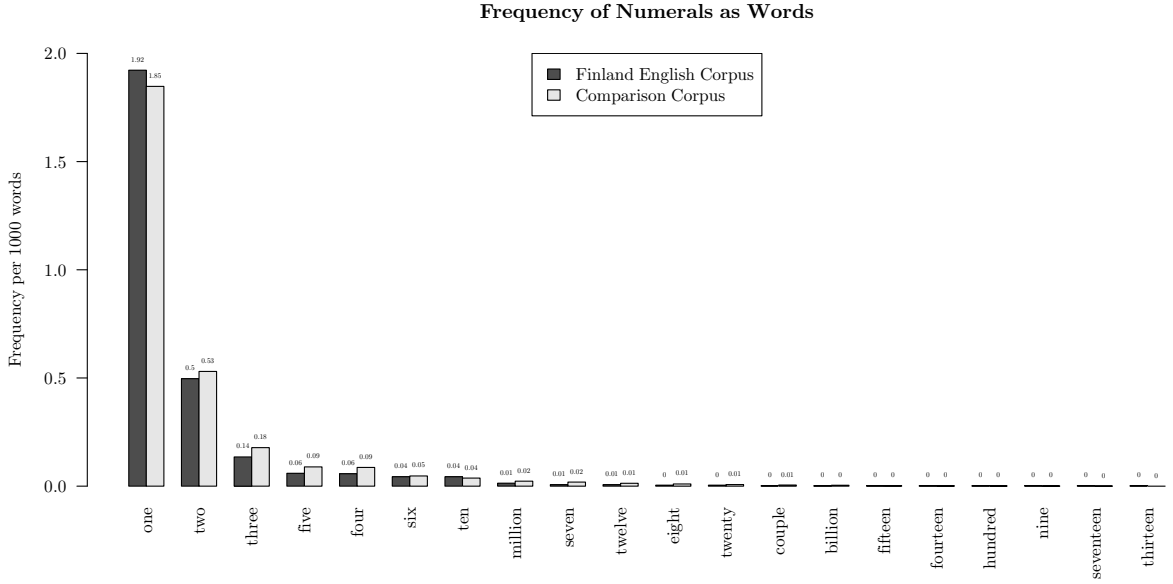

FIGURE 5.21: Frequency of Cardinal Numeral Words

to types such as *eleventh*, *twelfth*, *thirteenth*, *twentieth*, *hundredth*, *thousandth*, *millionth*, etc.

Variation in the use of number word types and numerical digits can reflect prescriptive norms that differ between languages, but also varies according to register and situational constraints. Biber et al. (1999) find that numerals are much more common in information–dense written registers such as news texts and academic writing than in conversation or in fiction. Ordinals are less common than cardinal numbers in general; of the cardinal numbers, digits are more common in news and academic prose, whereas word forms are more common in fiction writing.

The overall frequency of all numeral types, both cardinal and ordinal in digital, word, or combined representation, is 13.52 per thousand words in the Finland English Corpus and 13.62 per thousand words in the Comparison English Corpus. These rates are higher than

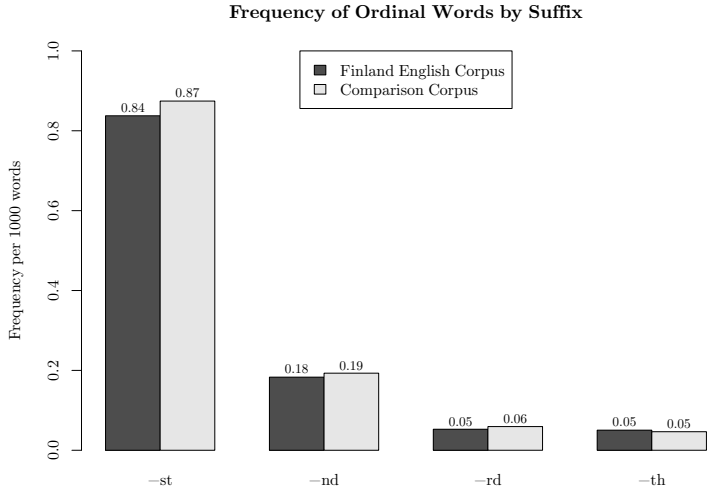

FIGURE 5.22: Frequency of Ordinal Words by Suffix

the rates reported by Biber et al. for all numerals in fiction (approximately 8 per thousand words), but not as high as the rate of numerals in spoken conversation (approximately 17 per thousand words; p. 279). News and academic texts exhibit rates of numeral use of approximately 25 and 23 per thousand words.

Biber et al. note that conversation shows a very high ratio of numerals and other quantifiers to nouns; although registers such as news reports or academic writing show overall higher rates of numeral and quantifier use per thousand words, they typically also feature far higher noun frequencies. Thus one cardinal form occurs per every seven nouns in conversation, per every approximately 15 nouns in news texts and academic writing, and per every 26 nouns in fiction (Biber et al. 1999: 235f., 279f.). In the Finland English and Comparison English data, the cardinal form to noun ratios are one cardinal form per 19.1 nouns and one cardinal form per 23 nouns, respectively.

It is suggested that an underlying communicative difference may be in play: Whereas in conversation noun phrases are most likely to be specified by quantitifiers, in information-

rich registers such as many written genres, noun phrases are more often specified by the definite article or deixis words that situate content within the discourse (such as relativizers). In short, marking for identity is more typical of information-dense written language; marking for quantity more typical of conversation.

The findings from the Finland English and Comparison English corpora suggest that the register of English–language Twitter communication again takes an intermediate position between canonical written genres such as news texts or academic writing and spoken conversation. The immediate, conversational nature of discourse for many users of Twitter is manifest in a relatively high ratio of nouns to numerals compared to written registers such as news texts or academic writing. The Finland English data show a lower noun–numeral ratio than the Comparison English data in part due to the lower overall frequency of nouns in the Finnish data. The ratio also reflects the more conversational nature and immediate concerns (i.e. the interactivity) of the discourse of Finland Twitter English compared to Twitter English without geographical specification.

5.4 Tweet Length, Lexical and Grammatical Features: Discussion

The findings from the analysis of selected lexical and grammatical features as they are manifest in the two principal corpora as well as in the gendered section of the Finland English Corpus help to situate Twitter English within the broader context of the parameters of spoken and written communication. More specifically, they allow a preliminary assessment of the extent to which Twitter English in Finland differs from global Social Media English or Twitter English and how an emergent Finland Social Media English variety could be characterized.

5.4.1 Discussion of Tweet Length, Lexical and Grammatical Features and Geography

The analyses undertaken in sections 5.1, 5.2, and 5.3 demonstrate some of the most salient differences between the variety of English used in Finland on Twitter and the corresponding non–geograpically–situated variety of Twitter English. The principal differences and their implications for the characterization of Finland Twitter English are summarized below.

Tweet Length

Corpus–based 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 (Xu, Ritter and Grishman 2013, Walkowska 2009: 149). 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).

Finnish user 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 ($\geq$ 6 characters)

words, as shown in section 5.1. Zipf (1935, 1949) 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). Yule (1939) proposed that sentence length in writing is a characteristic feature that could be used for e.g. authorship disambiguation by means of significance testing.

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. They find 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.

Piantadosi, Tily and Gibson (2011) discuss Zipf's observation that word length is inversely related to word frequency and the concommitant interpretation that infrequent words communicate more information content. They note that for a given set of contexts, the information provided by a word in a text can be quantified by the logarithm of the inverse mean frequency of the word. Contexts can be modeled by using n–grams.[52] Formalizing the model by quantifying informational content for words in bi–, tri– and quadgrams, they find that while information content is correlated to both word length and inverse word frequency, information content is a better predictor of word length than is word frequency.

Alis and Lim (2013), in a study of geo–tagged tweets from the United States, consider the relationship between length and some demographic variables. They find the strongest correlation to be the inverse relationship between Tweet length in characters and percent

[52]This is the entropy of a probability distribution according to the information model proposed by Shannon (1948); it is related to Mutual Information and similar collocation association measures. Piantadosi, Tily and Gibson note, however, that there are "many variables that may count as part of the 'context' for the purposes of language processing", including discourse context, local linguistic context, global world knowledge, and pragmatic considerations (p. 3526); such variables are much more difficult to adequately operationalize.

African–American population of a geographical region. They interpret the findings as evidence for a distinct vernacular language variety anchored in racial or ethnic identity.

The tweet length findings from the Finland and Comparison corpora can be interpreted both in the context of the communicative efficiency observations of Zipf as indications of the functional–pragmatic dynamics of language use online and in the context of sociolinguistic considerations pertaining to demographic traits of the users whose messages constitute the Finland English Corpus. Shorter tweets and shorter words generally contain less information content than do longer tweets and longer words. In general, Twitter discourse is less information oriented than the discourse of traditional text types such as news reports, academic writing, or fiction; it is more interactive. The shortness of tweets reflects the communicative functions most typical of Twitter, such as self–representation, often in abbreviated form, and interactivity. In this respect, the Finland English Tweets are even less informational and more interactive than the Comparison 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.[53] In a sense, Finland Twitter English is more "Twitter" than is non–Finland Twitter English.

Most Frequent Lexical Types

Section 5.2.1 shows that the most frequent lexical items in the two corpora are short grammatical function words such as prepositions, adverbs, and articles, as well as personal pronouns. Short function word types are more frequent in natural language and more uniformly distributed in texts than are content words; for this reason function words have been considered to be better suited for disambiguation tasks than content words (Mosteller and Wallace 1964, Damerau 1975). The higher frequency of function words in language

[53]This is expanded upon below in Chapter 6.

is reflected in their prominence among the most frequent lexical items in large, balanced corpora such as *COCA* or the *BNC*. Compared to *COCA* or the *BNC*, the two Twitter corpora contain relatively more pronouns and relatively fewer articles. This finding is in line with the results of previous research into lexical frequencies in CMC and can be best explained by the fact that many types of CMC, in addition to being situated along different communicative parameters than texts represented in large, balanced corpora, are often created in order to achieve interactive, rather than informational functionality. The "interactive" nature of Twitter English is further determined by the character limit constraint; Finland Twitter English is, based on the frequency of lexical types, more interactive and less informational than non–Finland Twitter English. For the most part, the content words that are most characteristic of either the Finland English Corpus or the Comparison English Corpus reflect chronological and geographical factors. The nature of of the Twitter platform as a vehicle for the propagation of relatively ephemeral content such as references to breaking news (Yang and Leskovec 2011) means that for some lexical items, large differences in frequencies between corpora may simply reflect the times at which the data was collected. Other lexical items are associated with geographical location: It is unsurprising that types such as *Finland, Helsinki*, or *Sweden* would be more common in a Finnish context. The high relative frequency of names or cultural entities such as *Obama* or *Steelers* may simply reflect the fact that much of the Comparison English Corpus data consists of American tweets. Non–standard word forms such as initialisms, emoticons, or symbols are among the most "Finnish" types, and are relatively more common than in the Comparison English data. In terms of identifying the styles of interactivity which characterize Finland Twitter English, this allows it to be situated in a larger context as a subtype of an emerging English variety.

Profanity and Taboo Words

The results from the short section pertaining to the use of profanity or otherwise taboo language (Section 5.2.4) show that Finland-based users of Twitter writing in English use words from this class at much higher rates than do non-Finland Twitter users writing in English. This could result from lower levels of inhibition amongst Finland English Twitter users for the use of English taboo language and profanity: Some studies have found that less "emotional force" is associated with taboo language by L2 speakers compared to L1 speakers (Dewaele 2004, Eilola and Havelka 2010). Taboo words and profanity, typically used as intensifiers or indexicals referencing emotional and affective states, are particularly useful for the highly interactive and relatively affect–intensive communication that is characteristic of Finland Twitter English (see the additional analyses in Chapter 6). A relative lack of usage constraints as well as the suitability of words from this class for the situational, discursive and communicational contexts characteristic for the medium may explain the high rates of use in the Finland English Corpus.

Emoticons

Emoticons have not figured as prominently in corpus–based studies of language as have other types of strings such as dictionary words. 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 text types such as blogs and online equivalents of more traditional text types like news reports or academic writing (Ptaszynski et al. 2011). Some early studies interpreted the symbols as affective indicators meant to convey contextual information that corresponds to spoken–language cues such as prosodic, stress, and intonation features (Herring 1999, 2013).

Ptaszynski (2007) compiled a corpus from the Japanese site 2ch.net to determine the frequencies of emoticon types and administered a survey in which participants were asked what types of linguistic resources they used to convey emotional states: lexical expressions, emoticons, character repitition (expressive lengthening), ASCII art, non–standard orthography, or other.[54] Corpus data and survey results confirm the importance of emoticons as a language resource in the community.

Schnoebelen (2012) discusses the interactive nature of emoticon use in an American English corpus of Twitter messages and considers the correlation between particular emoticon facial representations and lexical features. Although emoticons, as graphical representations of facial expressions, could be considered to correspond directly to user emotional states, Schnoebelen suggests that the way they are used in texts and patterned with other lexical items show that they are "interactive in nature, positioning audiences around propositions" (118).

The interpretation of emoticons as a linguistic resource whose meaning is contextualized by discourse considerations is shared by Vandergriff (2014). Analyzing emoticon use in a corpus of English–language IM data from American and Swedish university students, she finds that Swedish English users exhibit a rate of emoticon use almost double that of their American peers and suggests that emoticons can be used not only as affective indicators, but as contextualization cues and compensatory gestures for non–native–speaker competence.

The idea that emoticons are used for multiple communicative functions is somewhat similar to pragmatic interpretations of hashtag functionality by Zappavinga (2011) or Wikström (2014). According to these analyses, the dynamics of online communication on

[54]http://2chan.net is an anonymous discussion forum in which users can upload texts and image files; it was the inspiration for the popular American imageboard 4chan.net and many similarly formatted sites in other national contexts.

Twitter in terms of the relationship between specific character strings such as emoticons and communicative functions is not fixed and continues to evolve.

In the current study, the relative proportions of all emoticon use comprised by distinct emoticon types are similar in the Finland English and the Comparison English corpora and comparable to those reported by Schnoebelen 2012 (Table 5.12, Figure 5.5), suggesting that whatever the communicative or discourse functions of emoticon use may be, their type distributions are somewhat stable across cultural boundaries in English–language Twitter used from late 2008–2013.

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.[55] The analysis of the geographical patterning of emoticon use in Section 5.2.5.0.2 suggests that users in regions of Finland with higher per capita GDP are more likely to use emoticons than are users in poorer regions. If we further take into account the fact that all of the Finland English Corpus data is geo–encoded, and geo–encoded tweets are typically published using smart phones, it seems that the tendency to use emoticons in English is related to socio–economic factors such as income. The results of a correlation test with demographic data (Table 5.16) show a correlation between Russian–language tweet frequency and use of emoticons. However, in line with our interpretation of emoticon use as indexing primarily a socio–economic independent variable, this can be explained by the fact that Russian–language Twitter users in Finland may be

[55]Finland English: 23.87 per thousand tokens or 31.69 per 100 tweets; Comparison English: 6.79 per thousand tokens or 10.69 per hundred tweets. Vandergriff (2014) reports 2.18 emoticons per 100 "transmissions" for native speakers and 4.01 per 100 for non–native speakers. Genre considerations undoubtedly play a role: Vandergriff's data was compiled in the context of a collaborative writing task for American and Swedish university students, a type of writing that may be more formal than spontaneous IM communication or Twitter user messages.

wealthy Russian citizens of St. Petersburg who own property in Finland or frequently visit Finland for business, shopping, or leisure.[56]

Another consideration is the extent to which adoption of online services such as Twitter correlates with income. Particularly for geo–tagged tweets, the majority of which originate from smartphones, Twitter use may reflect the extent to which a population owns and uses smartphones. This statistic correlates quite strongly with measures of wealth such as per capita income. Mocanu et al. (2013) find that at least for those tweets that are geo–tagged, the wealthy country of Kuwait has the world's the highest per capita Twitter use.

Our data shows a strong correlation between GDP per capita and tweets per capita for the Finnish regions (Pearson's $\rho = 0.8593105$, Wilcoxon-Mann ranked sum test W = 0, p–value = $3.062e^{-08}$). The wealthiest area of the country, Uusimaa, shows far greater Twitter saturation than do relatively poor areas such as Karelia, Kainuu, Savo, or Lapland.

If emoticons are considered to be a linguistic resource with a range of communicative functions, and geo–encoded tweeting is a linguistic behavior that is associated most strongly with socio–economic considerations such as access to new technology, it may be the case that emoticon use in English messages in Finland corresponds to an evolving Twitter–based communicative functionality which emerges at the interface between language use and technological change. In the case of Twitter in Finland, this may reflect economic considerations.

Orthography and Expressive Lengthening

In the linguistics literature, orthography has traditionally been considered from the perspective of the correspondence between characters and speech sounds (Pike 1947). Studies of non–standard orthography, for the most part, have been limited to the investigation

[56] As remarked upon above, this interpretation seems reasonable for early 2013 when the Finland Twitter data was collected. As of early 2015, far fewer Russians visit Finland due to the economic and political consequences of the Ukraine crisis and related EU–Russia diplomatic friction.

of error occurrence rates, for example in corpora of student writing or in texts created by experimental subjects under specific conditions (Hotopf 1980, Mitton 1987). In recent years, orthography has received some attention in ethnography– and social anthropology–oriented sociolinguistics literature. Sebba (2007) analyzes orthographical standards, spelling reforms, and non–standard orthography in several languages, suggesting that there is a considerable range of orthographical variation in most written language and that orthography may express a range of individual, group, societal, and national orientations (5). In some cases, non–standard orthography may not represent faulty adherence to a standard, but rather a language choice designed to create "social meaning" for a language user (160). Althoug Sebba does not specifically analyze CMC, some analyses have suggested that non–standard orthography is a discourse feature characteristic of online text. Herring, considering the extent to which features such as non–standard grammar or orthography in online language reflect specific communicative or discourse maintenance concerns rather than failed attempts to produce standard forms, notes that "only a relatively small percentage of such features appears to be errors caused by inattention or lack of knowledge of the standard language forms" (2004:5). The analysis, however, was based on group e–mail data produced mainly by American academics or professionals, who would presumably have relatively good knowledge of standard English orthography (Herring 1998b). Other CMC genres such as chat and IM often exhibit high rates of use of non–standard orthography (Paolillo 2001, Tagliamonte and Denis 2008). The widespread orthographical variation in genres such as Twitter may represent an example of individuals and groups utilizing non–standard language variants to create social meaning.

Non–standard orthography is a frequent feature of the types in both the Finland English and Comparison English corpora. Distinguishing between non–standard orthography which represents lacunae in user knowledge or errors in user input, and non–standard orthography which represents stylistic choice, is a non–trivial endeavor. Ling (2005), con-

sidering the factors that may contribute to the use of non–standard orthography in SMS messages, remarks:

> What factors might promote the use of nonstandard spelling, capitalization, or punctuation? It could be any of several reasons. The writers might be ignorant of the standard usage. They might know it, but not be bothered to use it. They might be bothered, but don't have keyboard skills up to the task of typing it correctly. They might think they've typed it correctly, when actually they haven't, and failed to read their message through before sending it. They might make a conscious decision not to bother with the standard form, because they feel it is unimportant. They might, consciously or unconsciously, use the nonstandard form in order to accommodate to the usage of their peers. They might deliberately use it to create a special effect. Or some combination of these factors might apply. (qtd. in Crystal 2011: 61)

NLP research into spelling errors, however, has established that 80% of errors consist of strings with an edit distance of one from the intended string, using a Damerau–Levenshtein distance measure, and almost all spelling errors are within two edits of an intended lexical item (Mays, Damerau and Mercer 1991).

The analysis in Section 5.3.1 demonstrates that the Finland English Corpus exhibits rates of 1–edit–from–standard non–standard orthographical forms that are significantly higher than those of the Comparison English Corpus. Many of these types may represent input errors or lack of knowledge of standard forms, but it is impossible to determine the extent to which these forms are errors or instances of stylistic variation. For example, among the most frequent of these types, several (the types *cant, didnt, doesnt, isnt, wasnt*, and *todays*) omit an apostrophe. These types are one edit from the standard norms, and thus may represent "errors". On the other hand, Squires (2012) treats the apostrophe as a

sociolinguistic variable in an IM corpus, and finds that its use or omission correlates with social factors such as subject gender and gender of interlocutor.

Although it may not be possible to categorically distinguish between orthographical errors and non–standard orthography as a stylistic feature, some instances of non–standard orthography almost certainly represent user intentionality. Expressive lengthening (such as *yessss or looooool* is the repetition of individual characters in word strings. These are by definition 2 or more edits from the equivalent standard lexical item, and therefore unlikely to represent "errors".

Expressive lengthening in the Finland English Corpus is much more extensive than in the Comparison English Corpus. 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 whom phonological considerations of L1 Finnish influence the ways in which they combine particular phonemes or letter characters. An exploration of comparative letter character frequencies in Finnish and in English and their relationship to expressive lengthening in CMC would represent an interesting possibility for future research.

The interpretation of expressive lengthening as a stylistic resource, most often used to indicate or intensify expression of emotional affect, is reinforced by the research of Fry (1955) and Klett (1976), who show that longer vowel duration in spoken language is perceived as more emotional. The prevalence of this feature in the Finland English Corpus, according to this interpretation of the communicative function of the feature, demonstrates the primarily interactive nature of Finland Twitter English, a variety in which expression of affective stance plays a crucial role.

Grammatical Part-of-Speech Frequencies

An examination of the most frequent lexical items in the Finland English and Comparison English data sheds light on how Twitter English can be situated in the context of spoken and written English genres: It uses pronouns more frequently than typical written genres (i.e. it is more like speech), nouns somewhat less frequently than written genres (much less frequently, in the case of the Finland English Corpus), and articles less frequently than written genres. Twitter English, in terms of its communicative profile, is somewhat more like speech than it is like other written text genres; Finland Twitter English is more speech–like than is Comparison Twitter English, with which it shares a primarily interactive communicative orientation.

The comparison of the grammatical part–of–speech frequencies in the Finland English and Comparison English data (summarized in Table 5.17) provides further insight into the specific nature of an online English variety as it develops in the context of the communicative culture of Northern Europe. Hashtags and retweets are the two features that are most overrepresented in the Finland English Corpus, compared to the Comparison English Corpus. The four most "Finnish" features are, interestingly, three features which are unique to the Twitter language ecosystem (the hashtag, the retweet, and the username), and one feature, the interjection, which is strongly associated with informal types of online language: In Twitter data, interjections are most commonly emoticons or non–dictionary–word expressions of affective or emotional content.

In the case of hashtags, this may simply be an artifact of the time at which the two corpora were compiled, as noted above: Hashtags were not yet as widely used on Twitter in 2008–2009 compared to 2013. 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. As Wikström (2014) notes, hashtags may represent an example

of ways in which the functional roles 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. Although some commentators, such as Squires (2010), have cast a critical light on what they perceive to be "technological determinsm" in the interpretation of the development of CMC and online language, a technological moment in the evolving norms of language use on Twitter as they are manifest in frequency data is difficult to deny. Younger, more interaction–oriented, better educated and wealthier users are the most likely to use those features of language which are the most technologically determined: hashtags, retweets, usernames, and interjections, most of which are the symbolic affective resources known as emoticons.

Selected Grammatical Word Class Frequencies

Additional contours of the usage norms of this emerging interactive English variety are adduced by closely examining the frequencies of some specific word classes. As noted above, and confirming earlier research, the Finland English and Comparison English corpora take an intermediate position between spoken language and traditional written genres in terms of the frequencies of nouns, verbs, and pronouns. Finland Twitter English uses relatively few determiners such as articles compared to non-geographically-restricted Twitter English; this may reflect in part L1 interference effects with L2 English. Quantification in Finland Twitter English is similar to quantification in terms of the frequency distribution of numerical digits, but Finland Twitter English users are less likely to use quantifying determiners, cardinal number words, ordinal number words, or ordinals presented as combinations of digits and letter character suffixes. Overall, this may reflect a more interactive communicative orientation of Finland Twitter English compared to non-geographically-restricted Twitter English.

5.4.2 Discussion of Tweet Length, Lexical and Grammatical Features and Gender (Finland English Corpus)

The investigation of the correlation between gender and particular linguistic features has long been of sociolinguistic interest. Early sociolinguistic studies in the variationist paradigm empirically documented different rates of use for phonological features by males and females and typically attributed these differences to slightly different male and female patterns of orientation towards perceived linguistic norms. Studies such as Labov (1972) or Trudgill (1972, 1974) investigate patterns of phonological variation and gender and show that males are more likely to use non–standard phonological variants. Labov (1990) considers gender as a factor in a proposed ongoing reorganization of the vowel space in the Midwestern cities in the United States. Other sociolinguistic studies have investigated the interplay of lexical and grammatical variables and gender and typically attest a greater orientation of females towards standard norms. The articles in Coates (1998) provide an overview of much of the sociolinguistic research that deals with gender.

Alternative interpretations of perceived gendered language differences have been suggested, for example by Lakoff (1973) in a much-cited study. Lakoff asserts that women use certain word forms, such as diminutives or politeness markers, or syntactical constructions, such as hedges and tag questions, more frequently than men because they would otherwise be "ostracized, scolded, or made fun of"; gendered differences in feature frequencies thus reflect power differentials (47). However, Lakoff, who considered her research to be important for "those working in the women's liberation movement and other kinds of social reform", notes that "the data on which I am basing my claims have been gathered mainly by introspection: I have examined my own speech and that of my acquaintances, and have used my own intuitions in analyzing it" (45, 46). Despite her defense of introspection and intuition as valid empirical methods ("any procedure is at some point introspective:

the gatherer must analyze his data, after all"; 46), the conclusions do not present solid evidence for gendered difference in the frequencies of use of particular word classes or syntactic constructions. The corroboration (or refutation) of Lakoff's intuitions would be left to analysts conducting actual linguistic fieldwork or data analysis.

In the variationist paradigm, differences in gendered usage for some language variables have been interpreted inconsistently. Gender differences in frequency of language variables may reflect oppression of women, as suggested by Lakoff, or may reflect different strategies of discourse negotiation in direct interaction by males and females, corresponding to differential use of communicative functions that relate to categories such as orientation towards affective maintenance or solidarity (Holmes 1998). Explanatory factors such as "covert" and "overt" prestige have been proposed to suggest that males may be more oriented towards locally–based social networks, whereas females orient towards perceived mainstream norms (Labov 2001). Proposed gender-based sociolinguistic "universals", however, have been controversial, with some asserting that the category of gender is not a binary categorical universal (Cheshire 2002).

While the parameters of direct, face–to–face interaction may interact with gender to produce differences such as differential attention towards prestige– or non–prestige phonological realizations of some phonemes, it has been thought that writing, with its typically impersonal situational parameters, would exhibit no differences in gendered usage (Argamon et al. 2003: 323). Variationist studies often required time-intensive data collection and transcription of interviews for relatively small data sets. Corpus-based methods have allowed much more extensive quantitative analysis of gender differences in both spoken and written language.

Biber, Conrad and Reppen (1998) find that females use more modal verb forms than do males. Mehl and Pennebaker (2003), analyzing transcriptions of spoken conversations, find that females use first-person pronouns more frequently than do males. Argamon et al.

(2003) analyze the British National Corpus and find that even impersonal informational texts exhibit different rates of usage of some word classes by male and female authors. They interpret their results as broadly reflecting communicative dimensions of involvement and informational orientation suggested by Biber (1995 and 1998) and find that it is possible to determine the gender of an author from their data with 80% accuracy based on frequencies of word classes.

In a study of Instant Messager (IM) communication among students at an American university, Baron (2004) finds that in an 11,000 word corpus, females use fewer "CMC contractions" (abbreviations and non–pronounceable initialisms or acronyms such as *brb, btw*, or *g2g*), but more emoticons.

Herring and Paolillo (2006), in a study of the language of online blogs, find that determiners, demonstratives, and numbers are more used by males, whereas personal pronouns are more used by females.

A different analysis of a large number of online blogs using multidimensional techniques shows that males are more likely to use content–related grammatical variables (Argamon et al. 2007). Drawing on the communicative dimensions proposed by Biber, Argamon et al. find that females utilize linguistic resources that modulate interactions between speakers and audiences more often than males, who utilize features and linguistic resources that communicate propositional content more often.

Newman, Groom, Handelman and Pennebaker (2008), drawing from a corpus consisting of written texts and transcribed conversations, find that among other differences, males use more articles, numbers, longer words ($\geq$6 characters) and taboo words than do females, whereas females use more personal pronouns, emotion words, and intensive adverbs such as *very, strongly* or *really.*

Rao et al. (2010) analyze a Twitter dataset and find that women use emoticons, some types of punctuation (ellipses and multiple punctuation marks such as !! or ??), and

expressive lengthening more than do males. Burger et al. (2011) use machine-learning algorithms based on n-gram feature frequencies to predict the gender of blog authors (on a blogging platform where gender must be explicitly provided) linked in Twitter user profiles where gender is not explicitly provided. The most informative features, as determined by their algorithm, include content lexemes such as forms of the verb *to love* and emoticons. The authors find that their algorithm predicts user gender with better accuracy than do human subjects. Bamann, Eisenstein and Schnoebelen (2014) find that in a large corpus of United States Twitter messages, males tend to use, for example, numbers, "technology words" such as *api*, *ios*, *portal*, or *plugin*, sports-related terms, and profanity more frequently than females, who use pronouns, emotion words, CMC words such as *lol*, *omg* or *lmao* and emoticons more than males. They train a machine-learning algorithm to correctly identify gender from Twitter messages with up to 88% accuracy. Although the linguistic data show strong correlations between a number of linguistic features and gender/sex, to an extent that it is possible to distinguish gender based on written content with a high degree of discrimination, they draw upon ideas of social theorists such as Judith Butler and assert that gender is an interactive identity category that is "performed and constructed"; thus the "social meaning" of linguistic variation "constructs" gender as a "stylized repetition of acts" (138).

Females in the Finland English Corpus data are more likely to use non–standard orthographic forms from the set of 4269 Wikimedia misspellings. Among the most frequent types from this set, several (the types *cant, didnt, doesnt, isnt, wasnt, todays*, and *shouldnt*) omit an apostrophe. Squires (2012), investigating apostrophe use and gender in a corpus of IM, finds that males are less likely to use apostrophes than females. This tendency also holds true for the gendered portion of the Finland English Corpus, with females 1.37 times more likely to use apostrophes than males, despite having an overall higher likelihood for

the use of non–standard forms.[57] However, males seem more likely to use other types of punctuation symbols (Table 5.35).

Females are also much more likely to use language resources associated with the expression and situative orientation of emotional, interactive content, such as expressive lengthening and emoticons. Although a clear monotonic relationship between orthography and gender does not seem to explain the observed patterns of gendered variation for non–standard orthography, a preliminary finding of this study is that for Finland Twitter English, female discourse is situated closer towards the interactivity pole of an information–interactiveness axis, whereas male discourse is situated slightly nearer to the informational pole.

Males in the gendered section of the Finland English Corpus seem more likely to utilize grammatical part–of–speech categories that may have to do with the expression of proposition– and information–oriented discourse rather than interactive discourse. Features such as existential *there*, proper nouns, universal resource locators, hashtags, nouns, and numbers are more likely to be associated with informational content than with interactivity. Prepositions and other grammatical resources pertaining to physical location are also more "male" than female in the Finland English Corpus.

The female features in the gendered portion of the Finland English Corpus are highly interactive: Personal pronouns, retweets, possessive pronouns, and first– or second–person present verb forms are the most female forms, followed by interjections, Wh–words, usernames, and modals. These grammatical categories are used to engage in dialogue and mark personal expressions of orientation or stance.

It is interesting to note that collectively, the three features that are the most "Finnish" comprise one feature that is extremely "male" (hashtag use) and two features that are

[57]The relationship between apostrophe use and gender is significant according to a Chi-square goodness-of-fit test, $\chi^2 = 43.16, p - value = 5.04e^{-11}$.

extremely “female” (username and retweet use) in the gendered portion of the Finland English Corpus. It may demonstrate that although there are gendered differences in the ways in which linguistic resources are used online in Finland Twitter English, persons tweeting in English from Finland in both or any gender are quick to adopt those features which are most emblematic of the communication platform.

Chapter 6

Multi-Dimensional Analysis

In this chapter, quantitative methods of dimensionality reduction are used to reveal patterns of co-occurrence in the grammatical features of the data. A brief discussion of the methods and history of multidimensional analysis (pioneered in language studies primarily by Douglas Biber) is followed by a principal component analysis and factor analysis of the Finland English data and the Comparison English data. An interpretation of some of the communicative functions of grammatical features as they are used in Twitter, and how these differ for the two geographical groups that constitute the main interest of the study, follows.

"Multidimensional Analysis" refers to a quantitative corpus-based methodology developed by Douglas Biber and described in a series of articles and monographs from the 1980s and 1990s (Biber 1985, Biber 1986, Biber 1987, Biber 1988, Biber 1995), in which the functional discourse properties of different registers and genres are distinguished based on rates of co-occurrence of multiple linguistic features. Multivariate statistical techniques such as factor analysis are then used to identify the most salient patterns of co-occurrence. Multidimensional analyses have contributed to the study of register and genre by identifying many ways in which, for example, spoken and written language represent distinct varieties

or overlapping entities, and the quantitative procedures developed by Biber (such as the application of factor analysis, principal component analysis, or other dimensionality reduction statistical methods in linguistics) have been further developed and refined in work on, for example, quantitative stylistics and authorship disambiguation (such as those of Binongo and Smith 1999; Burrows 2001, 2002a, 2002b, 2003; or Hoover 2001, 2002, 2003a, 2003b, 2004). Increasingly, statistical programmers are developing code that automates some of the processing steps necessary to undertake such analyses (e.g. Eder, Kestemont, and Rybicki 2013).

In a historical perspective, Biber's original work from the 1980s represents not only an important contribution to the comparative study of register, genre and style, but also a significant methodological innovation in the application of computational and NLP procedures and the use of multivariate statistical methods in linguistics.

What is a "dimension", in a multi-dimensional analysis, and can the concept be applied to disambiguate Twitter datasets? In order to address this question, we will briefly describe the procedures and underlying theoretical justification for the methods developed in Biber 1988; Biber 1995 represents an application of the methodology to non-English language data.

First, in order to develop the key notion of "textual dimension", Biber considers the various configurations of functional and situational parameters that have traditionally been used to frame and study the resulting linguistic forms that are produced in speech and written texts. These are functional parameters (similar to those described above in Section 2.2) such as time/space orientation, participant attitudes, interactivity of communication, formality of situation, type of code employed, goal of communication, knowledge of interlocutors, shared knowledge by communicative participants, and others. Biber notes that traditionally, "researchers have given priority to functional dimensions such as formal/informal, restricted/elaborated, or involved/detached, and subsequently they have identified

the linguistic features associated with each dimension" (1988: 13). For example, text genres that may feature more interactivity (e.g. conversations) may exhibit some linguistic forms, such as personal pronouns, more frequently than do less interactive text genres, such as scientific articles in academic journals; conversations may exhibit other forms, such as verbal passive constructions, less frequently. In a traditional approach, one might look at the relative frequencies of these features in conversations compared to other text types in order to establish the fact that conversations are more interactive.

Biber suggests a different approach, in which the functional dimensions of variation are not pre-defined according to analysts' ideas about functional differences in discourse; rather, "quantitative techniques are used to identify the groups of features that actually co-occur in texts, and afterwards these groupings are interpreted in functional terms. The linguistic dimension rather than functional dimension is given priority" (13). The "frequent co-occurrence of a group of linguistic features in texts is indicative of an underlying function shared by those features" (64). This is the theoretical innovation upon which the quantitative multidimensional analyses are then developed.

Biber collects empirical frequency data for a wide range of linguistic features (in Biber 1988, 67 different linguistic features: 77f.) from 481 texts taken from the LOB Corpus of written British English and the London-Lund Corpus of Spoken (British) English. For each individual text, the normalized frequencies (per 1000 words) of the features are calculated for each of the 67 features. A correlation matrix is then constructed by calculating Pearson's r-scores for each feature combination, based on the mean feature frequency from the 481 different texts. The resulting matrix shows the correlation of each feature-feature combination.[1] These values range from -1 (no co-occurrence whatsoever in texts; i.e. total negative association) to 1 (complete co-occurrence, i.e. total positive association). A value

[1]Pearson's product–moment coefficient is commonly used, but matrices of covariance are also widely used in factor analysis; Biber utilizes correlation. For details consult Jolliffe 2002.

of 0 would indicate a neutral association: The feature is as likely (or unlikely) to co-occur with the other feature as with all the features in the corpus. Finally, a factor analysis is conducted, in which the normalized frequencies of the 67 features are treated as mathematical variables; their total variance is reduced to underlying factors which represent significant shared variance among the features that associate with each factor.[2]

The resulting factor groupings are then interpreted as the empirical "textual dimensions" which determine the range of variation for English texts, as they are represented in the sample data: The first dimension, "Involved versus Informational Production", is based upon positive co-occurrence of linguistic features such as 2nd-person pronouns, *be* as a main verb, or present tense forms, among others. Other features, such as the frequencies of noun forms or word length, have a strong negative association with this dimension (102, 129ff.). The second dimension, termed "Narrative versus Non-Narrative Concerns", is based on a positive association for the frequency of features such as past-tense verbs, 3rd-person pronouns, and perfect aspect. Present-tense verbs, for example, are negatively associated with this dimension. The third major dimension, "Explicit versus Situation-Dependent Reference", is strongly associated with wh-relative clause frequency, but negatively associated with time and place adverbial frequencies. Additional dimensions are adduced based on the remaining four factors, which account for much less of the total variation in the data.

6.1 Choice of Features to be Analyzed

Biber utilizes the concept of dimensions to potentially identify the configurations of co-occurrence of linguistic features for all possible types of discourse, including a great variety of written and spoken text types. A provisional identification of some of the salient functional and situational parameters of social media and Twitter has been provided above

[2]Biber 1988 chooses 7 factors based on the structure of the unrotated eigenvalues for the preliminary factors (83f.)

in Section 2.2; a dimensional analysis of feature co-occurrence may shed light on those discourse strategies that are most characteristic of Twitter.

However, certain text–immanent conditions of Twitter messages partially determine the parameters within which feature variables can be manifest in a multi-dimensional analysis. The first is the length restriction of tweets to 140 characters. This parameter was set in 2006 to encourage intercompatibility with SMS messaging services, which were often restricted to 160 characters; Twitter reserved 20 characters for additional meta-information such as username and url links.

The character limit for tweets sets a practical upper limit on the clausal complexity of messages, making an analysis of some syntactic structures, such as complex multi-clause sentences, irrelevant: They simply do not occur with high frequency in the data. For example, syntactical constructions such as present participial clauses (*Sitting at my computer,*) can be combined with pied-piping relative clauses (*the way in which I will write this tweet*), followed by a copula, a prepositional phrase (*is by typing on my smartphone,*) and one or more sentential relative or subordinate clauses (*which is slower than typing on a keyboard, although much easier to do when one is standing in the bus.*) In this analysis, we have not investigated the combinatorial possibilities of clausal constructions, mainly because multiple-clause sentences are longer than is the norm for Twitter, but also due to the inherent difficulty of automatically parsing clause constructions, particularly in formats such as Twitter that exhibit non-standard syntax and orthography to a significant degree, as well as widespread use of symbols such as emoticons. For this reason our dimensional analysis utilizes fewer features than do Biber's, and focus on those linguistic features which are relatively frequent, even in short messages.

6.2 Principal Component Analysis and Factor Analysis

For principal component analysis (PCA), a large number of variables is reduced to a much smaller number of variables while retaining as much of the variation of the data set as possible. While PCA calculates eigenvectors for the covariance of a number of variables and uses eigenvalues to explain the proportion of variance explained by each component, factor analysis creates a model of presumed underlying variables and solves for multiple equations (see Jolliffe 2002, 150ff.). The presumed model of the underlying variability takes the following form:

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + e_1 \\ x_1 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2m} f_m + e_2 \\ &\vdots \\ x_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pm} f_m + e_p \end{aligned} \tag{6.1}$$

where $x_1, x_2, \ldots, x_p$ represent the random variables in the data; $f_1, f_2, \ldots, f_m$ the factors in the presumed underlying model; $\lambda_{jk}, j = 1, 2, \ldots, p$, and $k = 1, 2, \ldots, m$ constants (the factor loadings); and $e_j, j = 1, 2, \ldots, p$ unique error values (Jolliffe 2002: 151). The set of linear equations underlying the factor analysis model only has a solution if the corresponding matrix $\Lambda f + e$ has a non-zero determinant, i.e. is invertible.[3]

Treating individual tweets (in both the Finland English and Comparison Corpus data sets) as texts results in correlation matrices (created from determining the Pearson prod-

[3]A matrix with identical rows or columns, or rows or columns which are equivalent by row or column operations, has a non-invertible determinant and is thus computationally singular; its set of linear equations cannot be solved.

uct–moment correlation over 32,916 and 181,861 values for 37 variables) with 609,000 and 3.364 million unique cells, respectively.[4]

Many tweets in the data are relatively short. If we were to consider only the tags corresponding to the grammatical features, a short tweet could, for example, correspond to a sequence such as PRP VBN RB UH, indicating the message consists of a personal pronoun, followed by a non-3rd-person singular present verbal form, an adverb, and an interjection. This tweet would be represented by a numerical vector of feature frequencies with four *1* entries and 33 *0* entries (for the 37 distinct grammatical features). A correlation matrix created from a large number of tweets would inevitably feature a number of tweets with this specific pattern of features and thus result in some individual rows being equivalent. A factor analysis of the frequency data would then be impossible due to the equivalence or linear dependence of the matrix and the resulting insolubility of the corresponding linear equations.

There are various methods for surmounting this obstacle. One could aggregate the data into larger sections or chunks. The normalized feature frequencies in the equal chunks (which are much longer than individual tweets) can then be used to create a correlation matrix; if the chunks are large enough, and the distribution of features not highly irregular, the chance that any two rows in the resulting matrix would exhibit linear dependency would be effectively zero.[5] If any obvious linear dependencies remain, they can be manually identified and, if shown to be only minimally useful (e.g. because they result from the application of a tag only a handful of times in the entire data), removed.

A second approach would be to conduct a principal component analysis. PCA accounts for all the variance in the data by constructing a linear composite of variance based on

[4]Pearson's correlation quantifies the relationship between two continuous random variables; there is no assumption that the variables are normally distributed.

[5]Regularizing linguistic data for statistical analysis by dividing it into equal–length chunks was first suggested by Markov (2006[1913]); it is a common method in lexical statistics. See Baayen (2008: 135ff.).

observed variables. Factor Analysis, on the other hand, accounts only for shared variance by constructing a linear model of underlying factors that are presumed to account for the shared variance. PCA can be utilized for matrices that are non-invertible. In practice, PCA and factor analysis often will result in similar patterns of variation being identified in the data.

Other options for dimensionality reduction exist, for example the substitution of parametrized values derived from minimum residuals rather than eigenvalues in the covariance or correlation matrices in the case of linear dependency in the data. Kline (2014) describes the linear algebraic procedures used for various approaches. Minimum residuals-based factor analysis code has been developed for *R* in, for example, the *psych* package (Revelle 2014). For further details, Jolliffe (2002) provides a discussion of the differences between factor analysis and principal component analysis. Baayen (2008, ch. 5) introduces several statistical dimensionality reduction procedures and some of the code necessary for their implementation in *R*.

6.2.1 Dimensionality Reduction and Structure of Data

The decision was made undertake both of the main dimensionality reduction procedures described above. In order to conduct an exploratory factor analysis, feature values derived from 100 equal–length "chunks" of tweet data were used to construct correlation matrices of the individual variables. PCA was also undertaken, based on correlations derived from feature values per tweet.

Both approaches are useful for measuring and visualizing the variability of features in the data. Feature frequencies per chunk or per tweet were standardized by converting

them to z–scores (i.e. distance from the mean value for that feature, expressed in units of standard deviation).[6]

6.3 Factor Analysis

A factor analysis requires solving a set of linear equations that represent the underlying shared variance in a data set. The factor analysis in the following subsection was conducted in *R* using the maximum likelihood extraction method and varimax orthogonal rotation.

6.3.1 Factor Analysis of the Finland English Corpus

As noted above, factor analysis is only possible for non–invertible matrices that consist of non–identical columns and rows. In order to facilitate an exploratory factor analysis, the Finland English Corpus was divided into 100 chunks of approximately 4400 tokens each and the Comparison English Corpus into 100 chunks of approximately 29,000 tokens: These chunk lengths are large enough so that each row in the derived correlation matrix has a unique value for each of the 37 grammatical feature variables.

The results of the factor analysis with 37 variables can be seen in Table 6.1 (only factors with a loading ≥ 0.3 are shown).

The first factor has strong negative loadings on tags associated with singular proper nouns and urls and somewhat strong negative loadings on tags associated with punctuation, singular or mass nouns, hashtags, and prepositions. A strong positive loading is associated with non-3$^{\text{rd}}$-person singular present verb forms; almost as strong is the positive loading on

[6]Converting to z-scores does not change the factor loadings for a factor analysis, which are already based on a product-moment correlation, but does affect component loadings in a PCA based on a whether covariance or correlation is being used to calculate component eigenvectors. Scaling a PCA (i.e. using a correlation matrix) allows us to better grasp the interrelationship between the variables and how their co-occurrence may reflect discourse functions. Not scaling a PCA (i.e. using a covariance matrix) can better show the extent to which some variables (e.g. nouns) are highly frequent and others (e.g. comparative adverbs) are relatively infrequent; expressed in terms of eigenvalues.

TABLE 6.1: Exploratory Factor Analysis for the Finland English Corpus Data: Factor Loadings for 37 Variables and Seven Factors (Chunks as Data)

	Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
1	Punctuation (*: ; ... + - = <> / /*)	-0.50						
2	Modal verb	0.58						
3	Proper noun, singular	-0.74			-0.36			
4	Personal pronoun	0.78				-0.60		
5	Universal Resource Locator	-0.72			-0.47			
6	Verb, base form	0.68						
7	Verb, non-3rd person singular present	0.70				-0.31		
8	Quotation mark		-0.59					
9	Retweet		-0.51					
10	Username (preceded by @)	0.32	-0.53					
11	Determiner			0.82				
12	Noun, singular or mass	-0.38	0.38	0.51				
13	Interjection		-0.43	-0.50				
14	Adverb				0.56			
15	Hashtag	-0.46					-0.80	
16	*. ! ?*							-0.61
17	Comma					0.38		
18	Coordinating conjunction				0.50			
19	Cardinal number							
20	Existential *there*							
21	Preposition	-0.47	0.48					
22	Adjective							
23	Adjective, comparative				0.35			
24	Adjective, superlative							
25	Noun, plural					0.47		0.39
26	Possessive pronoun		0.38		0.38			
27	Adverb, comparative							
28	Adverb, superlative							
29	Particle							
30	*to*		0.41					
31	Verb, past tense				0.34	-0.47		
32	Verb, gerund or present participle		0.43					
33	Verb, past participle		0.32					
34	Verb, 3rd person singular present							0.32
35	Wh-determiner							
36	Wh-pronoun							
37	Wh-adverb			0.31	0.34			
	Cumulative Variance	0.12	0.18	0.23	0.28	0.33	0.36	0.39
	$\chi^2 = 760.75$, df = 428, p-value = $5.49e^{-21}$							

verbal base forms. The positive association with modal verbs and usernames is somewhat strong. This factor may represent, to an extent, discourse that expresses information about attitudinal or modality stances of the user: An example from the Finland English Corpus that fits these parameters is *I think I might be sick*. Dialogic discourse directed at

other users could also fall under the parameters of this factor;[7] for example the message from the Finland English Corpus *@britt_ underwood can u follow?*, whereas messages with somewhat denser informational content, expressed in the form of proper nouns such as place names, noun or prepositional phrases, or explicitly marked by a hashtag or a url address, are disfavored.

The second factor has positive loadings for prepositions, possessive pronouns, the word *to*, and gerund or participial forms, and negative loadings for quotation marks, retweets, usernames, and interjections. The first three negative loadings suggest that the extent of discourse reference to other user messages or non-immediate information is limited; the negative loading on interjections suggests relatively little overt affective orientation towards the proposition expressed in the tweet user message. The positive loadings suggest compound verbal forms are used in messages circumscribed by this factor. As pronouns do not have a significant loading for this factor, it can circumscribe tweet messages that do not correspond to standard sentence structure, for example by omission of subject pronouns. Examples from the Finland English Corpus that correspond to these parameters include user messages like *cleaning my room turned out to be more difficult than ever planned* and *gym this morning, and now lounging around watching all my videos on YouTube.*

The third factor has strong positive loadings for determiners and moderately strong positive loadings for singular and mass nouns and Wh-adverbs, as well as a negative loading for interjections. The simplest tweet user message type circumscribed by these factor loadings is a short direct question message; examples from the Finland English Corpus include tweet user messages such as *where's this place?* and *where did the weekend go?*.

The fourth factor has positive loadings for adverbs, conjunctions, comparative adjectives, possessive pronouns, past tense verbs, and Wh-adverbs, and negative loadings for

[7]The factor also has weak positive loadings associated with tags such as Wh-pronoun and Wh-adverb that represent question words; because the values (0.24 and 0.21) are not > 0.3, they are not displayed in the table.

proper nouns and urls. Based on the loading on verbal forms, this factor could represent user status reports or commentary on events that he or she has experienced, expressed using several phrases or clauses. The negative loadings for proper nouns and urls again suggest that specific place names (whether physical or online) do not figure prominently in the messages circumscribed by this factor.

Factor five has positive loadings on commas and plural nouns, and negative loadings on personal pronouns, verbal base forms, and past tense verbs. The factor is somewhat difficult to interpret in terms of its putative communicative function based on these parameters alone, but it may be the case that this factor circumscribes messages that contain impersonal informational content expressed mainly in noun phrases in present-tense verb forms, perhaps with some clausal or phrasal complexity.

The sixth factor, consisting only of a strongly negative loading for the feature hashtag, cannot be interpreted in terms of communicative function. Equally problematic is the seventh factor, which consists of positive loadings for plural nouns and 3$^{\text{rd}}$–person singular present verb forms and a stronger negative association for the period, exclamation mark, and question mark.

As shown in Table 6.1, seven factors account for only 39% of the total variance in the data. This value is not particularly high, but is not too different from the figure of 51% in Biber (1988). Our preliminary interpretation of these factors can be kept in mind in the following principal component analysis.

The loadings of the first two factors are plotted in Figure 6.1. The spatial position of the features can be interpreted as corresponding to the extent to which the features play shared or different roles in the communicative functions that may underlie the two factors. A prominent group on the right–hand side of the figure shows modals (MD), verbal infinitives (VB), personal pronouns (PRP) and non-3$^{\text{rd}}$-person singular present verbal forms in close

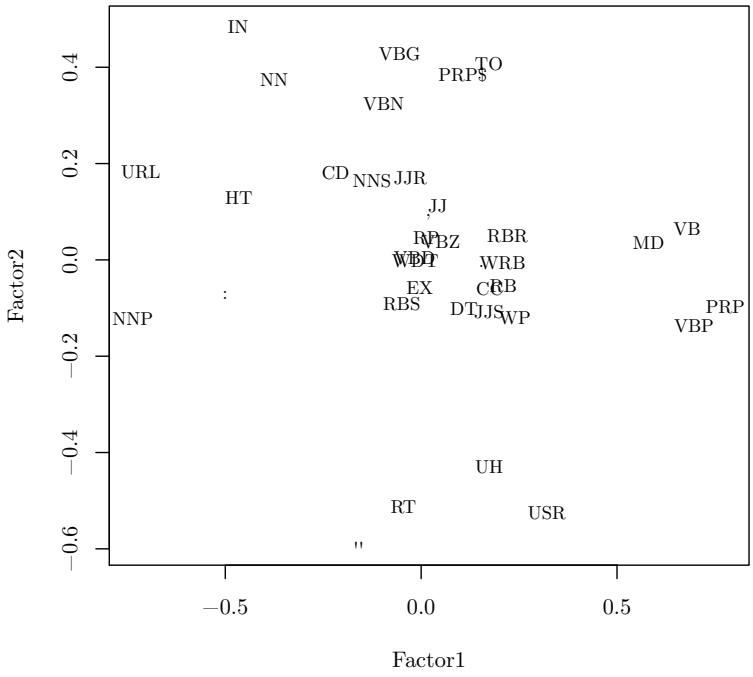

FIGURE 6.1: Plot of Factors 1 and 2 for the Finland English Corpus (Chunks as Data)

association. These features are primary resources used to give expression to personal attitudinal and epistemic stance and used to interact with other users.

The features interjections (UH), usernames (USR), retweets (RT) and quotation marks (“) form another group near the center bottom of Figure 6.1. As usernames and retweets are specific to Twitter communication, and the interjection tag in the Finland English data is most frequently applied to emoticon types, these features can be interpreted as indexing specifically Twitter–based patterns of user interactivity. The use of these features is a salient characteristic of the variety Finland Twitter English.

In the top left quadrant of the figure the tags prepositions (IN), nouns (NN), urls (URL) and hashtags (HT) can be found, in proximity to the tags numbers (CD) and plural

nouns (NNS). These features are typically used to express informational content; their spatial proximity according to a factor analysis suggests they might share an underlying communicative function.[8]

6.3.2 Factor Analysis of the Comparison English Corpus

TABLE 6.2: Exploratory Factor Analysis for the Comparison English Corpus Data: Factor Loadings for 37 Variables and Seven Factors (Chunks as Data)

	Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
1	Punctuation (*: ; ... + - = <> / /*)	-0.78	-0.36					
2	Coordinating conjunction	0.58						
3	Cardinal number	-0.54	-0.31					
4	Adjective	0.58	0.38					
5	Modal verb	0.72						
6	Proper noun, singular	-0.86	-0.31					-0.36
7	Personal pronoun	0.87				-0.39		
8	Possessive pronoun	0.50						0.37
9	Adverb	0.74						
10	Universal Resource Locator	-0.81	-0.31					
11	Username (preceded by @)	0.56				-0.39		
12	Verb, base form	0.82						
13	Verb, past tense	0.57						
14	Verb, non-3rd person singular present	0.91						
15	Wh-determiner	0.52						
16	Wh-pronoun	0.60						
17	Wh-adverb	0.74						
18	Particle	0.30	0.60					
19	*to*	0.37	0.67		0.35			-0.34
20	Verb, gerund or present participle		0.83					
21	Determiner			0.75				
22	Existential *there*		-0.49	0.55				
23	Preposition		0.30	0.69	0.32			
24	Verb, past participle			0.64				
25	Hashtag				0.62			
26	Noun, plural				0.60	0.33		
27	Quotation mark					0.66		
28	Noun, singular or mass		0.51	0.38		0.52		
29	Adverb, superlative						0.56	
30	Comma			0.38		0.34	-0.40	
31	*. ! ?*	0.47			-0.35			
32	Adjective, comparative	0.41						
33	Adjective, superlative	0.31					0.34	
34	Adverb, comparative	0.38			0.43			
35	Retweet				0.46	-0.47		
36	Interjection	0.31		-0.47				0.44
37	Verb, 3rd person singular present							
	Cumulative Variance	0.26	0.35	0.42	0.48	0.53	0.56	0.59
	$\chi^2 = 916.88$, df = 428, p-value = $1.05e^{-37}$							

[8]The determiner tag (DT) is located quite close to the center of Figure 6.1. The weak association between determiners and nouns in the Finland English data may reflect the overall low use of determiners such as articles in Finland Twitter English; this may be a result of linguistic interference.

The same procedure was used to conduct a factor analysis of grammatical feature frequencies in the Comparison English Corpus. The features with factor loadings ≥ 0.3 are shown in Table 6.2, and a plot of the features for the first two factors in Figure 6.2.

For the Comparison Corpus, the first factor has strong positive loadings for modal verbs, personal pronouns, adverbs, verbal base forms, non-3rd-person singular present verb forms, and Wh–adverbs. Moderately strong positive loadings are present for the tags punctuation (colon, semicolon, ellipsis, basic mathematical symbols) and other punctuation (period, exclamation mark and question mark), for conjunctions, adjectives, possessive pronouns, usernames, past-tense verb forms, Wh–determiners and Wh–pronouns, phrasal verb particles, *to*, comparatives, and interjections. Strongly negative loadings, for the first factor, are present for other punctuation (colon, semicolon, ellipsis, square brackets, etc.), proper nouns, and urls; cardinal numbers have a moderately strong negative loading.

The comparison tweets are somewhat longer, on average, than the Finnish tweets (mean 78.99 characters and 15.73 tokens vs. 71.10 characters and 13.27 tokens), so it is not surprising that a factor analysis shows more features as likely to co-occur in a single tweet. The combination of positive and negative loadings for the first factor may circumscribe user reports and updates of their own personal activities in present and past tenses (as evident from the loadings for personal pronouns and some verbal forms) as well as dialogic modes such as questions posed to other Twitter users (as evident from the loadings on Wh–words).

The relatively higher number of features associated with the first factor for the Comparison English data suggests that tweets from the Comparison English Corpus comprise more grammatically and syntactically complex discourse structures than those of the Finland English Corpus. The features implicated in the second factor shed additional light on the discourse nature of the Comparison Corpus tweets.

For the second factor, present participial and gerund forms have a strong positive loading, followed by moderately strong positive loadings for *to*, phrasal particles, nouns, adjectives, and prepositions. The features punctuation (colon, semicolon, ellipsis, basic mathematical symbols) and other punctuation (period, exclamation mark, question mark), numbers, proper nouns, urls, and existential *there* have moderately strong negative loadings for the factor.

This interpretation would then allow the first factor to be interpreted as a broader category that includes affective and interactive communicative functions, whereas the second factor presents informational content in the present tense, either as non–sentence participial and gerund phrases without subject pronouns, or using the third person (this is suggested by the strong positive factor loadings for past tense verbs and personal pronouns in factor 1). Specific information, such as proper nouns, quantities and urls are disfavored by this factor.

The third factor in the Comparison English Corpus shows a strongly positive factor loading for determiners, and moderately strong positive loadings for existential *there*, prepositions, past participles, singular nouns, and commas. The only negative loading is for interjections. This factor is somewhat difficult to interpret, but could represent informational content of an impersonal nature. A tweet such as *There's nothing left to eat in the house, need to go shopping!* might fit into this group.

Factor four has moderately positive loadings for *to*, prepositions, hashtags, plural nouns, comparative adverbs, and retweets. Other punctuation (period, exclamation mark, question mark) has a negative loading. This factor may circumscribe tweets which are sentence fragments commenting on topics or other Twitter discourse.

The fifth factor includes moderately strong loadings for plural nouns, quotation marks, singular or mass nouns, and commas, and moderately strong negative loadings for personal pronouns, usernames, and retweets. The combination of features suggests verbal neutrality

as to tense, mood and aspect. The dimension seems to suggest short informational messages without explicit reference to the self, other persons, other Twitter messages or other Twitter users.

Factor six, which has positive loadings for superlatives and a negative loading for commas, can be interpreted as a communicative context in which short propositional superlative relations are introduced without clausal or extensive adjectival qualification. This may not represent a distinct communicative function, but rather a mode of expression with some overlap in the functionality with factors one and two.

The final factor has a negative loading on proper nouns and *to*, and positive loadings on possessive pronouns and interjections. Many types of messages could be circumscribed by this dimension, for example, non-sentence short self-referential affective stance indicators consisting of emoticons.

Although the specific variables associated with each factor for the Comparison English Corpus factor analysis are not the same as for the Finland English Corpus factor analysis, there are correlations between the items with shared polarity. This is manifest as the clustering of the individual variables in Figure 6.2: A cluster in the upper half of the center of the figure groups elements common in descriptive or informational phrases, such as nouns (NN), prepositions (IN), plural nouns (NNS) and 3rd-person-singular present verb forms (VBZ); determiners (DT) are situated not far away.

Strong positive loadings are associated with personal pronouns and non-3rd-person-singular present verb forms; these word forms are used for one of the principal communicative functions of Twitter: short personal status reports such as *I am (x)*. The close proximity of these tags (PRP and VBP) on the right hand side of Figure 6.2 to modals (MD), Wh-words (WP, WDT, WRB) and usernames (USR) suggest that the interactive and dialogic communicative functions of Twitter discourse utilize similar grammatical forms

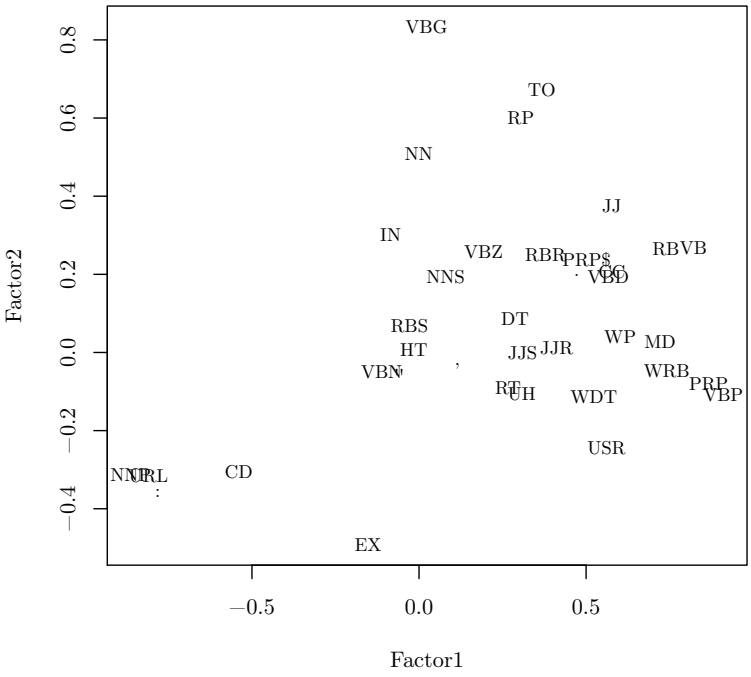

FIGURE 6.2: Plot of Factors 1 and 2 for the Comparison English Corpus (Chunks as Data)

and operate as markers of communicational interactivity; a similar grouping of personal pronouns with verbal forms is found in the Finland English Corpus factor analysis.

An additional visual cluster common to both datasets groups Twitter features such as retweets (RT) and usernames (USR) in relatively close proximity; these features are found to be in association with the interjection tag (UH). As this tag is often applied to the emoticon types in the data, the feature cluster can be thought of as marking CMC or Twitter–specific communicativity. Once again, the same grouping was discovered in the Finland English Corpus.

In both data sets, urls (URL), proper nouns (NNP), punctuation (:), and, to a certain extent, hashtags (HT) occupy the same space. Proper nouns, colons and urls are associated

with automatically generated text created by apps that append specifying information to geo-encoded user texts, while hashtags are user-generated explicit content specifiers.

In summary, a factor analysis of the two principal corpora shows that grammatical features are used in remarkably similar patterns in the discourse of the corpora, suggesting that the same communicative functions may underlie the feature frequencies.

6.4 Principal Component Analysis

Principal component analysis reduces data dimensionality by orthogonally transforming values from covariance or correlation matrices, similar to factor analysis, but differs in that it considers the total variance of a data set and not only the portion of variance in a data set that is shared by all of the variables under consideration. As PCA does not require covariance or correlation matrices that are non–invertible for the PCA conducted on the Finland English and Comparison English data, each individual tweet in the following analysis was modeled as a variable using functionality in *R*.

6.4.1 Principal Component Analysis of the Finland English Corpus

TABLE 6.3: Principal Component Analysis of the Finland English Data, Summary of the the First Sixteen Components

	Importance of Components							
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	2.019	1.489	1.204	1.188	1.118	1.088	1.055	1.048
Proportion of Variance	0.110	0.059	0.039	0.038	0.033	0.032	0.030	0.029
Cumulative Proportion	0.110	0.170	0.209	0.247	0.281	0.313	0.343	0.373
	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16
Standard deviation	1.034	1.016	1.014	1.003	0.998	0.995	0.989	0.986
Proportion of Variance	0.028	0.027	0.027	0.027	0.026	0.026	0.026	0.026
Cumulative Proportion	0.402	0.430	0.458	0.485	0.512	0.539	0.565	0.591

Table 6.3 provides a summary of individual feature variance and the total variance accounted for by the first sixteen components in a principal component analysis of the

TABLE 6.4: Principal Component Analysis of the Finland English Data, Loadings of the the First Two Components

Loadings of Components 1 and 2, Finland English PCA					
Feature	Comp. 1	Comp. 2	Feature	Comp. 1	Comp. 2
Quotation mark		0.107	Adverb	-0.257	
Comma	-0.161	0.118	Adverb, comparative		
Other punctuation marks (. ? !)	-0.217		Adverb, superlative		
Punctuation (: ; ... + - = <> / /)		0.232	Particle		
Coordinating conjunction	-0.253		Retweet		
Cardinal number		0.207	*to*	-0.211	
Determiner	-0.267	0.198	Interjection		-0.215
Existential *there*			Universal Resource Locator	0.146	0.374
Hashtag	0.104	0.204	Username (preceded by @)		-0.204
Preposition or subordinating conjunction	-0.243	0.317	Verb, base form	-0.276	-0.179
Adjective	-0.217	0.102	Verb, past tense	-0.173	
Adjective, comparative			Verb, gerund or present participle	-0.135	
Adjective, superlative			Verb, past participle	-0.117	
Modal	-0.170	-0.168	Verb, non-3rd-person singular present	-0.237	-0.235
Noun, singular or mass	-0.280	0.266	Verb, 3rd-person singular present	-0.147	0.105
Proper noun, singular		0.344	Wh-determiner		
Noun, plural	-0.167	0.160	Wh-pronoun		
Personal pronoun	-0.327	-0.267	Wh-adverb	-0.113	-0.104
Possessive pronoun	-0.168				

Finland English Corpus; these components collectively explain 59% of the total variance in the data. The loadings for the first two components are shown in Table 6.4 and plotted in the subfigure on the left in Figure 6.3. The biplot indicates where each individual tweet in the corpus falls in terms of the two components that account for the largest proportion of the total variance in the Finland English data and shows the locations calculated for the grammatical features for the first two components. Tweets are represented in the space as small circles. On the subfigure on the right, the z–axis shows the density (number of tweets) for components 1 and 2.

For the most part, the component loadings have relatively low values in the range $-0.2 < x < 0.2$, but some variables, such as pronouns, nouns, adjectives, determiners, hashtags, proper nouns, urls, base verbal forms, adverbs and non-3rd-person inflected verbal forms, show stronger loadings on the first two components. For the Finland data, some of the remarks made about the clustering of grammatical features for the exploratory factor

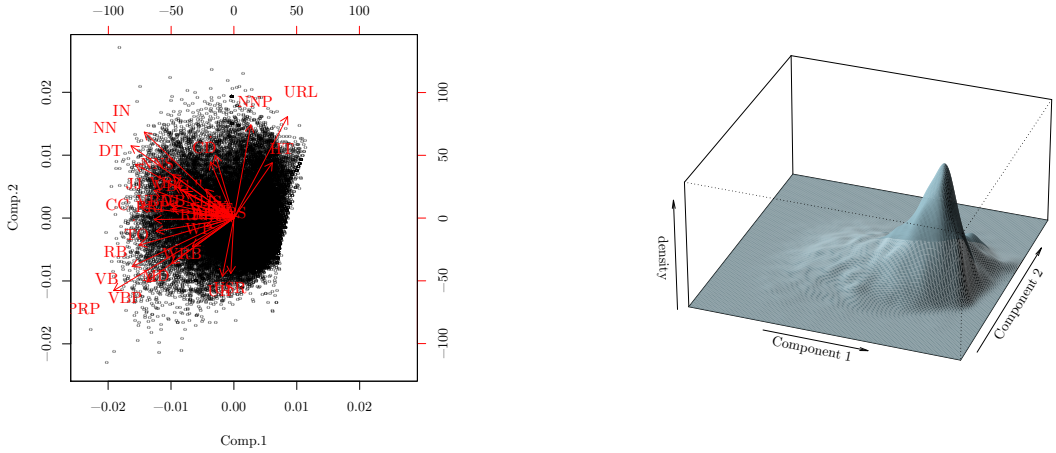

FIGURE 6.3: PCA biplot and Density for Components 1 and 2 of the Finland English Corpus (Tweets as Data)

analysis seem to fit for the principal component analysis as well. Determiners, prepositions and nouns (DT, IN, NN) have similar negative loadings on the first component, i.e. they co–occur in tweets and are found close to one another in the upper left quadrant of the left–hand subfigure in Figure 6.3. Non-3$^{\text{rd}}$-person-singular present verbal forms (VBP), adverbs (RB), and uninflected verbal forms (VB) represent another cluster, associated with with personal pronouns (PRP). Interjections/emoticons (UH) co–occur with usernames (USR). Finally, hashtags (HT), web addresses (URL) and proper nouns (NNP) have positive loadings for the first two components, indicating they may play a role in shared communicative functionality, such as communication of specific information in the form of e.g. real or virtual place names.

TABLE 6.5: Principal Component Analysis of the Comparison English Data, Summary of the the First Sixteen Components

	Importance of Components							
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	1.897	1.485	1.253	1.198	1.103	1.071	1.057	1.045
Proportion of Variance	0.097	0.059	0.042	0.038	0.032	0.031	0.030	0.029
Cumulative Proportion	0.097	0.157	0.199	0.238	0.271	0.302	0.332	0.362
	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16
Standard deviation	1.031	1.017	1.011	1.009	1.003	0.999	0.997	0.993
Proportion of Variance	0.028	0.027	0.027	0.027	0.027	0.026	0.026	0.026
Cumulative Proportion	0.390	0.418	0.446	0.473	0.501	0.528	0.555	0.581

TABLE 6.6: Principal Component Analysis of the Comparison English data, Loadings of the the First Two Components

Loadings of Components 1 and 2, Comparison English PCA					
Feature	Comp. 1	Comp. 2	Feature	Comp. 1	Comp. 2
Quotation mark		-0.106	Adverb	-0.242	
Comma	-0.180	-0.132	Adverb, comparative		
Other punctuation marks (*. ? !*)	-0.288	0.131	Adverb, superlative		
Punctuation (*: ; ... + - = <> / |*)		-0.409	Particle	-0.100	
Coordinating conjunction	-0.218		Retweet		
Cardinal number		-0.215	*to*	-0.200	
Determiner	-0.269	-0.198	Interjection		0.140
Existential *there*			Universal Resource Locator	0.173	-0.376
Hashtag			Username (preceded by *@*)		0.289
Preposition or subordinating conjunction	-0.255	-0.282	Verb, base form	-0.279	
Adjective	-0.205		Verb, past tense	-0.179	
Adjective, comparative			Verb, gerund or present participle		
Adjective, superlative			Verb, past participle		
Modal	-0.178		Verb, non-3rd person singular present	-0.240	0.149
Noun, singular or mass	-0.265	-0.269	Verb, 3rd person singular present	-0.107	-0.107
Proper noun, singular		-0.402	Wh-determiner		
Noun, plural	-0.164	-0.153	Wh-pronoun		
Personal pronoun	-0.338	0.180	Wh-adverb	-0.104	
Possessive pronoun	-0.150				

6.4.2 Principal Component Analysis of the Comparison English Corpus

In the Comparison Corpus, the first sixteen principal component account for a proportion of the variance of the dataset (58%) similar to that of the Finland English Corpus (Figure 6.4).

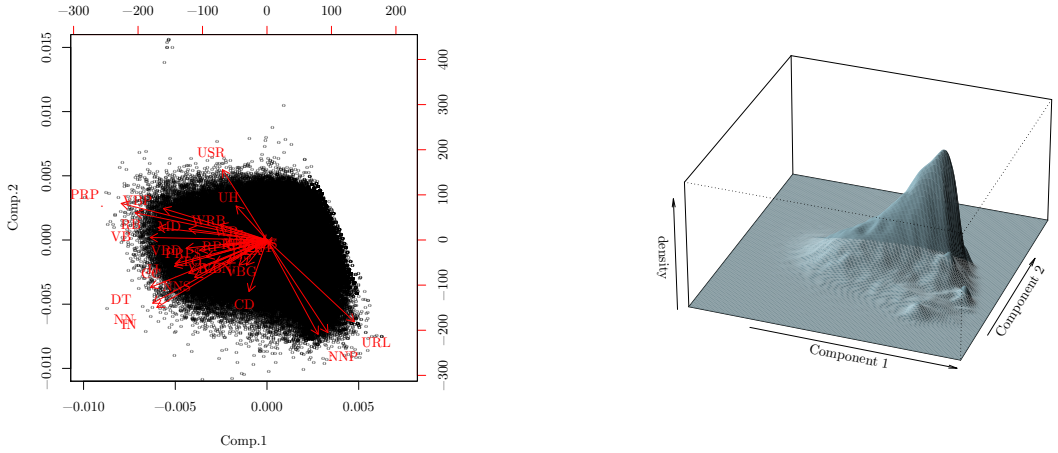

FIGURE 6.4: PCA Biplot and Density for Components 1 and 2 of the Comparison English Corpus (Tweets as Data)

Component loadings for the Comparison Corpus data differ from those of the Finland English Corpus data, but the loadings for individual features tend to share polarity with the same features as for the Finland English data.

As was the case for the Finnish data, the clustering suggests that the phrasal constituents prepositions, determiners and nouns co-occur and share communicative function: A cluster consisting of the tags DT, NN and IN can be found in the lower left quadrant of the left-hand subfigure in Figure 6.4. Adjectives are situated in the same quadrant, but as the eigenvalues associated with adjectival features are much smaller, the tag labels are situated quite close to the center of the plot. As is the case in the Finnish data, proper nouns and urls have similar factor loadings in the first two components.

6.5 Discussion of Factor Analysis and Principal Component Analysis

Multifactorial dimensionality reduction techniques such as exploratory factor analysis and principal component analysis demonstrate the extent to which grammatical features share underlying communicative functions as they can be mapped by eigenvectors derived from feature correlation or covariance matrices.

In a number of multi–dimensional analyses, Biber and others have demonstrated how texts from different genres or language comprising the different communication channels of speech and writing can vary according to aggregate measures of grammatical feature frequencies, as quantified using exploratory factor analysis; these differences may represent the presence or absence of underlying configurations of communicative functionality in different types of communication (Biber 1987; 1988; 1995; Biber, Conrad and Cortes 2003; Biber, Conrad and Cortes 2004). In this section, a multidimensional approach demonstrates that English as it is used on Twitter in Finland and as it is used on Twitter in general exhibits a patterning of co–occurrence of grammatical features that suggests common underlying communicative functions. Finland Twitter English emerges in a multi–dimensional perspective from the relatively distinct separation of interactive and informational communicative functions for individual factor groups compared to the Comparison English Corpus, where, although the same groupings are apparent, individual factor groups are less able to distinguish communicative functions. This may have to do with the fact that discourse in the Comparison English Corpus is relatively more complex. While the frequencies of individual grammatical features in the the Finland English Corpus and the Comparison English Corpus differ, the ways in which features co–occur is remarkably similar.

Chapter 7

Word Clusters: Lexical and Grammatical Bundles

Frequentist analysis of the role of multiple–word units of linguistic structure in the organization of discourse has become possible with the widespread use of computers for text analysis and the digitization of large numbers of texts. Some of the earlier theoretical perspectives that suggested n–grams play a role in discourse organization include those of Z. Harris. His *String Analysis of Sentence Structure* (1962) suggests that string analysis, at that time undertaken on the Univac computer, may complement grammatical approaches such as constituent or transformational analysis in the attempt to describe the sentences of a language.

> it is possible to decompose each sentence into elementary strings which combine (to form a sentence) in accordance with specified rules. If in a given sentence we find a sequence of words which cannot be assigned to any known string formula occurring in it in accordance with some known rule, then a new string or rule of occurrence has to be set up. The intention is that a few classes of strings, with

> simple rules describing how they occur in relation to each other, will suffice to characterize all sentences of the language. (Harris 1962: 9-10)

Mathematical Structures of Language (1968) further builds upon this approach, suggesting that 6–grams from a limited set of word categories may be sufficient for the description of most types of discourse. Summarizing this perspective in a 1982 paper, Harris notes that although discourse is typically formulated at the level of the sentence, "discourse... is not a matter of detailed restrictions on the sentence–structure sequence"; it is not dependent on sets of categorical rules for obtaining sentence equivalence through constituent transformation and substitution, but rather on "the fact that words recur in particular positions relative to other recurring words, within the word–class sequences which constitute the sentences of the discourse" (Harris 1982: 232).

The empirical investigation of the role of n–grams in discourse was not pursued substantially in subsequent decades, possibly in favor of conceptions of language in which formal theories of transformational grammar such as that of Harris' influential student Chomsky played a more important role. In an era of limited access to computing resources and modest processor and memory capabilities, a statistics–based frequentist approach to n–grams would be impractical. Baayen (2003: 230) remarks on the relative lack of interest in statistical approaches to linguistic analysis in the 1960s–1980s as a reflection of the relatively unsophisticated computer processor memory architecture of the time, in which statistical analyses of large amounts of data would have represented a near impossibility:

> Not surprisingly, the linguistic theories of the time took formal languages as the model for language, emphasizing the generative capacity of language, denying any role of importance to probability and statistics, and elevating economy of storage in memory to a central theorem.

As computers became more affordable and their processing power and memory increased, quantitative and statistical considerations of language began to inform work in language acquisition, perception, and production in the 1970s and 1980s (Bod, Hay and Jannedy 2003). For example, quantitative analysis of sociolinguistic variables using logistic regression was pioneered by Cedergren and Sankoff with the VARBRUL program in 1975.

Linguists associated with the "London School" and M. A. K. Halliday, such as John Sinclair, argued that a consideration of multiple-word units should inform theoretical conceptions of language and its modalities of use. Sinclair's "Principle of Idiom" suggests that the unit of meaning is not only the single word:

> A language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments. To some extent this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation. (Sinclair 1991: 110)

Semi–preconstructed phrases, as they are manifest in collocational tendencies and n–gram frequencies and analyzed as "lexical bundles", play an important role in written and spoken language.

The first empirical study of multi-word sequences seems to have been Altenberg and Eeg–Olofsson 1990, based on material from the London–Lund Corpus. Biber and others, building upon the earlier ideas of Sinclair and Harris, have proposed to analyze n–grams as lexical bundles, or sequences of words that occur frequently in natural discourse and constitute lexical building blocks in various communicative contexts (Biber et al. 1999: 990–991). Typically, lexical bundles are not idiomatic in meaning and not perceptually salient; on the contrary, the meaning of a lexical bundle is transparent from the individual

words contained in it (Biber 2006: 134). According to Biber, "the functions and meanings expressed by these lexical bundles differ dramatically across registers and academic disciplines, depending on the typical purposes of each" (2006: 174). It has been suggested that 4–word lexical bundles have a more readily recognizable range of structures and functions than 3–word bundles and 5–word bundles (Hyland 2008, Chen and Baker 2010). In this section frequent 4–gram lexical bundles in the Finland English and Comparison English corpora will be considered, as well as frequent sequences of grammatical word classes as suggested by Harris (1982).

7.1 Additional Processing Steps

Existing lexical bundle literature typically examines the frequencies of sequences of word types; punctuation and pseudo–words such as emoticons are usually not considered. The tokenization procedure of the CMU Twitter Tagger, used to annotate the Finland English and Comparison English corpora, treats punctuation and punctuation–based pseudo–words as distinct tokens; such tokens often represent meaningful units, such as emoticons. However, in some cases punctuation in tweets results from apps that manage images or location automatically by adding sequences including punctuation to user messages. For example, some of the most frequent 4–gram token collocations in the Finland English Corpus (if punctuation tokens are retained) are sequences such as *[pic] :*, *w 2 others)*, or *at (hel)*. Bundles such as these may play a role in the organization and maintenance of online discourse, but as they do not represent active linguistic construction by human users, they provide relatively little insight into the dynamics of Finland Twitter English as it is used by people. Thus, the decision was made to additionally filter punctuation from the data prior to calculating n–gram frequencies. After removing punctuation and filtering for lexical elements introduced by smartphone apps, frequencies of lexical bundles were calculated

for 4–word sequences. In addition, n–gram frequencies for grammatical features according to grammatical part of speech tags were calculated.

7.2 Lexical Bundles

For the Finland English and Comparison English data, lexical bundles were considered to be 4–gram word sequences that occur at a rate of at least 10 per million words (0.01 per thousand words).

7.2.1 Lexical Bundles in the Finland English Corpus

Using this measure, the Finland English Corpus shows an overall rate of 14.66 lexical bundles per thousand words. This rate is higher than rates reported in Biber, Conrad and Cortes for genres such as textbooks, academic prose, conversation, and classroom teaching, which range from approximately 2.5 to 8 lexical bundles per thousand words (2003: 380).

Some lexical bundles in the Finland English Corpus occur relatively frequently. In the Finland English data, after filtering for the most common automated tweets, the most frequent word 4–grams occur at a rate of almost 0.16 per thousand words. Biber, Conrad and Cortes, in their study of lexical bundles in academic discourse, analyze 4–token bundles that occur with a frequency of at least 40 per million words (0.04 per thousand words), but note that some bundles have frequencies of up to 0.2 per thousand words (2004: 376). In general, spoken conversation has higher overall frequencies of lexical bundles than does written language; the higher frequencies in the Finland English Corpus and Comaprison English Corpus could therefore reflect the somewhat conversational nature of Twitter, in which, like speech, "everyday language use is composed of prefabricated expressions" (Biber, Conrad and Cortes 2004: 372).

Table 7.1: Most Frequent Lexical Bundles in the Finland English Corpus

	Lexical bundle	Frequency per thousand words		Lexical bundle	Frequency per thousand words
1	can u follow me	0.158	11	please can u follow	0.057
2	i dont want to	0.126	12	can u help me	0.055
3	love you so much	0.110	13	having a good time	0.048
4	on my way to	0.089	14	is going to be	0.046
5	can you follow me	0.069	15	love u so much	0.046
6	i just want to	0.069	16	thank you so much	0.046
7	at the same time	0.066	17	i have no idea	0.043
8	i want to go	0.064	18	i have to go	0.043
9	cant wait to see	0.062	19	i wish i could	0.043
10	i dont know what	0.060	20	want to go to	0.043

Lexical bundles do not necessarily correspond to coherent grammatical or syntactic constituents such as a phrases or a clauses: They frequently overlap clause– or phrase boundaries. Biber, Conrad and Cortes suggest that lexical bundles that represent elements of verbal clauses are more frequent in spoken conversation, whereas bundles that represent noun phrases are more common in written texts (2004: 377). The most common bundles from the Finland English Corpus (Table 7.1) are weighted towards verbal clause elements: Fifteen of the twenty most frequent bundles contain verbal phrase elements; five contain nominal or prepositional phrase elements. The "orality" of Twitter discourse is in line with previous findings about the discourse properties of CMC genres such as chat or IM, and may reflect some of the parameters of production of Twitter user messages, such as real–time communication, interactivity, and a topical focus on personal concerns.

In terms of content, a prominent feature of the Finland English data as it is manifest in the most frequent lexical bundles is the frequency of words pertaining to the self: First-person pronouns (including possessives) appear eleven times in the twenty most frequent bundles, but second-person pronouns are also common, appearing seven times.

The most frequent types in the Finland English corpus reflect configurations of online interaction typical for Twitter use. Three of the twenty most frequent types include the

verb "follow" in the form of a request to other users; this refers to a setting in the Twitter interface that highlights the user messages of a particular user.

Biber, Conrad, and Cortes propose a taxonomy of the discourse functions of lexical bundles: The basic classes correspond to the discourse functions stance expression, discourse organization, referential expression, and "special conversational functions" such as politeness expressions (2004: 389ff.). They find that classroom teaching utilizes more stance and discourse organizing sequences than does normal conversation, but, surprisingly, classroom teaching also has higher frequencies of bundles with referential function than do textbooks.

The proposed taxonomy is useful for some of the most frequent lexical bundles in the Finland English Corpus. Stance expressions are well represented in the Finland English bundles, with attitudinal or modality stance expressions the most common type overall (*i dont want to*, *i just want to*, *i want to go*, *cant wait to see*, *is going to be*, *i have to go*, *i wish i could*, and *want to go to*), expressing volition/desire, prediction, or deontic modality. Direct requests, a subcategory of attitudinal/modality stance expressions, are associated with the "follow" user setting, and exemplified by the types *can u follow me*, *can you follow me*, and *please can u follow*, as well as *can u help me*. Epistemic stance expressions such as *i don't know what* or *i have no idea* also feature prominently in the Finland English data.[1]

There are three referential expression bundles among the 20 most frequent types in the Finland English data: *on my way to* and *at the same time* indicate place or time deixis, and *having a good time*, an adverbial, specifies an attribute (occurring most commonly as a complement of the personal pronoun *I*).

Finally, the types *love you so much*, *love u so much*, and *thank you so much* represent expressions of personal relationship maintenance or politeness; in the taxonomy proposed by Biber, Conrad and Cortes, they would be considered stance bundles.

[1]Epistemic stance refers to a speaker's attitude towards the truth value of a proposition. See Lyons (1977/1: 787).

Notably absent from the most frequent lexical bundles from the Finland data are n–grams that have discourse organization functions such as sequences containing demonstrative pronouns or Wh–words.

Overall, the Finland English discourse, as it is manifest in the most frequent word 4-grams, prominently features personal attitudinal or epistemic functions (mostly in conjunction with first-person pronouns), with direct requests and expressions of politeness or relationship maintenance additionally among the most frequent types. Other discourse functions, such as direct reference to entities beyond the immediate domain of the communicator, or topic focus/elaboration of discourse content already introduced, figure less prominently or not at all in the data.

7.2.2 Lexical Bundles in the Comparison English Corpus

Lexical bundles occur in the Comparison English Corpus at a rates of 0.486 per thousand words, less than half that of the Finland English Corpus.[2]

TABLE 7.2: Most Frequent Lexical Bundles in the Comparison English Corpus

	Lexical bundle	Frequency per thousand words		Lexical bundle	Frequency per thousand words
1	is going to be	0.076	11	the end of the	0.038
2	i am going to	0.072	12	its going to be	0.037
3	just got back from	0.060	13	on the way to	0.037
4	on my way to	0.056	14	at the same time	0.034
5	going to be a	0.053	15	the first time in	0.029
6	trying to figure out	0.049	16	i wish i could	0.029
7	getting ready to go	0.045	17	i have no idea	0.028
8	the rest of the	0.045	18	i cant wait to	0.027
9	i dont want to	0.043	19	in the middle of	0.027
10	cant wait to see	0.038	20	what do you think	0.027

In terms of the functional properties of the most common bundles (Table 7.2), the Comparison English Corpus has a slighly different profile from that of the Finland English Corpus. Like in the Finland English Corpus, lexical bundles incorporating verbal phrase or

[2]For the Comparison Corpus data, the cutoff frequency for a word sequence to be considered a lexical bundle was 34 occurrences; this reflects the fact that the Comparison English Corpus has almost seven times as many tokens as the Finland English Corpus.

clause elements are more common than lexical bundles comprising noun phrase elements: Thirteen of the twenty most frequent bundles in the Comparison English Corpus are verbal phrase- or clause-based; seven represent constituents of nominal or prepositional phrases.

There are some similarities between the functional properties of the word 4–gram types most common in both corpora. Like in the Finland English Corpus, stance expressions feature prominently among the most frequent word 4–grams in the Comparison English Corpus. The types *is going to be*, *going to be a*, *i dont want to*, *cant wait to see*, *its going to be* and *i wish i could*, and *i cant wait to* represent volition, prediction, or deontic modal stance expressions; *i am going to* is more common as an expression of modal futurity, although it could also represent referential deixis in the indicative mood.

The types *just got back from*, *on my way to*, *on the way to*, *at the same time*, and *the first time in* are referential expression specifiers with place and time deixis functions. The types *trying to figure out* and *getting ready to go* are adverbial referential specifiers used as complements, most often to the personal pronoun *I* (although the former may index epistemic stance and the latter verbal tense/aspect/mood functions).

The types *the rest of the*, *in the middle of*, and *what do you think* have a discourse organization function; they are interpretable in the context of prior or subsequent constituent structure.

The most frequent Comparison English Corpus bundles index attitudinal and epistemic stance modality, as do the Finland English Corpus bundles, but feature a somewhat broader range of communicative functionality in terms of discourse organization. Finland English bundles more directly make reference to the language medium of Twitter (*can u follow me*), whereas Comparison English functions do not.

7.3 Grammatical Bundles

The data was also considered in terms of the frequency of sequences of grammatical features as determined by the CMU Twitter Tagger. Compiled in this way, the data have a slightly different profile. Part–of–speech sequences may show some functional or structural properties more clearly than lexeme bundles, whose functional roles as e.g. epistemic, attitudinal, or deontic stance markers cannot always be classified categorically by the presence or absence or formal elements.

An interesting aspect of the most frequent part–of–speech sequences is the fact that many of them would be considered ungrammatical according to conventional standards of English text production. Twitter user messages are often constructed in a way that represents spoken language; in the tradition of rhetoric and classical grammarians this is referred to as *anacoluthon.*[3]

7.3.1 Grammatical bundles in the Finland English Corpus

Table 7.3: Most Frequent Grammatical Bundles in the Finland English Corpus

	Sequence	Frequency per thousand words		Sequence	Frequency per thousand words
1	HT HT HT HT	5.586	11	DT NN IN DT	1.126
2	NNP NNP NNP NNP	3.758	12	DT JJ NN IN	1.115
3	NN IN DT NN	2.231	13	IN DT NN NN	1.096
4	PRP VBP TO VB	1.751	14	NN IN PRP$ NN	1.078
5	IN DT JJ NN	1.462	15	USR USR USR USR	1.037
6	PRP VBP DT NN	1.389	16	TO VB DT NN	0.952
7	NNP IN CD NNS	1.325	17	PRP MD VB PRP	0.913
8	IN DT NN IN	1.295	18	IN NNP NNP NNP	0.899
9	NNP NNP NNP URL	1.284	19	PRP VBP PRP VBP	0.826
10	IN CD NNS URL	1.172	20	NNP NNP IN CD	0.815

[3] "We can hardly conclude even so desultory a survey of grammatical misdemeanours as this has been without mentioning the most notorious of all. The anacoluthon is a failure to follow on, an unconscious departure from the grammatical scheme with which a sentence was started, the getting switched off, imperceptibly to the writer, very noticeably to his readers, from one syntax track to another". (Fowler and Fowler 1962 [1931]: 371).

The most frequent grammatical category 4–gram sequences from the Finland English data are shown in Table 7.3.[4] The most common sequence is a series of hashtags. Four singular proper nouns in sequence is the second most common type in this set; this, however, may be an artifact of the tokenization and tagging code of the CMU Twitter Tagger: Word forms that are capitalized and don't exist in the model data are assigned this tag. The sequence noun-preposition-determiner-noun is next most common, followed by personal pronoun-verb (non-3rd-person-singular-present)-*to*-verb (base form), preposition-determiner-adjective-noun, personal pronoun-verb (non-3rd-person-singular-present)- determiner-noun, singular proper noun–preposition-cardinal number-plural noun, preposition-determiner-noun-preposition, singular proper noun-singular proper noun-singular proper noun-url, and preposition-cardinal number-plural noun–url.

[4]A description of the tagset can be found in Table 3.6.

TABLE 7.4: Most Frequent Lexical Bundles by Most Frequent Grammatical Bundles in the Finland English Corpus

	1 HT HT HT HT	2 NNP NNP NNP NNP	3 NN IN DT NN	4 PRP VBP TO VB	5 IN DT JJN NN
1	thankyoujustinfromfinland thankyoujustinfromfinland thankyoujustinfromfinland thankyoujustinfromfinland	helsingin messukeskus helsinki exhibition	girl in the world	i want to go	at the same time
2	weather instaweather instaweatherpro sky	helsinki exhibition convention centre	game of the season	i have to go	of this amazing family
3	instaweather instaweatherpro sky outdoors	messukeskus helsinki exhibition convention	middle of the night	i want to be	for the first time
4	instaweatherpro sky outdoors nature	world wide live cam	side of the road	i want to see	in a long time
5	sky outdoors nature instagood	hard rock cafe helsinki	snow on the ground	i need to get	for a long time
6	incase incase incase incase	management events oy helsinki	son of a bitch	i have to do	on the bright side
7	outdoors nature instagood photooftheday	engholm husky design lodge	alert with no sleep	i have to wake	on the other hand
8	nature instagood photooftheday instamood	home sweet home kajaaninlinnantie	breakfast on the terrace	i want to do	for the whole day
9	amazing beautiful awesome sky	adecco finland oy tampere	cashier on the bus	i get to see	in the first place
10	rt rt rt rt	finnish science center heureka	city in the world	i go to sleep	after a long time
	6 PRP VBP DT NN	7 NNP IN CD NNS	8 IN DT NN IN	9 NNP NNP NNP URL	10 IN CD NNS URL
1	i have no idea	hel w 25 others	in the middle of	wide live cam (url)	in 17 days (url)
2	i know the feeling	hel w 30 others	for the rest of	aalto design factory (url)	in 20 minutes (url)
3	you have no idea	hel w 18 others	at the end of	aamukahvit bmw service (url)	in 30 minutes (url)
4	i get a follow	hel w 22 others	in a couple of	abc lappeenranta viipurinportti (url)	in 6 days (url)
5	i hate the fact	kinopalatsi w 2 others	in the world if	academic work helsinki (url)	in six languages (url)
6	i love the way	hel w 15 others	for a couple of	advertising agency satumaa (url)1gr	of two components (url)
7	i feel a bit	hel w 35 others	for the sake of	air balloon austria (url)	w 10 others (url)
8	i have a feeling	hel w 7 others	in the face with	aleksanteri ii patsas (url)	w 10 others (url)
9	i have no life	door w 2 others	on a date with	aleksis kiven katu (url)	w 10 others (url)
10	i love the rain	hel w 14 others	on the ground in	alta lufthavn alf (url)	w 10 others (url)

4–grams have a range of discourse functions they can potentially instantiate. Table 7.4 shows the ten most frequent sequences of grammatical word classes, and for each sequence, the ten most frequent lexeme sequences that correspond to the sequence of grammatical word classes.[5] The importance of the Twitter–specific word class hashtag, which can have various grammatical and discourse organization functions, is clear in the Finland English Corpus. In terms of topicality, the most frequent 4–hashtag sequences refer to the entertainer Justin Bieber, who gave a concert in Helsinki in 2013, or to topical content popular among the Finnish Twitter userbase (weather, outdoors, nature, sky, etc.).

4–grams consisting of proper noun sequences mostly reference places (*Helsinki Exhibition Convention Centre*, *Hard Rock Café Helsinki*), or companies (*Nokia oyj Nokia oyj*, *Adecco Finland Oy Tampere*).

Noun-preposition-determiner-noun sequences are a type that can reference a broad range of content; frequent types include various head nouns (*girl*, *game*, *city*) and prepositional phrases that serve as time or place modifiers (*in the world*, *of the season*, *of the night*, *on the terrace*, etc.).

The most frequent lexical item sequences in the word–class sequence personal pronoun-verb (non-3rd-person singular present)-*to*-verb (3rd-person singular present or base form), for the Finland data, consist exclusively of epistemic or attitudinal stance bundles with the first person pronoun. The word 4–grams most frequent for this sequence demonstrate again the primarily interactive nature of Twitter discourse for Finland–based users tweeting in English: reporting on their own desires, wishes, plans, and obligations.

Preposition-determiner-adjective-noun 4-grams typically have a time noun as the head element (*time*, *day*), with the other constituents providing specification. These sequences mainly have a discourse organizational function.

[5]Urls have been substituted with (url).

As is the case for the other grammatical n-grams containing verbal phrase elements, personal pronoun-verb (non-3rd-person-singular)-determiner-noun sequences are used in the Finland English Corpus by users mainly to express their various epistemic or attitudinal stances, mostly in the first person.

The sequence proper noun-preposition-cardinal number-plural noun is a sequence that is automatically added to user messages by an app for geo–tagging tweets. As it is not user–initiated in the same way as the other most common bundles, it will not be further considered.

The sequence preposition-determiner-noun-preposition represents a discourse organization bundle. Although discourse organization bundles are not among the most frequent types of lexical bundles, here we can see that this versatile sequence can express locational or spatial relations (*in the middle of*, *on the ground with*, *on a date with*), temporal relations (*for the rest of*, *in a couple of*, etc.), or various framing attributes (*for a couple of*, *for the sake of*).

The ninth and tenth most frequent part–of–speech 4–grams include urls; the proper noun-proper noun-proper noun-url sequence typically provides a web address for a specific organization, place, or company, and the preposition-cardinal number-noun-url sequence represents an automated addition to a user text generated by an app.

The most common grammatical bundles in the Finland English Corpus emphasize the interactive and self-referential discourse typical of Finland Twitter English.

7.3.2 Grammatical bundles in the Comparison English Corpus

TABLE 7.5: Most frequent grammatical bundles in the Comparison English Corpus

	Sequence	Frequency per thousand words		Sequence	Frequency per thousand words
1	NN IN DT NN	3.570	11	NN IN NNP NNP	1.353
2	NNP NNP NNP NNP	3.467	12	DT NN NN IN	1.218
3	IN DT NN NN	2.416	13	DT NN IN NN	1.217
4	IN DT NN IN	2.190	14	PRP VBP TO VB	1.186
5	IN DT JJ NN	2.187	15	DT NN IN NNP	1.144
6	DT NN IN DT	1.926	16	JJ NN IN DT	1.137
7	DT JJ NN IN	1.863	17	NN IN PRP$ NN	1.118
8	TO VB DT NN	1.459	18	DT NN NN NN	1.103
9	PRP VBP DT NN	1.388	19	URL NNP NNP NNP	1.090
10	DT JJ NN NN	1.364	20	NNS IN DT NN	1.081

The 20 most frequent grammatical bundles in the Comparison English Corpus are shown in Table 7.5. For this data the Twitter–specific categories hashtags and usernames play no role; the tag URL is present in only one type among the twenty most frequent types. The most frequent type is the sequence noun-preposition-determiner-noun, followed by four proper nouns in sequence. Common content–related and discourse organization bundles such as preposition-determiner-noun-noun or preposition-determiner-noun-preposition follow. Sequences based on nominal phrases, like determiner-noun-preposition-determiner or determiner-adjective-noun-preposition are also relatively common. The eighth, ninth, and tenth most frequent sequences of grammatical word classes are *to*-verb (base form)- determiner- noun, personal pronoun-verb (non-3rd-person-singular-present)-determiner-noun, and determiner- adjective-noun-noun. Frequent grammatical bundle types typically consist of noun phrase constituents; only three types in the twenty most frequent types include verbal constituents.

TABLE 7.6: Most Frequent Lexical Bundles by Most Frequent Grammatical Bundles in the Comparison English Corpus

	1 NN IN DT NN	2 NNP NNP NNP NNP	3 IN DT NN NN	4 IN DT NN IN	5 IN DT JJ NN
1	rest of the day	charles hotel the bar	on the m1 northbound	in the middle of	for the first time
2	end of the day	the charles hotel the	on the m1 southbound	at the end of	at the same time
3	preview of the week	organic herb herbal cs	on the way home	for the rest of	in a long time
4	quote of the day	the temple bar kultfabrik	in the parking lot	on the a5 from	for a long time
5	product of the week	postgazette ed bouchettes steelers	for a bike ride	on the a46 from	after a long day
6	word of the day	michigan ave chicago il	in the mail today	for a couple of	on the other hand
7	middle of the night	nfl video nfl gameday	at the grocery store	at the top of	in a good way
8	quote for the week	ap the pittsburgh steelers	on the plus side	on the phone with	in the first place
9	rest of the week	360 hd dvd player	like a pirate day	in the midst of	on the other side
10	room with a view	xbox 360 hd dvd	in the living room	by the end of	for a little while
	6 VBG TO NNP NNP	7 DT NN IN DT	8 DT JJ NN IN	9 TO VB DT NN	10 NN IN NNP NNP
1	going to whole foods	the rest of the	the first time in	to start the day	m1 towards lichfield south
2	going to philly standards	the end of the	a big fan of	to get some sleep	southbound between junctions j6
3	going to santa barbra	a bit of a	a little bit of	to take a nap	country for old men
4	going to st george	the middle of the	the first day of	to get some work	northbound between junctions j6
5	moving to new zealand	the name of the	the other side of	to take a shower	northbound between junctions j4
6	accding to uc davis	a preview of the	no such thing as	to see a movie	clockwise between junctions j16
7	according 2 jeska linden	the top of the	the last day of	to be a part	coffee from dunkin donuts
8	according to army times	a look at the	a beautiful day in	to hit the hay	eastbound between junctions j3
9	according to cnn oneal	the end of a	the first time since	to work this morning	northbound between junctions j
10	according to indus news	a lot of the	the other way around	to do some work	northbound between junctions j25

For the Comparison English data, grammatical category 4-grams and the most frequent corresponding word 4-grams reflect the topicality of the user messages, which are often internet- and media-based. The most common lexical sequences in the grammatical sequence noun-preposition-determiner-noun (Table 7.6) feature a time noun (*week*, *night*, *day*) in a prepositional phrase attached to a specifier head noun which qualifies the temporal aspect of the prepositional phrase (*rest, end, middle*; the nouns in this sequence otherwise represent a disparate class.

Sequences of four proper nouns in the Comparison English data are mainly names from media, entertainment, and sports.

Preposition-determiner-noun-noun sequences often refer to specific places (*on the m1 northbound, on the m1 southbound*, *in the postal mailbox*, *at the grocery store*, *in the living room*, *in the post office*) or events (*for a bike ride*, *for a conference call*).

The sequence preposition-determiner-noun-preposition includes mainly lexical bundles with discourse organization function such as *in the middle of*, *at the end of*, *for a couple of*, or *by the end of*. Types that can function as object complements of the copula (*on the phone with*) are also present.

The preposition-determiner-adjective-noun sequence has similar functions, based on the most frequent corresponding lexical bundles. The head noun is most commonly a time noun (*time*, *day*, *while*), and the preceding elements typically iteration or duration modifiers (*for a long*, *for a little*, *for the first*).

The verb (participle or gerund)-*to*-proper noun-proper noun sequence, in our data, is most commonly used to indicate movement, frequently as an object complement of a copular verb. Among the most frequent lexical sequences of this grammatical bundle type are also several that attribute information to specific sources.

The sequence determiner–noun–preposition–determiner is another discourse organization bundle. The most common lexical bundle types in this sequence consist of an article; a spatial, temporal, or quantity specifier (*rest*, *end*, *bit*, *middle*, *top*); the prepositional attributive sequence *of the*.

Determiner-adjective-noun-preposition sequences in the Comparison English data can function as discourse organizers (*a big fan of*, *a little bit of*, *the first day of, the last day of*); the indefinite article suggests a role as a subject complement of the copula (*a beautiful day in*, *a good day for*, *a great meeting with*).

The sequence *to*-verb-determiner-noun consists of the infinitive forms of common transitive verbs (*start*, *get*, *take*, *see*) with object complements such as the time nouns *day* or *morning* or other (often sleep–related) nouns such as *sleep, nap*, or *hay.*

Finally, the noun-preposition-proper noun-proper noun sequence in the Comparison English data consists mainly of tweets that, despite filtering, represent semi-automated traffic reports for British motorways; in some of these the CMU Twitter Tagger has misapplied the proper noun tag, possibly due to capitalization.

To summarize, the Finland English Corpus tweets show patterns of use of lexical and grammatical bundles that are much more indicative of a primarily interactive communicative orientation, as evidenced by higher use of personal pronouns and verbal forms. The propensity of Finland Twitter English users to utilize the Twitter–specific hashtag and to make relatively frequent use of urls in grammatical bundles demonstrates the extent to which this user group has taken advantage of specific communication technology modalities in order to satisfy functional demands.

The lexical and grammatical bundles most typical of the Comparison English Corpus, although also oriented primarily towards interactive communication, are more informational in nature, as evidenced by the higher frequency of specific information or discourse organization bundles such as those containing proper nouns or determiner–noun sequences.

Chapter 8

Concluding Remarks

Communication technology continues to evolve at a rapid pace, and as interaction in online and virtual space constitute an increasing proportion of linguistic contact for much of the world's population, the dynamics of online language use in our mediated worlds represent an important site for the study of English.

The compilation of relatively large corpora of online English as it is used in specific local contexts where it has not traditionally been a language of daily communication can serve as a useful tool for probing the dynamics of English varieties as they continue to diversify globally. For Finland, as a relatively prosperous society with relatively high levels of educational attainment, English plays a prominent role in the online daily life of many internet users, and this role is reflected in rates of use of English in popular social media such as Twitter.

In the second chapter of this study, communicative and situational parameters of media and communication channels and their effects on communicative functionality were discussed. It can be demonstrated that although CMC constitutes a diverse range of forms, some of the situational and communicative parameters of genres such as chat, IM, and Twitter correspond to those of spoken language, while others are more in line with

the parameters of other text-based communicative forms, whether online or in traditional media.

In the third chapter, a description was provided of the procedures used to collect a large geo–encoded Twitter database, clean the data of characters that can't be rendered, automatically detect the language of the user messages, automatically disambiguate a subset of the data for user gender, and automatically tag user message tokens with part–of–speech tags.[1] Conducting these steps primarily by using scripting in the statistical and programming environment *R* has the benefit of making the data accessible for a wide range of analytical approaches, including statistical measures and various forms of data visualization, including GIS. The Finland Twitter data collected in the project was augmented with some demographic data from the Finnish national statistical service.

Twitter is a communication platform with a global extent whose userbase shows a high degree of orientation to the developing social and communicative norms of global online culture, including extensive use of English and familiarity with global media figures. An overview of the language breakdown of Twitter messages geo–located to Finland (Chapter 4) shows that English plays an important role in Finland on the platform: English is used almost as often as Finnish in geo–located Finnish Tweets, and at a much higher rate than the second official language of the country, Swedish. High levels of exposure to English and knowledge of English are typical for younger persons in relatively prosperous societies who spend time online e.g. gaming or utilizing social media (Reinders and Wattana 2011). The language data from Finland Twitter support this interpretation.

[1] Although a geo–encoded Twitter database was only collected for Finland for this project, the procedures as described would allow the synchronous collection of Twitter data from multiple national contexts, making a detailed comparison of the features of online varieties of Twitter English possible. Such a database would make the typological comparison of online English varieties in local or national contexts possible. This would be a step towards creating a richer picture of the current status of global English as it is manifest in informal online language.

The demographics of use of Twitter suggest that younger people are overrepresented as users of the service (Owoputi et al. 2013, Eisenstein 2013) and that Twitter has been adopted more extensively in societies that are relatively prosperous (Mocanu et al. 2012). A consideration of the lexical features most characteristic of Finland Twitter English, as measured in comparison to Twitter English with no geographical specification by the odds ratio θ, shows that Finnish users tweet frequently about popular–culture entities that are associated with a youthful consumer demographic; they also utilize lexical items associated with informal internet language and the interactive negotiation of affective stance, such as emoticons, usernames, initialisms, and profanity (Section 5.2) at a much higher rate. These feature frequencies may index rates of adoption of technological innovation (and subsequent novel use of language features) as it is reflected economically: The use of emoticons, for example, correlates positively with GDP and educational level within Finland, according to data from Finland's official statistics bureau. As the Finland English Corpus consists solely of tweets tagged with geo–coordinates, and Twitter smartphone apps automatically send geo–coordinates, Finland Twitter English may be a variety whose relatively well–to–do young language users interact online using smartphones or similar devices and utilize these lexical features in order to mark identity.

The communicative function of any linguistic exchange is in part determined by situational and contextual parameters. Online communication in CMC genres such as Twitter represents a type of language communication whose configuration of parameters, like that of chat, IM, or spoken language, is more suited to personal stance expression and interactivity than the communication of informational content.

The predominantly informal nature of Twitter communication among Finland–based users is evidenced by aggregate rates of use of grammatical features, including non–standard features such as expressive lengthening. The discourse of Finland Twitter English shows high rates of use of personal pronouns, particularly first and second person singular forms.

High personal pronoun frequencies are typical for spoken language and CMC genres such as chat and IM, but the discourse of Finland Twitter English references the self and conversational interlocutors at a higher rate than does global Twitter English. The personal pronouns *I* and *you* and corresponding verbal forms, typically modals or verbs expressing stance propositions, are more frequent in Finland Twitter English than in global Twitter English. The discourse of the variety makes relatively less use of lexical and grammatical features associated with the communication of informational content, such as determiner–noun sequences, proper nouns, numbers, prepositional phrases, or complex verb forms.

Multidimensional quantitative techniques such as exploratory factor analysis or principal component analysis of aggregate feature frequencies have traditionally been used to identify the functional communicative dimensions of language genres or registers and show the extent to which they overlap or differ from one another. In this study, multidimensional techniques do provide some evidence for differentiation of communicative dimensions in English Twitter data, particularly for the Finland English data. The most important finding from this section, however, is the strong evidence from both the Finland English and the Comparison English corpora that those grammatical features most closely associated with the interactive style typical of Twitter tend to co-occur and can be located in vector space in close proximity to one another. This is especially true of grammatical features that are specific to Twitter: the hashtag, the username, and the retweet.

These features, along with features such as emoticons and expressive lengthening, are among the most characteristic of Finland Twitter English as a variety. They tend to signal affective or interactive content, but can also be used in unexpected ways to negotiate discourse–related concerns. The extensive use of these features in the variety is an indication of the multiple ways in which Finland Twitter English users construct identities and meanings in English at the interface of user interactivity and technological innovation.

Bibliography

Alis, C. and M. Lim (2013). "Spatio-Temporal Variation of Conversational Utterances on Twitter". In: *PLoS ONE* 8.10.

Altenberg, B. and M. Eeg-Olofsson (1990). "Phraseology in spoken English". In: *Theory and Practice in Corpus Linguistics*. Ed. by J. Aarts and W. Meijs. Amsterdam: Rodopi, pp. 1–26.

Androutsopoulos, J. (2003). "Online Gemeinschaften und Sprachvariation. Soziolinguistische Perspektiven auf Sprache im Internet". In: *Zeitschrift für Germanistische Linguistik* 31.2, pp. 173–197.

— (2006). "Introduction: Sociolinguistics and computer-mediated communication". In: *Journal of Sociolinguistics* 10.4, pp. 419–438.

Androutsopoulos, J. and V. Hinnenkamp (2001). "Code–Switching in der bilingualen Chat-Kommunikation: Ein explorativer Blick auf #hellas und #turks". In: *Chat-Kommunikation: Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation*. Ed. by M. Beißwenger. Stuttgart: Ibidem, pp. 367–402.

Anthony, L. (2014). *AntConc (Version 3.4.3W)*. Computer Software. Tokyo, Japan: Waseda University.

Argamon, S., M. Koppel, J. Fine, and A. Shimoni (2003). "Gender, genre, and writing style in formal written texts". In: *Text* 23, pp. 321–346.

Argamon, S., M. Koppel, J. Pennebaker, and J. Schler (2007). "Mining the blogosphere: Age, gender, and the varieties of self-expression". In: *First Monday* 12.9. URL: http://firstmonday.org/ojs/index.php/fm/article/view/2003.

Aronoff, M. (1985). "Orthography and linguistic theory". In: *Language* 61, pp. 28–72.

Ascher, D., P. Dubois, K. Hinsen, J. Hugunin, and T. Oliphant (2001). *Numerical Python*. Tech. rep. UCRL-MA-128569. Livermore, CA: Lawrence Livermore National Laboratory. URL: http://www.numpy.org.

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

— (2003). "Probabilistic Approaches to Morphology". In: *Probabilistic Linguistics*. Ed. by R. Bod, J. Hay, and S. Jannedy. Cambridge, MA: MIT Press, pp. 229–288.

— (2008). *Analyzing Linguistic Data: A practical introduction to statistics*. Cambridge, UK: Cambridge University Press.

— (2013). *LanguageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*. R package version 1.4.1. URL: http://CRAN.R-project.org/package=languageR.

Bamman, D., J. Eisenstein, and T. Schnoebelen (2014). "Gender Identity and Lexical Variation in Social Media". In: *Journal of Sociolinguistics* 18.2, pp. 135–160.

Baron, N. S. (2004). "See you online: Gender issues in college student use of instant messaging". In: *Journal of Language and Social Psychology* 23.4, pp. 397–423.

Baroni, M. and S. Evert (2009). "Statistical methods for corpus exploitation". In: *Corpus Linguistics: An International Handbook*. Ed. by A. Lüdeling and M. Kytö. Vol. 2. Berlin: Mouton de Gruyter, pp. 777–803.

Barzilay, R. and L. Lee (2004). "Catching the drift: Probabilistic content models, with applications to generation and summarization". In: *Proceedings of HLT-NAACL*. Boston, MA: Association for Computational Linguistics, pp. 113–120.

Batra, S. and D. Rao (2010). *Entity Based Sentiment Analysis on Twitter*. Tech. rep. Department of Computer Science, Stanford University. URL: http://nlp.stanford.edu/courses/cs224n/2010/reports/drao-sidbatra.pdf.

Bayraktar, M., B. Say, and V. Akman (1998). "An Analysis of English Punctuation". In: *International Journal of Corpus Linguistics* 3.1, pp. 33–57.

Becker, R. A., A. R. Wilks, R. Brownrigg, and T. P. Minka (2014). *Maps: Draw Geographical Maps*. R package version 2.3-9. URL: http://CRAN.R-project.org/package=maps.

Biber, D. (1985). "Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses". In: *Linguistics* 23, pp. 337–360.

— (1986). "Spoken and written textual dimensions in English: Resolving the contradictory findings". In: *Language* 62, pp. 384–414.

— (1987). "A textual comparison of British and American writing". In: *American Speech* 62, pp. 99–119.

— (1988). *Variation Across Speech and Writing*. Cambridge, UK: Cambridge University Press.

— (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge, UK: Cambridge University Press.

— (2006). *University Language. A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

— (2009). "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing". In: *International Journal of Corpus Linguistics* 14, pp. 275–311.

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

Biber, D., S. Conrad, and V. Cortes (2003). "Lexical bundles in speech and writing: An initial taxonomy". In: *Corpus linguistics by the Lune: A Festschrift for Geoffrey Leech*. Ed. by A. Wilson, P. Rayson, and T. McEnery. Frankfurt/Main: Peter Lang, pp. 71–92.

— (2004). "If you look at...: Lexical bundles in university teaching and textbooks". In: *Applied Linguistics* 25.3, pp. 371–405.

Biber, D., S. Conrad, and R. Reppen (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press.

Biber, D., S. Johannson, G. Leech, S. Conrad, and E. Finegan (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Biemann, C., F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski, and T. Zesch (2013). "Scalable construction of high-quality web corpora". In: *Journal of Language Technology and Computational Linguistics* 28.2, pp. 23–59.

Binongo, J. N. G. and M. W. A. Smith (1999). "The application of principal component analysis to stylometry". In: *Literary and Linguistic Computing* 14, pp. 445–465.

Bird, S., E. Loper, and E. Klein (2009). *Natural Language Processing with Python*. Newton, MA: O'Reilly.

Bivand, R., T. Keitt, and B. Rowlingson (2014). *Rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 0.9-1. URL: `http://CRAN.R-project.org/package=rgdal`.

Bivand, R., E. Pebesma, and V. Gomez-Rubio (2013). *Applied Spatial Data Analysis with R*. 2nd ed. New York: Springer.

Bod, R., J. Hay, and S. Jannedy (2003). "Introduction". In: *Probabilistic Linguistics*. Ed. by R. Bod, J. Hay, and S. Jannedy. Cambridge, MA: MIT Press, pp. 1–10.

Bollen, J., H. Mao, and X. Zeng (2011). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1, pp. 1–8. URL: `http://www.sciencedirect.com/science/article/pii/S187775031100007X`.

Boyd, D. (2014). *Bibliography of research on twitter & microblogging*. URL: www.danah.org/researchBibs/twitter.php.

Boyd, D., S. Golder, and G. Lotan (2010). "Tweet tweet retweet: Conversational aspects of retweeting on Twitter". In: *Proceedings of HICSS-43*. Kauai, HI: IEEE Computer Society. URL: http://www.danah.org/papers/TweetTweetRetweet.pdf.

Burger, J., J. Henderson, G. Kim, and G. Zarrella (2011). "Discriminating gender on Twitter". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP. Edinburgh, UK, pp. 1301–1309. URL: http://aclweb.org/anthology//D/D11/D11-1120.pdf.

Burrows, J. (2002a). " 'Delta' : A measure of stylistic difference and a guide to likely authorship". In: *Literary and Linguistic Computing* 17, pp. 267–287.

— (2002b). "The Englishing of Juvenal: Computational stylistics and translated texts". In: *Style* 36, pp. 677–699.

— (2003). "Questions of authorship: Attribution and beyond". In: *Computers and the Humanities* 37, pp. 5–32.

Cedergren, H. and D. Sankoff (1974). "Variable rules: Performance as a statistical reflection of competence". In: *Language* 50.2, pp. 333–355.

Chambers, J. K. and P. Trudgill (1998). *Dialectology*. Cambridge, UK: Cambridge University Press.

Chen, Y. and P. Baker (2010). "Lexical bundles in L1 and L2 academic writing". In: *Language Learning & Technology* 14.2, pp. 30–49.

Cheshire, J. (2002). "Sex and gender in variationist research". In: *Handbook of Language Variation and Change*. Ed. by J. K. Chambers, P. Trudgill, and N. Schilling-Estes. Oxford, UK: Blackwell, pp. 423–443.

Chesterman, A. (1991). *On Definiteness: A Study with Special Reference to English and Finnish*. Cambridge, UK: Cambridge University Press.

Chovanec, J. (2014). *Pragmatics of Tense and Time in News: From Canonical Headlines to Online News Texts*. Amsterdam: John Benjamins.

Coates, J., ed. (1998). *Language and Gender: A Reader*. Oxford, UK: Blackwell.

Conrad, S. (1999). "Lexical bundles in conversation and academic prose". In: *Out of Corpora: Studies in Honor of Stig Johansson*. Ed. by H. Hasselgard and S. Oksefjell. Amsterdam: Rodopi, pp. 181–189.

Crystal, D. (2001). *English as a Global Language*. 2nd ed. Cambridge, UK: Cambridge University Press.

— (2006). *Language and the Internet*. 2nd ed. Cambridge, UK: Cambridge University Press.

— (2011). *Internet Linguistics*. London: Routledge.

Dahl, Ö. (2000). "The grammar of future time reference in European languages". In: *Tense and Aspect in the Languages of Europe*. Ed. by Ö. Dahl. Berlin: Mouton de Gruyter, pp. 309–328.

Damerau, F. J. (1975). "The use of function word frequencies as indicators of style". In: *Computers and the Humanities* 9.6, pp. 271–280.

Davies, M. and D. Gardner (2010). *A Frequency Dictionary of Contemporary American English*. London/New York: Routledge.

Derczynski, L., A. Ritter, S. Clark, and K. Bontcheva (2013). "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria: Incoma, pp. 198–206. URL: http://aclweb.org/anthology/R13-1026.

Dewaele, J.-M. (2004). "The emotional force of swearwords and taboo words in the speech of multilinguals". In: *Journal of Multilingual and Multicultural Development* 25.2-3, pp. 204–222. URL: http://www.ingentaconnect.com/content/mm/jmmd/2004/00000025/F0020002/art00008.

Dunning, T. (1993). "Accurate methods for the statistics of surprise and coincidence". In: *Computational Linguistics* 19.1, pp. 61–74.

Dürscheid, C. (2005). "Medien, Kommunikationsformen, kommunikative Gattungen". In: *Linguistik Online* 22, pp. 3–16.

Eder, M., M. Kestemont, and J. Rybicki (2013). "Stylometry with R: A suite of tools". In: Lincoln: University of Nebraska-Lincoln, pp. 487–489.

Eilola, T. and J. Havelka (2010). "Affective norms for 210 British English and Finnish nouns". In: *Behavior Research Methods* 42.1, pp. 134–140.

Eisenstein, J. (2013). "What to do about bad language on the internet". In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Atlanta, GA. URL: http://www.cc.gatech.edu/~jeisenst/papers/naacl2013-badlanguage.pdf.

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

Ellegård, A. (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.

Ellis, N. (2002). "Frequency effects in language processing". In: *Studies in Second Language Acquisition* 24, pp. 143–188.

European Commission (2011). "Flash Eurobarometer 313: User language preference online". In: URL: http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

Evert, S. (2004). *Association Measures*. URL: http://www.collocations.de/AM/.

— (2008). "Corpora and collocations". In: *Corpus Linguistics: an International Handbook*. Ed. by A. Lüdeling and M. Kytö. Vol. 2. Berlin: Mouton de Gruyter, pp. 1212–1248.

Evert, S. and M. Baroni (2007). "ZipfR: Word frequency distributions in R". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Prague, pp. 29–32. URL: http://zipfr.r-forge.r-project.org/).

— (2012). *ZipfR: Statistical models for word frequency distributions*. R package version 0.6-6. URL: http://CRAN.R-project.org/package=zipfR.

Feinerer, I., K. Hornik, and D. Meyer (2008). "Text mining infrastructure in R". In: *Journal of Statistical Software* 25.5, pp. 1–54. URL: http://www.jstatsoft.org/v25/i05/.

Fengxiang, F. (2007). "A corpus based quantitative study on the change of TTR, word length and sentence length of the English language". In: *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday*. Ed. by Peter Grzybek and Reinhard Köhler. Berlin/New York: Mouton de Gruyter, pp. 123–130.

Fowler, H. W. and F. G. Fowler (1962 [1931]). *The King's English*. 3rd ed. Oxford: Clarendon.

French, N. R. and J. C. Steinberg (1947). "Factors governing the intelligibility of speech sounds". In: *The Journal of the Acoustical Society of America* 19.1, pp. 90–119.

Fry, D. (1955). "Duration and intensity as physical correlates of linguistic stress". In: *Journal of the Acoustic Society of America* 27, pp. 765–768.

Gentry, J. (2012). *TwitteR: R based Twitter client*. R package version 0.99.19. URL: http://CRAN.R-project.org/package=twitteR.

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, Pittsburgh, PA. URL: http://www.ark.cs.cmu.edu/TweetNLP/annot_guidelines.pdf.

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". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA: Association for Computational Linguistics, pp. 42–47.

Gleason, H. (1965). *Linguistics and English Grammar*. New York: Holt, Rinehart & Winston.

Gonçalves, B. and D. Sánchez (2014). "Crowdsourcing dialect characterization through Twitter". In: *PLoS ONE* 9.11. DOI: 10.1371/journal.pone.0112074.

Harris, Z. (1962). *String Analysis of Sentence Structure*. Den Haag: Mouton.

— (1968). *Mathematical Structures of Language*. New York: Interscience.

— (1982). "Discourse and sublanguage". In: *Sublanguage: Studies of Language in Restricted Semantic Domains*. Ed. by R. Kittredge and J. Lehrberger. Berlin; New York: De Gruyter, pp. 231–236.

Hecht, B., L. Hong, B. Suh, and E. H. Chi (2011). "Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, pp. 237–246.

Hentschel, E. (1998). "Communication on IRC". In: *Linguistik Online* 1. URL: http://www.linguistik-online.de/inhalt1_98.html.

Herring, S. (1998). "Le style du courrier électronique: Variabilité et changement". In: *Terminogramme* 84-85, pp. 9–16.

— (1999). "Interactional coherence in CMC". In: *Journal of Computer-Mediated Communication* 4.4. URL: http://jcmc.indiana.edu/vol4/issue4/herring.html.

— (2004). "Computer-mediated discourse analysis: An approach to researching online behavior". In: *Designing for Virtual Communities in the Service of Learning*. Ed. by S. Barab, R. Kling, and J. Gray. New York: Cambridge University Press, pp. 338–376.

Herring, S. (2007). "A faceted classification scheme for computer-mediated discourse". In: *Language@Internet* 4.1. URL: http://www.languageatinternet.org/articles/2007/761.

— (2013). "Discourse in Web 2.0: Familiar, reconfigured, and emergent". In: *Discourse 2.0: Language and New Media*. Ed. by D. Tannen and A. M. Trester. Washington, DC: Georgetown University Press. URL: http://books.google.com/books?id=lhvgId6o-p4C.

Herring, S. and J. Paolillo (2006). "Gender and genre variation in weblogs". In: *Journal of Sociolinguistics* 10.4, pp. 439–459.

Hill, T. (1998). "The first digit phenomenon". In: *American Scientist* 86.4, pp. 358–363.

Hiltz, S. R. and M. Turoff (1978). *The Network Nation: Human Communication via Computer*. Reading, MA: Addison-Wesley.

Holmes, J. (1993). "Women's talk: The question of sociolinguistic universals". In: *Australian Journal of Communications* 20, pp. 125–149.

Honeycutt, C. and S. Herring (2009). "Beyond microblogging: Conversation and collaboration via Twitter". In: *Proceedings of the 42nd Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE, pp. 1–10.

Hoover, D. (2001). "Statistical stylistics and authorship attribution: An empirical investigation". In: *Literary and Linguistic Computing* 16, pp. 421–444.

— (2002). "Frequent word sequences and statistical stylistics". In: *Literary and Linguistic Computing* 17, pp. 157–180.

— (2003a). "Frequent collocations and authorial style". In: *Literary and Linguistic Computing* 18, pp. 261–286.

— (2003b). "Multivariate analysis and the study of style variation". In: *Literary and Linguistic Computing* 18, pp. 341–360.

Hoover, D. (2004). "Testing Burrows's Delta". In: *Literary and Linguistic Computing* 19, pp. 453–475.

Hotopf, N. (1980). "Slips of the pen". In: *Cognitive Processes in Spelling*. Ed. by U. Frith. London: Academic Press, pp. 287–307.

Hundt, M. (2008). "Text corpora". In: *Corpus Linguistics: An International Handbook*. Ed. by A. Lüdeling and M. Kytö. Vol. 1. Berlin: De Gruyter, pp. 168–187.

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

Hyland, K. (2008). "As can be seen: Lexical bundles and disciplinary variation". In: *English for Specific Purposes* 27, pp. 4–21.

Jiang, L., M. Yu, M. Zhou, X. Liu, and T. Zhao (2011). "Target-dependent Twitter sentiment classification". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. HLT '11. Stroudsburg, PA: Association for Computational Linguistics, pp. 151–160. ISBN: 978-1-932432-87-9. URL: http://dl.acm.org/citation.cfm?id=2002472.2002492.

Jolliffe, I. (2002). *Principal Component Analysis*. 2nd ed. New York: Springer.

Jones, B. (1996). *What's the Point? A (Computational) Theory of Punctuation*. PhD thesis. University of Edinburgh.

Jurafsky, D. (2003). "Probabilistic modeling in psycholinguistics: Linguistic comprehension and production". In: *Probabilistic Linguistics*. Ed. by R. Bod, J. Hay, and S. Jannedy. Cambridge, MA: MIT Press, pp. 39–95.

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

Kahle, D. and H. Wickham (2013). *Ggmap: A Package for Spatial Visualization with Google Maps and OpenStreetMap*. R package version 2.3. URL: http://CRAN.R-project.org/package=ggmap.

Kapidzic, S. and S. Herring (2011). "Gender, communication, and self-presentation in teen chatrooms revisited: Have patterns changed?" In: *Journal of Computer-Mediated Communication* 17.1, pp. 39–59.

Karttunen, L. (2006). "Numbers and Finnish numerals". In: *SKY Journal of Linguistics* 19, pp. 407–421.

Kawakami, M. (2008). "顔文字が表す感情と強調に関するデータベース [The database of 31 Japanese emoticon with their emotions and emphases]". In: *Human Science Research Bulletin of Osaka Shoin Women's University* 7. URL: http://ci.nii.ac.jp/naid/110006629179/en.

Kelly, R. (2009). *Pear analytics Twitter study*. www.pearanalytics.com. URL: https://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf.

Klatt, D. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence". In: *Journal of the Acoustic Society of America* 59, pp. 1208–1221.

Kline, P. (2014). *An Easy Guide to Factor Analysis*. Oxford: Routledge.

Kochen, M. (1978). "Long-term implications of electronic information exchanges for information science". In: *Bulletin of the American Society for Information Science* 4.5, pp. 22–23.

Kretzschmar, W. A. K. (2009). *The Linguistics of Speech*. Cambridge, UK: Cambridge University Press.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

— (1990). "The intersection of sex and social class in the course of linguistic change". In: *Language Variation and Change* 2, pp. 205–254.

— (2001). *Principles of Linguistic Change, Vol. II: Social Factors*. Oxford: Blackwell.

Lakoff, R. (1973). "Language and woman's place". In: *Language in Society* 2, pp. 45–80.

Lang, D. T. (2012). *XML: Tools for Parsing and Generating XML within R and S-Plus*. R package version 3.9-4.1. URL: http://CRAN.R-project.org/package=xml.

Larsson, L. G. (1983). "Studien zum Partitivgebrauch in den ostseefinnischen Sprachen." In: *Acta Universitatis Upsaliensis, Studia Uralica et Altaica Upsaliensia* 15.

Leech, G., R. Garside, and M. Bryant. "CLAWS4: The tagging of the British National Corpus". In: *Proceedings of the 15th International Conference on Computational Linguistics*. Kyoto, Japan: ACM, pp. 622–628.

Leech, G., P. Rayson, and A. Wilson (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow, UK: Longman.

Leiss, E. (2007). "Covert patterns of definiteness/indefiniteness and aspectuality in Old Icelandic, Gothic, and Old High German." In: *Nominal Determination: Typology, Context, Constraints, and Historical Emergence.* Ed. by E. Stark, E. Leiss, and W. Abraham. Amsterdam: John Benjamins, pp. 73–102.

Ling, R. (2005). "An analysis of SMS use by a random sample of Norwegians". In: *Mobile Communications: Renegotiation of the Social Sphere*. Ed. by R. Ling and P. Pedersen. London: Springer, pp. 335–349.

Lui, M. and T. Baldwin (2012). "Langid.py: An off-the-shelf language identification tool". In: *50th Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics. Stroudsburg, PA, pp. 25–30.

Magdy, A., T. M. Ghanem, M. Musleh, and M. F. Mokbel (2014). "Exploiting geo-tagged tweets to understand localized language diversity". In: *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*. GeoRich'14. Snowbird, UT, USA: ACM, 2:1–6. URL: http://doi.acm.org/10.1145/2619112.2619114.

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

Markov, A. A. (2006 [1913]). "An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains". In: *Science in Context* 19. transl. by G. Custance and D. Link, pp. 591–600.

Markwardt, A. (1942). *Introduction to the English Language*. New York: Oxford University Press.

Maynor, N. (1994). "The language of electronic mail: Written speech?" In: *Centennial Usage Studies*. Ed. by M. Montgomery and G. D. Little. Tuscaloosa: University of Alabama Press, pp. 48–54.

Mays E., F. Damerau and R. Mercer (1991). "Context based spelling correction". In: *Information Processing and Management* 27.5, pp. 495–505.

McEnery, T. (2005). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. New York: Routledge.

McEnery, T., R. Xiao, and Y. Tono (2006). *Corpus-Based Language Studies*. 2006. London: Routledge.

McIlroy, D. (2014). *Mapproj: Map Projections*. R package version 1.2-2. Packaged for R by R. Brownrigg and T. Minka and transitioned to Plan 9 codebase by R. Bivan. URL: `http://CRAN.R-project.org/package=mapproj`.

Mehl, M. and J. Pennebaker (2003). "The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations". In: *Journal of Personality and Social Psychology* 84, pp. 857–870.

Mitton, R. (1987). "Spelling checkers, spelling correctors, and the misspellings of poor spellers". In: *Information Processing and Management* 23.5, pp. 495–505.

Mocanu, D., A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani (2013). "The Twitter of Babel: Mapping world languages through microblogging platforms". In: *PLoS ONE* 8.4.

Mosteller, F. and D. Wallace (1963). "Inference in an authorship problem". In: *Journal of the American Statistical Association* 58, pp. 275–309.

Murthy, D. (2011). "Twitter: Microphone for the masses?" In: *Media, Culture & Society* 33.5, pp. 779–789.

— (2013). *Twitter: Social Communication in the Twitter Age*. Cambridge, UK: Polity.

Nation, P. (2006). "Vocabulary: Second language". In: *Encyclopedia of Language and Linguistics*. Ed. by K. Brown. 2nd Edition. Oxford: Elsevier, pp. 448–454.

Nelson, M. (2010). "Building a written corpus: What are the basics?" In: *The Routledge Handbook of Corpus Linguistics*. Ed. by A. O. Keeffe and M. McCarthy. Oxford, UK: Routledge, pp. 53–65.

Newman, M., C. Groom, L. Handelman, and J. Pennebaker (2008). "Gender differences in language use: An analysis of 14,000 text samples". In: *Discourse Processes* 45, pp. 211–236.

Noswearing.com (2015). *List of Bad Words*. URL: `http://www.noswearing.com/dictionary`.

Nunberg, G. (1990). *The Linguistics of Punctuation*. Palo Alto: CSLI.

Official Statistics of Finland (2013a). *Polytechnic education*. Appendix table 1. Completed polytechnic degrees by Region of Education in 2012. Helsinki. URL: `http://www.stat.fi/til/akop/2012/akop_2012_2013-04-10_tau_001_en.html`.

— (2013b). *Population*. Area, population and GDP by region. Helsinki. URL: `http://www.tilastokeskus.fi/tup/suoluk/suoluk_vaesto_en.html`.

— (2013c). *University education*. University students 2012, Appendix table 2. Completed university degrees by region of education in 2012; University degrees 2012, Appendix table 2. New university students and total number of students in universities by region in 2012. Helsinki. URL: `http://www.stat.fi/til/yop/2012/01/yop_2012_01_2013-04-23_tau_002_en.html`.

Official Statistics of Finland (2013d). *Vocational education*. Completers of curriculum-based vocational qualifications in the calendar year 2012, Appendix table 1. New students in preparatory education for a skills examination, students and completers of a qualification by region in 2012. Helsinki. URL: http://www.stat.fi/til/aop/2012/03/aop_2012_03_2013-11-06_tau_001_en.html.

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". In: *Proceedings of NAACL-HLT*. Vol. 2013. NAACL-HLT, pp. 380–390.

Paice, C. D. (1990). "Another stemmer". In: *ACM SIGIR Forum* 24.3, pp. 56–61.

Pang, B. and L. Lee (2008). "Opinion mining and sentiment analysis". In: *Foundations and Trends in Information Retrieval* 2.1-2, pp. 1–135.

Paolillo, J. C. (2001). "Language variation on Internet Relay Chat: A social network approach". In: *Journal of Sociolinguistics* 5.2, pp. 180–213.

Paul, H. (1886). *Principien der Sprachgeschichte*. 2., Erweiterte Auflage. Halle: Max Niemeyer.

Pebesma, E. and R. Bivand (2005). "Classes and methods for spatial data in R". In: *R News* 5.2. URL: http://cran.r-project.org/doc/Rnews.

Pennebaker, J., M. Francis, and R. Booth (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Mahway, NJ: Lawrence Erlbaum Associates.

— (2007). *Linguistic Inquiry and Word Count: LIWC2007*. Austin, TX: LIWC.net.

Piantadosi, S. T., H. Tily, and E. Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proc. Natl. Acad. Sci. USA* 108.9, pp. 3526–3529. eprint: http://www.pnas.org/content/108/9/3526.full.pdf+html. URL: http://www.pnas.org/content/108/9/3526.abstract.

Pike, K. (1947). *Phonemics: A Technique for Reducing Languages to Writing*. Ann Arbor: University of Michigan Press.

Porter, M. (1980). "An algorithm for suffix stripping." In: *Program* 14.3, pp. 130–137.

Ptaszynski, M. (2006). “萌える言語—インターネット掲示板の上の日本語会話における感情表現の構造と記号論的機能の分析・「２ちゃんねる」電子掲示板を例として・[Boisterous language - Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum 2channel]”. MA thesis. Adam Mickiewicz University, Poznan, Poland.

Ptaszynski, M., R. Rzepka, K. Araki, and Y. Momouchi (2011). “Research on emoticons: Review of the field and proposal of research framework”. In: *Proceedings of The Seventeenth Annual Meeting of The Association for Natural Language Processing (NLP-2011)*. Organized Session on Un-Natural Language Processing. Association for Natural Language Processing. Toyohashi, Japan, pp. 1159–1162.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: `http://www.R-project.org`.

Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta (2010). “Classifying latent user attributes in Twitter”. In: *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. ACM, pp. 37–44.

Reinders, H. and S. Wattana (2011). “Learn English or die : The effects of digital games on interaction and willingness to communicate in a foreign language”. In: *Digital Culture and Education* 3.1, pp. 3–29.

Revelle, W. (2014). *Psych: Procedures for Personality and Psychological Research*. R package version 1.4.8. URL: `http://CRAN.R-project.org/package=psych`.

Rice, R. and G. Love (1987). “Electronic emotion: Socioemotional content in a computer-mediated communication network”. In: *Communication Research* 14.1, pp. 85–108.

Ritter, A., C. Cherry, and B. Dolan (2010). “Unsupervised modeling of Twitter conversations”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT ’10. Los An-

geles, California: Association for Computational Linguistics, pp. 172–180. URL: http://dl.acm.org/citation.cfm?id=1857999.1858019.

Ritter, A., S. Clark, M. Etzioni, and O. Etzioni (2011). "Named entity recognition in tweets: An experimental study". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: ACM, pp. 1524–1534. URL: http://dl.acm.org/citation.cfm?id=2145432.2145595.

Roesslein, J. (2013). *Tweepy*. Python programming language module. URL: https://github.com/tweepy/tweepy.

Rossum, G. van and F. L. Drake, eds. (2006). *Python Reference Manual*. URL: http://docs.python.org/ref/ref.html.

Sankoff D., S. Tagliamonte and E. Smith (2005). *GoldVarb X: A Variable Rule Application for Macintosh and Windows*. URL: http://individual.utoronto.ca/tagliamonte/goldvarb.html.

Santorini, B. (1995). *Part-of-speech Tagging Guidelines for the Penn Treebank Project*. Tech. rep. Department of Computer and Information, University of Pennsylvania. URL: https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.ps.

Sapir, E. (1921). *Language*. New York: Harcourt Brace and World.

Sass, E. (2011). *Average Twitter User Sends Half a Tweet Per Day*. Mediapost.com. URL: http://www.mediapost.com/publications/article/160712/average-twitter-user-sends-half-a-tweet-per-day.html.

Say, B. and V. Akman (1997). "Current approaches to punctuation in compuational linguistics". In: *Computers and the Humanities* 30, pp. 457–469.

— (1998). *Dashes as Typographical Cues for the Information Structure*. Tech. rep. Center for Research in Cognitive Science and Department of Psychology, National Chung Cheng University, Chiayi, Taiwan, pp. 209–223. URL: http://cogprints.org/209/.

Schnoebelen, T. (2011). *Emotions are Relational: Positioning and the Use of Affective Linguistic Resources*. PhD Thesis. Stanford University.

— (2012). "Do You Smile with Your Nose? Stylistic Variation in Twitter Emoticons". In: *University of Pennsylvania Working Papers in Linguistics* 18.2, pp. 115–125. URL: http://repository.upenn.edu/pwpl/vol18/iss2/14.

Scott, M. (1996). *WordSmith Tools*. Oxford: Oxford University Press.

— (2012). *WordSmith Tools, Version 6*. Computer software. URL: http://lexically.net.

Sebba, M. (2007). *Spelling and Society: The Culture and Politics of Orthography around the World*. Cambridge, UK: Cambridge University Press.

— (2010). "Discourses in transit". In: *Semiotic Landscapes: Language, Image, Space*. Ed. by A. Jaworski and C. Thurlow. London: Continuum, pp. 59–76.

Seshagiri, A. (2014). *The languages of Twitter users*. The New York Times: Bits. URL: http://bits.blogs.nytimes.com/2014/03/09/the-languages-of-twitter-users/?_r=0.

Shannon, C. (1948). "A mathematical theory of communication". In: *Bell System Technical Journal* 27, pp. 379–423; 623–656.

Sigurd, B., M. Eeg-Olofsson, and J. Van Weijer (2004). "Word length, sentence length and frequency – Zipf revisited". In: *Studia Linguistica* 58, pp. 37–52.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sonderegger, S. (1983). "Leistung und Aufgabe der Dialektologie im Rahmen der Sprachgeschichtsschreibung des Deutschen". In: *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung, 2. Halbband*. Ed. by Werner Besch. Berlin/New York: de Gruyter, pp. 1526–1558.

Squires, L. (2010). "Enregistering internet language". In: *Language in Society* 39, pp. 457 –492.

Squires, L. (2012). "Whos punctuating what? Sociolinguistic variation in instant messaging". In: *Orthography as Social Action: Scripts, Spelling, Identity and Power*. Ed. by A. Jaffe. Boston/Berlin, pp. 289–324.

Statistic Brain (2014). *Twitter Statistics*. URL: http://www.statisticbrain.com/twitter-statistics/.

Stolcke, A., N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer (2000). "Dialogue act modeling for automatic tagging and recognition of conversational speech". In: *Computational Linguistics* 26.3, pp. 339–373.

Sweet, H., ed. (1885). *The Oldest English Texts, with Introductions and a Glossary*. London: N. Trübner.

Tagliamonte, S. and D. Denis (2008). "Linguistic ruin? Lol! Instant messaging and teen language". In: *American Speech* 83.1, pp. 3–34. URL: http://www.jstor.org/stable/40281250.

Takagi, H. (1999). *Japanese Smileys (Emoticons)*. Hiroette.com. URL: http://club.pep.ne.jp/~hiroette/en/facemarks/body.html.

Toutanova, K., D. Klein, C. Manning, and Y. Singert (2003). "Feature-rich part-of-speech tagging with a cyclic dependency network". In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics)*. URL: http://aclweb.org/anthology/N03-1033.

Trudgill, P. (1972). "Sex, covert prestige and linguistic change in the urban British English of East Anglia." In: *Language in Society* 1, pp. 179–195.

— (1974). *The Social Differentiation of English in Norwich*. Cambridge, UK: Cambridge University Press.

Truss, L. (2003). *Eats, Shoots & Leaves: The Zero Tolerance Approach to Punctuation*. London: Profile.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe (2010). "Predicting elections with Twitter: What 140 characters reveal about political sentiment". In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178–185.

Turney, P. D. (2002). "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424. URL: http://arxiv.org/abs/cs.LG/0212032.

Twitter (2015). *Certification Requirements*. URL: https://biz.twitter.com/partners/certification-requirements.

Vandergriff, I. (2014). "A pragmatic investigation of emoticon use in nonnative/native speaker text chat". In: *Language@Internet* 11.4. URL: http://www.languageatinternet.org/articles/2014/vandergriff.

Walkowska, J. (2009). "Gathering and analysis of a corpus of Polish SMS dialogues". In: *Challenging Problems of Science. Computer Science: Recent Advances in Intelligent Information Systems*. Ed. by M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, and K. Trojanowski, pp. 145–157.

Wexler, P. (1976). "On the non–lexical expression of determinedness (with special reference to Russian and Finnish)". In: *Studia Linguistica* 30, pp. 34–67.

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

— (2012). *Stringr: Make It Easier to Work with Strings*. R package version 0.6.2. URL: http://CRAN.R-project.org/package=stringr.

Wiebe, J., T. Wilson, and C. Cardie (2005). "Annotating expressions of opinions and emotions in language". In: *Language Resources and Evaluation* 39.2-3, pp. 165–210.

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

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

Wu, S., J. Hofman, W. Mason, and D. Watts (2011). "Who says what to whom on Twitter". In: *Proceedings of the 20th World Wide Web Conference (WWW '11)*. Hyderabad, India: ACM, pp. 705–714.

Xu, W., A. Ritter, and R. Grishman (2013). "Gathering and generating paraphrases from Twitter with application to normalization". In: *Proceedings of ACL 2013 Workshop on Building and Using Comparable Corpora*. Sofia, Bulgaria: Association of Computational Linguistics. URL: https://www.cs.nyu.edu/~xuwei/publications/ACL2013_BUCC.pdf.

Yang, J. and J. Leskovec (2011). "Patterns of temporal variation in online media". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. New York, NY, USA: ACM, pp. 177–186. URL: http://doi.acm.org/10.1145/1935826.1935863.

Yleisraadio (Finnish State Broadcasting Service) (2013). *Census Reveals Number of Finnish Tweeters*. URL: http://yle.fi/uutiset/census_reveals_number_of_finnish_tweeters/6503554?origin=rss.

Yule, U. (1939). "On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship". In: *Biometrika* 30, pp. 363–390.

Zappavigna, M. (2011). "Ambient affiliation: A linguistic perspective on Twitter". In: *New Media and Society* 13.5, pp. 788–806.

Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies*. Amsterdam: Elsevier.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

— (1949). *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*. New York: Hafner.

Appendices

Appendix A

Code for Data Collection and Analysis

A.1 Code in Python

A.1.1 Finland Twitter Corpus Compilation Code

```
# -*- coding: utf-8 -*-
import sys
import tweepy
import unicodedata
import codecs

consumer_key=""   #obtained from Twitter API
consumer_secret="" #obtained from Twitter API
access_key = ""  #obtained from Twitter API
access_secret = ""   #obtained from Twitter API

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)

class CustomStreamListener(tweepy.StreamListener):
```

```
    def on_status(self, status):
        try:
            print "%s\n%s\n%s\n%s\n%s\n%s\n\n" % (status.text,
                                        status.author.screen_name,
                                        status.created_at,
                                        status.source,status.coordinates,status.place)
            with codecs.open('test1.txt', 'ab','utf-8') as f:

                        newline=' NEWLINE'
                        linebreak='\r\n'
                        mylist=(status.text, status.author.screen_name, status.created_at, status
                            .source,status.coordinates,status.place)
                        f.write (("%s\t%s\t%s\t%s\t%s\t%s\t") % (mylist)+linebreak)
        except Exception, e:
            print >> sys.stderr, 'Encountered Exception:', e
            pass

    def on_error(self, status_code):
        print >> sys.stderr, 'Encountered error with status code:', status_code
        return True # Don't kill the stream

    def on_timeout(self):
        print >> sys.stderr, 'Timeout...'
        return True # Don't kill the stream

sapi = tweepy.streaming.Stream(auth, CustomStreamListener())
sapi.filter(locations=[21,60,29,70])
```

A.1.2 Language Identification Code

```
import langid
x=combined.txt    #textfile with one tweet user message per line
lang=[]
for w in y:  lang.append(langid.classify(w))
zipped=zip(y,lang)  #merges tweet and lang ID on lines as duples
z=open('b1linesat.txt','a')
```

```
for item in zipped:
  z.write("%s,%s\n"%item) \#writes as normal textfile
z.close()
```

A.2 Code in *R*

```
library(ggplot2)
library(rgdal)
library(sp)
library(mapproj)
library(maps)
library(mapdata)
library(maptools)  #for shapefiles
library(scales)
load(<path to gadm shapefile for Finland>)
fin.adm2.spdf <- get("gadm")
fin.adm2.df <- fortify(fin.adm2.spdf, region = "NAME_2")
bb<-readLines(<path to Finland Twitter data>)
bbb<-strsplit(bb,"\t",perl=T,useBytes=F)
b1<-as.data.frame(do.call("rbind", bbb),fill=T,stringsAsFactors=F,row.names=F)
df<-b1[,1:5]
aa<-readLines(<path to filtered Finland tweet data>)
aa<-head(aa,-1) #useless extra row
aaa<-strsplit(aa,",\\(\\'")
aaa1<-sapply(aaa,"[",1)

qqa<-grep("^\".*\"$",aaa1)
aaa1.1<-gsub("^.","",aaa1[qqa])
aaa1.2<-gsub(".$","",aaa1.1)
aaa1.3<-gsub("\"\"","\"",aaa1.2)
aaa1[qqa]<-aaa1.3

aaa2<-sapply(aaa,"[",2)
df$V1<-aaa1
df$V6<-aaa2
aa1<-gsub(".*,\\('","",aa) #a list with all characters after lang name.
```

```
aa2<-substr(aa1,1,2) #lang
aa3<-substr(aa1,6,10) #probability of langid
uu<-which(aa3=="1.0)")
aa3[uu]<-"1.000"
aa4<-as.numeric(aa3)
df$V7<-aa2
df$V8<-aa4
b8<-grep("2013",df$V3)              #to remove some contaminated rows in the data
df2<-df[b8,]                        #"
b9<-which(df2$V5=="None")           #"
df3<-df2[-b9,]                      #"
b10<-grep("coordinates",df3$V5)     #"
df4<-df3[b10,]                      #"
ll<-df4$V5
ll1<-sub("\\{u'type': u'Point', u'coordinates': \\[","",ll)
ll2<-sub("\\]\\}","",ll1)
ll3<-strsplit(ll2,",")
library(stringr)
lat<-as.numeric(str_trim(sapply(ll3,"[[",2)))
lon<-as.numeric(str_trim(sapply(ll3,"[[",1)))
df5<-data.frame(df4,lon,lat)
b11<-which(df5$lat<=58)
df6<-df5[-b11,]
b12<-which(df6$lat>=71)
df7<-df6[-b12,]
b13<-which(df7$lon<=19)
df8<-df7[-b13,]
b14<-which(df8$lon>=31)
df9<-df8[-b14,]
#####################Languages in the unfiltered data
df9l<-df9[,c(1,7:8)]
langs<-c("English", "Finnish", "Swedish", "Russian","Spanish","French",   "Norwegian","German","
    Estonian",   "Italian", "Dutch","Danish","Indonesian","Polish","Norwegian (Nynorsk)","Maltese
    ", "Bulgarian", "Serbian","Ukranian","Portuguese", "Lithuanian","Malaysian", "Tagalog","
    Macedonian","Turkish","Norwegian (Bokmål)",
 "Swahili","Japanese","Slovenian","Basque","Belarusian","Kazakh",  "Afrikaans","Latvian","Kyrgyz
     ","Mongolian","Ikinyarwanda","Welsh","Romanian","Korean","Latin","Croatian", "Icelandic","
     Mandarin",
```

```
"Malagasy","Esperanto","Luxembourgish","Volapük","Xhosa","Slovak", "Hungarian","Breton","Javanese","Catalan","Czech","Azeri","Arabic",
 "Quechua","Galician","Albanian","Zulu","Bosnian","Occitan","Walloon","Irish","Northern Sami","Thai",
  "Aragonese","Hatian Creole","Farsi",  "Georgian","Faroese","Vietnamese","Hindi","Hebrew","Kurdish","Bengali","Amharic","Urdu","Armenian","Marathi","Khmer","Nepali","Panjbi","Pashto","Sinhala","Greek","Oriya","Tamil","Assamese","Gujarati","Lao","Malayalam","Uyghur")
lang.order<-match(df91$V7,names(sort(table(df91$V7),decreasing=T)))
df91$langs<-langs[lang.order]

lang.plot<-barplot(sort(table(df91$langs),decreasing=T)[1:20],ylim=c(0,57000),las=2,cex.names=.7)
title("Number of Tweets by Language (Automatic Detection by langid.py),\nFinland Corpus, Top 20 Languages")
text(lang.plot,sort(table(df91$langs),decreasing=T)[1:20]+1500,labels=round(sort(table(df91$langs),decreasing=T)[1:20],digits=2),cex=.5,col="black")

filtered<-df91[df91$V8>.6,]   ###Languages with >.6 probabilistic accuracy
filtered.plot<-barplot(sort(table(filtered$langs),decreasing=T)[1:20],ylim=c(0,55000),las=2,cex.names=.7)
title("Number of Tweets by Language (Automatic Detection by langid.py),\nFinland Corpus Top 20 Languages, Probabilistic Accuracy > 0.6")
text(filtered.plot,sort(table(filtered$langs),decreasing=T)[1:20]+1500,labels=round(sort(table(filtered$langs),decreasing=T)[1:20],digits=2),cex=.5,col="black")

sum(sort(table(filtered$langs),decreasing=T)[1:4])/length(filtered[,1])   #percent of filtered tweets that are in English, Finnish, Swedish and Russian

###Geocoding the unfiltered tweets
lonA.all<-df9$lon;latA.all<-df9$lat
lonlatA.all<-data.frame(lonA.all,latA.all)

coordinates(lonlatA.all)<-~lonA.all+latA.all #creates Spatial Points Frame
proj4string(lonlatA.all)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0")
      #use the proj4string of the gadm projection
lonlatA.all<-spTransform(lonlatA.all,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0"))#as above
overA.all<-over(lonlatA.all,as(fin.adm2.spdf,"SpatialPolygons"))
overA.all.df<-data.frame(table(overA.all))
nrtweetsA.all = overA.all.df$Freq
```

```
nrtweetsA.all.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweetsA.all)
meanlength.A.all<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(overA.all==x),1])))) #low
    because lots of tweets out of Finnish borders; overA.all is unfiltered for automated tweets,
    over.all is filtered.
names(nrtweetsA.all.df)[1]<-"id"
df9$province<-fin.adm2.spdf$NAME_2[overA.all]

rm(df8,df7,df6,df5,df4,df3,df2,df,aaa2,aaa1,aaa,aa,b1,bbb,bb,b8,b9,b10,b11,b12,b13,b14,df91,lang.
    order,langs,lang.plot, filtered, filtered.plot)

comb7a<-readLines(<path to tweet textfile>)
dfc1<-readLines(<path to PoS tagged textfile>)
dfc2<-gsub(".*,\\('","",dfc1) #a list with all characters after lang name.
dfc3<-substr(dfc2,1,2) #lang
dfc4<-substr(dfc2,6,10) #probability of langid
dfc.prob<-which(dfc4=="1.0)")
dfc4[dfc.prob]<-"1.000"
dfc5<-as.numeric(dfc4)
dfc6<-data.frame(comb7a,dfc3,dfc5,stringsAsFactors=F)

qqc<-grep("^\".*\"$",dfc6[,1])
dfc6qs<-gsub("^.","",dfc6[qqc,1])
dfc6qs1<-gsub(".$","",dfc6qs)
dfc6qs2<-gsub("\"\"","\"",dfc6qs1)
dfc6[qqc,1]<-dfc6qs2

incx<-which(dfc6[,2]=="en" & dfc6[,3]>=.600)
#dfc6[incx,] is Comparison English Corpus with probability of "en" tag greater than 60%

rm(dfc5,dfc4,dfc.prob,dfc3,dfc2,dfc1,comb7a,dfc6qs,dfc6qs1,dfc6qs2,qqc,dfc6.1,dfc6.2,dfc6.3,dfc6
    .4,dfc6.5)

####Filtering out the automated tweets
df9.0<-df9
del<-grep("°C. Rain today",df9.0$V1)
df9.1<-df9.0[-del,]
bongi<-grep("BONGI",df9.1$V1)
df9.2<-df9.1[-bongi,]
del1<-grep("with #Endomondo. See it here",df9.2[,1])
```

```
df9.3<-df9.2[-del1,]
del2<-grep("speed up time and get an early listen to @ddlovato's new",df9.3[,1])
df9.4<-df9.3[-del2,]
del3<-grep("AustinMahone I would be the happiest girl",df9.4[,1])
df9.5<-df9.4[-del3,]
del4<-grep("helsinkiairport",df9.5[,1])
df9.6<-df9.5#[-del4,]                          #This element added to message by iphone app, left
      in
del5<-grep("Modis Flood",df9.6[,1])
df9.7<-df9.6[-del5,]
del6<-grep("RÄPPII \\(official",df9.7[,1])
df9.8<-df9.7[-del6,]                           #no longer in basis data
del7<-grep("Just posted a photo",df9.8[,1])
df9.9<-df9.8[-del7,]
del8<-grep("as the mayor of",df9.9[,1])
df9.10<-df9.9[-del8,]
del9<-grep("justinbieber OMG I can't believe u r in Finland",df9.10[,1])
df9.11<-df9.10[-del9,]
del10<-grep("dankanter Please follow me!!\\?!\\? I love u so much\\?",df9.11[,1])
df9.12<-df9.11[-del10,]
del11<-grep("w/ 2 others",df9.12[,1])  #This element added to message by iphone app, left in so
     as to not lose main body of text
df9.13<-df9.12#[-del11,]                       #
del12<-grep("w/ 3 others",df9.13[,1])  #
df9.14<-df9.13#[-del12,]                       #
del13<-grep("Method And Apparatus For",df9.14[,1])
df9.15<-df9.14[-del13,]
del14<-grep("danielsahyounie Please Skip follow me I need",df9.15[,1])
df9.16<-df9.15[-del14,]
del15<-grep("Finland I'm about to pass out! Ilysm",df9.16[,1])
df9.160<-df9.16#[-del15,]                        #del15 is same as del9
del15.1<-grep("Harry_Styles PLEASE FOLLOW ME ILYSM",df9.160[,1])
df9.161<-df9.160[-del15.1,]
del16<-grep("NiallOfficial please hug me",df9.161[,1])
df9.162<-df9.161[-del16,]
del17<-grep("@AustinMahone U are my world",df9.162[,1])

df9.17<-df9.162[-del17,]
```

```
rm(df9.161,df9.160,df9.16,del16,del15,del14,del13,del12,del11,del10,del9,del8,del7,del6,del5,del4
    ,del3,del2,del1,del,bongi,df9.15,df9.14,df9.13,df9.12,df9.11,df9.10,df9.9,df9.8,df9.7,df9.6,
    df9.5,df9.4,df9.3,df9.2,df9.1,df9.0,df9,ll,ll1,ll2,ll3)
####### Tweet Frequencies from the filtered frame
###### All tweets
lon.all<-df9.17$lon;lat.all<-df9.17$lat
lonlat.all<-data.frame(lon.all,lat.all)
###### English tweets
en.inx<-which(df9.17$V7=="en" & df9.17$V8>=.600) #prob greater than 60% and tagged with 'fi'
df10.en<-df9.17[en.inx,]
lon.en<-df10.en$lon;lat.en<-df10.en$lat
lonlat.en<-data.frame(lon.en,lat.en)
###### Finnish tweets
fi.inx<-which(df9.17$V7=="fi" & df9.17$V8>=.600) #prob greater than 60% and tagged with 'fi'
df10.fi<-df9.17[fi.inx,]
lon.fi<-df10.fi$lon;lat.fi<-df10.fi$lat
lonlat.fi<-data.frame(lon.fi,lat.fi)
####### Swedish tweets
sv.inx<-which(df9.17$V7=="sv" & df9.17$V8>=.600) #prob greater than 60% and tagged with 'sv'
df10.sv<-df9.17[sv.inx,]
lon.sv<-df10.sv$lon;lat.sv<-df10.sv$lat
lonlat.sv<-data.frame(lon.sv,lat.sv)
####### Russian tweets
ru.inx<-which(df9.17$V7=="ru" & df9.17$V8>=.600) #prob greater than 60% and tagged with 'ru'
df10.ru<-df9.17[ru.inx,]
lon.ru<-df10.ru$lon;lat.ru<-df10.ru$lat
lonlat.ru<-data.frame(lon.ru,lat.ru)
#######Other language tweets
Nototh.inx<-sort(c(ru.inx,en.inx,sv.inx,fi.inx),decreasing=F)
df10.oth<-df9.17[-Nototh.inx,]
lon.oth<-df10.oth$lon;lat.oth<-df10.oth$lat
lonlat.oth<-data.frame(lon.oth,lat.oth)
#######Other language tweets all probabilities
Nototh2.inx<-c(which(df9.17$V7=="en"),which(df9.17$V7=="fi"),which(df9.17$V7=="sv"),which(df9.17$
    V7=="ru"))
df10.oth2<-df9.17[-Nototh2.inx,]
lon.oth2<-df10.oth2$lon;lat.oth2<-df10.oth2$lat
lonlat.oth2<-data.frame(lon.oth2,lat.oth2)

```

```
############ Overlaying the tweets into the Finland Geographical Polygon borders
### Overlaying all Tweets
coordinates(lonlat.all)<-~lon.all+lat.all #creates Spatial Points Frame
proj4string(lonlat.all)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0")
    #use the proj4string of the gadm projection
lonlat.all<-spTransform(lonlat.all,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
    towgs84=0,0,0"))#as above
over.all<-over(lonlat.all,as(fin.adm2.spdf,"SpatialPolygons"))
over.all.df<-data.frame(table(over.all))
nrtweets.all = over.all.df$Freq
nrtweets.all.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.all)
meanlength.all<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(over.all==x),1]))))   ###low
     because NA's outside Finnish borders
names(nrtweets.all.df)[1]<-"id"

negF<-which(is.na(over.all)==T)    #the tweets that do not fall within Finnish borders
df9.18<-df9.17[-negF,]              #frame with all the tweets within Finnish borders
df9.19<-df9.18[which(df9.18[,8]>.6),] #within Finnish borders and lang.id>.6
length(df9.19[which(df9.19[,7]=="ru"|df9.19[,7]=="sv"|df9.19[,7]=="en"|df9.19[,7]=="fi"),1])/
    length(df9.19[,1])                  #percent of above either fi,en,sv,ru
length(df9.19[which(df9.19[,7]=="fi"),1])/length(df9.19[,1])  #percent of above Finnish
length(unique(df9.19[which(df9.19[,7]=="fi"),2]))/length(unique(df9.19[,2])) #percent of unique
    users with at least one tweet in finnish
length(df9.19[which(df9.19[,7]=="en"),1])/length(df9.19[,1])  #percent of above english
length(unique(df9.19[which(df9.19[,7]=="en"),2]))/length(unique(df9.19[,2])) #percent of unique
    users with at least one tweet in english
length(df9.19[which(df9.19[,7]=="sv"),1])/length(df9.19[,1])  #percent of above swedish
length(unique(df9.19[which(df9.19[,7]=="sv"),2]))/length(unique(df9.19[,2])) #percent of unique
    users with at least one tweet in swedish
length(df9.19[which(df9.19[,7]=="de"),1])/length(df9.19[,1])  #percent of above russian
length(unique(df9.19[which(df9.19[,7]=="ru"),2]))/length(unique(df9.19[,2])) #percent of unique
    users with at least one tweet in russian

### Overlaying the English Tweets
coordinates(lonlat.en)<-~lon.en+lat.en #creates Spatial Points Frame
proj4string(lonlat.en)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0") #
    use the proj4string of the gadm projection
lonlat.en<-spTransform(lonlat.en,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
    towgs84=0,0,0"))#as above
```

```
over.en<-over(lonlat.en,as(fin.adm2.spdf,"SpatialPolygons"))
over.en.df<-data.frame(table(over.en))
nrtweets.en = over.en.df$Freq
nrtweets.en.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.en)
meanlength.en<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(over.en==x),1]))))
names(nrtweets.en.df)[1]<-"id"

### Overlaying the Finnish Tweets
coordinates(lonlat.fi)<-~lon.fi+lat.fi #creates Spatial Points Frame
proj4string(lonlat.fi)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0") #
    use the proj4string of the gadm projection
lonlat.fi<-spTransform(lonlat.fi,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
    towgs84=0,0,0"))#as above
over.fi<-over(lonlat.fi,as(fin.adm2.spdf,"SpatialPolygons"))
over.fi.df<-data.frame(table(over.fi))
nrtweets.fi = over.fi.df$Freq
nrtweets.fi.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.fi)
meanlength.fi<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(over.fi==x),1]))))
names(nrtweets.fi.df)[1]<-"id"

### Overlaying the Swedish Tweets
coordinates(lonlat.sv)<-~lon.sv+lat.sv
proj4string(lonlat.sv)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0") #
    use the proj4string of the gadm projection
lonlat.sv<-spTransform(lonlat.sv,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
    towgs84=0,0,0"))#as above
over.sv<-over(lonlat.sv,as(fin.adm2.spdf,"SpatialPolygons"))
#newfactor <- factor(oldfactor, exclude=NULL)                    #Necessary because some
    polygons zero frequency
over.sv.f<-factor(over.sv,levels=as.character(c(seq(1:21))))       #"
over.sv.f.df<-data.frame(table(over.sv.f))
nrtweets.f.sv = over.sv.f.df$Freq
nrtweets.f.sv.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.f.sv)
meanlength.sv.f<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(over.sv.f==x),1]))))
names(nrtweets.f.sv.df)[1]<-"id"

### Overlaying the Russian Tweets
coordinates(lonlat.ru)<-~lon.ru+lat.ru
```

```
proj4string(lonlat.ru)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0") #
    use the proj4string of the gadm projection
lonlat.ru<-spTransform(lonlat.ru,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
    towgs84=0,0,0"))#as above
over.ru<-over(lonlat.ru,as(fin.adm2.spdf,"SpatialPolygons"))
over.ru.f<-factor(over.ru,levels=as.character(c(seq(1:21))))        #Necessary because some
    polygons zero frequency
over.ru.f.df<-data.frame(table(over.ru.f))
nrtweets.f.ru = over.ru.f.df$Freq
nrtweets.f.ru.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.f.ru)
meanlength.ru.f<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(over.ru.f==x),1]))))
names(nrtweets.f.ru.df)[1]<-"id"

### Overlaying the Other Tweets(language != en,fi,sv,ru>0.6)
coordinates(lonlat.oth)<-~lon.oth+lat.oth
proj4string(lonlat.oth)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0")
    #use the proj4string of the gadm projection
lonlat.oth<-spTransform(lonlat.oth,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
    towgs84=0,0,0"))#as above
over.oth<-over(lonlat.oth,as(fin.adm2.spdf,"SpatialPolygons"))
over.oth.f<-factor(over.oth,levels=as.character(c(seq(1:21))))       #Necessary because some
    polygons zero frequency
over.oth.f.df<-data.frame(table(over.oth.f))
nrtweets.f.oth = over.oth.f.df$Freq
nrtweets.f.oth.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.f.oth)
meanlength.f.oth.f<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(over.oth.f==x),1]))))
names(nrtweets.f.oth.df)[1]<-"id"

### Overlaying the Other Tweets(language != en,fi,sv,ru>0)
coordinates(lonlat.oth2)<-~lon.oth2+lat.oth2
proj4string(lonlat.oth2)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0")
     #use the proj4string of the gadm projection
lonlat.oth2<-spTransform(lonlat.oth2,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
    towgs84=0,0,0"))#as above
over.oth2<-over(lonlat.oth2,as(fin.adm2.spdf,"SpatialPolygons"))
over.oth2.f<-factor(over.oth2,levels=as.character(c(seq(1:21))))      #Necessary because some
    polygons zero frequency
over.oth2.f.df<-data.frame(table(over.oth2.f))
nrtweets.f.oth2 = over.oth2.f.df$Freq
```

```
nrtweets.f.oth2.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.f.oth2)
meanlength.f.oth2.f<-unlist(lapply(1:21,function(x) mean(nchar(df11[which(over.oth2.f==x),1]))))
names(nrtweets.f.oth2.df)[1]<-"id"

##########  Getting the PoS tagged text back into the larger frame
#### This creates a frame in which each word in English tweets has a PoS tag, used for finding most frequent words for each PoS type
fPoS<-readLines(<path to Finland PoS data>,encoding="UTF-8")          #Penn treebank tags
fPoS1<-strsplit(fPoS,"\t")                                             #
fPoS2<-as.data.frame(do.call("rbind", fPoS1),stringsAsFactors=F)       #  just tokens
#fPoSD1<-strsplit(fPoSD,"\t")                                          #
#fPoSD2<-as.data.frame(do.call("rbind", fPoSD1),stringsAsFactors=F)     #  just tokens

fPoS1a<-sapply(fPoS1,"[",2)                    ######To get the PoS Sequences by Tweet
fPoS1a[is.na(fPoS1a)]<-""
fPoS1b<-paste(fPoS1a,collapse=" ")
fPoS1c<-strsplit(fPoS1b,"  ")
fPoS1d<-strsplit(unlist(fPoS1c," "),"  ")
fPoS1e<-sapply(fPoS1d,unlist)

fPoS2a<-sapply(fPoS1,"[",1)                    ######To get the word Sequences by Tweet
fPoS2a[is.na(fPoS2a)]<-""
fPoS2b<-paste(fPoS2a,collapse=" ")
fPoS2c<-strsplit(fPoS2b,"  ")
fPoS2d<-strsplit(unlist(fPoS2c," "),"  ")
fPoS2e<-sapply(fPoS2d,unlist)
######This creates a frame for each English tweet with the frequency of each PoS tag and total number of tokens
lev <- c(sort(unique(unlist(fPoS1d))),"TOK")
G1 <- do.call(rbind,lapply(fPoS1d,function(x,lev){ table(factor(x,levels = lev,ordered = TRUE))}, lev = lev))
G2<-as.data.frame(G1)
G2$TOK<-unlist(lapply(fPoS1d, function(x) (sum(table(factor(x,levels=lev))))))

df10<-df9.17[en.inx,]              #This is all the filtered English tweets
df11<-data.frame(df10,unlist(fPoS1c),unlist(fPoS2c),G2,stringsAsFactors=F)          #This new frame has for each English tweet the original metadata, lat/long plus PoS and individual token info
names(df11)[12:19]<-c("PoSseq","Tokens","''","-LRB-","-RRB-",",",".",":")
```

```
rm(fPoS1a,fPoS1b,fPoS1c,fPoS1d,fPoS1e,fPoS2a,fPoS2b,fPoS2c,fPoS2d,fPoS2e,fPoS,fPoS1)

###### Same framework for Comparison Corpus
cPoS<-readLines(<path to PoS-tagged Comparison data>,encoding="UTF-8")
cPoS1<-strsplit(cPoS,"\t")
cPoS2<-as.data.frame(do.call("rbind", cPoS1),stringsAsFactors=F)

cPoS1a<-sapply(cPoS1,"[",2)                    ######To get the PoS Sequences by Tweet
cPoS1a[is.na(cPoS1a)]<-""
cPoS1b<-paste(cPoS1a,collapse=" ")
cPoS1c<-strsplit(cPoS1b,"  ")
cPoS1d<-strsplit(unlist(cPoS1c," "),"  ")
cPoS1e<-sapply(cPoS1d,unlist)

cPoS2a<-sapply(cPoS1,"[",1)                    ######To get the word Sequences by Tweet
cPoS2a[is.na(cPoS2a)]<-""
cPoS2b<-paste(cPoS2a,collapse=" ")
cPoS2c<-strsplit(cPoS2b,"  ")
cPoS2d<-strsplit(unlist(cPoS2c," "),"  ")
cPoS2e<-sapply(cPoS2d,unlist,rec=F)

cEngText<-readLines(<path to Comparison data>,encoding="UTF-8")
cPoSb<-readLines(<path to Comparison data PoS tags >,encoding="UTF-8")
cPoSb1<-paste(cPoSb,collapse=" ")
cPoSb2<-strsplit(cPoSb1,"  ")
cPoSb3<-strsplit(unlist(cPoSb2," "),"  ")

levc <- c(sort(unique(unlist(cPoS1d))),"TOK")
G1c <- do.call(rbind,lapply(cPoS1d,function(x,levc){ table(factor(x,levels = levc,ordered = TRUE)
    )},lev = levc))
G2c<-as.data.frame(G1c)
G2c$TOK<-unlist(lapply(cPoS1d, function(x) (sum(table(factor(x,levels=levc))))))

dfc10<-cEngText          #  the same as dfc6[incx,] above, was saved as ctwa2.txt
dfc11<-data.frame(dfc10,unlist(cPoS1c),unlist(cPoS2c),G2c,stringsAsFactors=F)
names(dfc11)[1:9]<-c("V1","PoSseq","Tokens","''","-LRB-","-RRB-",",",".",":")

rm(cPoS1a,cPoS1b,cPoS1c,cPoS1d,cPoS1e,cPoS2a,cPoS2b,cPoS2c,cPoS2d,cPoS2e,cPoS,cPoS1)
```

```
362
363 neg<-grep("Listening to",dfc11[,1])
364 dfc11.1<-dfc11[-neg,]
365 neg<-grep("? listening to",dfc11.1[,1])
366 dfc11.2<-dfc11.1[-neg,]
367 neg<-grep("New blog post:",dfc11.2[,1])
368 dfc11.3<-dfc11.2[-neg,]
369 neg<-grep("team hyjak -",dfc11.3[,1])
370 dfc11.4<-dfc11.3[-neg,]
371 neg<-grep("James Walker",dfc11.4[,1])
372 dfc11.5<-dfc11.4[-neg,]
373 neg<-grep("#listening to",dfc11.5[,1])
374 dfc11.6<-dfc11.5[-neg,]
375 dfc11<-dfc11.6
376 dfc11<-dfc11[,-23]
377 rm(dfc11.1,dfc11.2,dfc11.3,dfc11.4,dfc11.5,dfc11.6)
378 finwcap<-unlist(strsplit(df11[,13]," "));finw<-tolower(finwcap)
379 compwcap<-unlist(strsplit(dfc11[,3]," "));compw<-tolower(compwcap)
380 finp<-unlist(strsplit(df11[,12]," "))
381 compp<-unlist(strsplit(dfc11[,2]," "))
382 finw.del.np<-gsub("[[:punct:]]","",finw);finw.np<-finw.del.np[-which(finw.del.np=="")]
383 compw.del.np<-gsub("[[:punct:]]","",compw);compw.np<-compw.del.np[-which(compw.del.np=="")]
384
385 #### English tweets from Finland with at least 1 tag UH
386 uhs<-df11[df11[,41]>0,]
387 lon.uhs<-df11$lon;lat.uhs<-df11$lat
388 lonlat.uhs<-data.frame(lon.uhs,lat.uhs)
389
390 coordinates(lonlat.uhs)<-~lon.uhs+lat.uhs #creates Spatial Points Frame
391 proj4string(lonlat.uhs)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0")
       #use the proj4string of the gadm projection
392 lonlat.uhs<-spTransform(lonlat.uhs,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +
       towgs84=0,0,0"))#as above
393 over.uhs<-over(lonlat.uhs,as(fin.adm2.spdf,"SpatialPolygons"))
394 over.uhs.df<-data.frame(table(over.uhs))
395 nrtweets.uhs = over.uhs.df$Freq
396 nrtweets.uhs.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.uhs)
397 names(nrtweets.uhs.df)[1]<-"id"
398
```

```
######### Merging the data into a single frame
demog.dat<-read.table(<path to file with demographic statistics>,header=T,stringsAsFactors=T)
nrtweets.1000<-nrtweets.all*1000/demog.dat$Population
demog.dat1<-data.frame(demog.dat,nrtweets.all,meanlength.all,nrtweets.en,meanlength.en,nrtweets.
    fi,meanlength.fi,nrtweets.f.sv,meanlength.sv.f,nrtweets.f.ru,meanlength.ru.f,nrtweets.f.oth,
    meanlength.f.oth.f,nrtweets.f.oth2,meanlength.f.oth2.f,nrtweets.uhs,nrtweets.1000)
frame.all<-merge(fin.adm2.df,demog.dat1,by.y="id",all.x=T)

######### MAPPING THE TWEETS
########  Location of Tweets in the data set as points (faulty geocoordinates filtered out)
lonlat2<-data.frame(lonlat.all)
lonlat3a<-lonlat2[(which(lonlat2$lon>21& lonlat2$lon<31)),]                #Additional bbox
    filtering
lonlat4a<-lonlat3a[(which(lonlat3a$lat>60 & lonlat3a$lat<70)),]

pPoints <- ggplot(frame.all, aes(x = long, y = lat,group=group)) + geom_polygon(fill="bisque")+
    geom_point(data=lonlat4a,aes(group=NULL,x=lon.all,y=lat.all),size=1,pch=16,color="red")+
geom_path(color="black") + labs(x=" ", y=" ")  + coord_map() + ggtitle("Location of Tweets
    Collected in the Data Set")

###Center of each region; for labelling on map
fin.adm2.centroids.df<- data.frame(long = coordinates(fin.adm2.spdf)[, 1],lat = coordinates(fin.
    adm2.spdf)[, 2])
fin.adm2.centroids.df[, 'ID_2'] <- fin.adm2.spdf@data[,'ID_2']
fin.adm2.centroids.df[, 'NAME_2'] <- fin.adm2.spdf@data[,'NAME_2']

##Number of Tweets by Province
p.tweetP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill=
    nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.all,digits=1), x =
    long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("",low = "#ffffcc", high = "#ff4444",space = "Lab", na.
    value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Number of Tweets")

##Number of Tweets by Province Population
p.tweet1000P <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill=
    nrtweets.1000)) +
```

```
geom_text(data = fin.adm2.centroids.df, aes(label = round(nrtweets.1000,digits=1), x = long, y =
    lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("",low = "#ffffcc", high = "#ff4444",space = "Lab", na.
    value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Number of Tweets per 1000 Population")

##Proportion of Tweets in English

p.enP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill=
    nrtweets.en*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.en*100/demog.dat1$
    nrtweets.all,digits=1), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in English")

##Proportion of Tweets in Finnish
p.fiP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.fi*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.fi*100/demog.dat1$
    nrtweets.all,digits=1), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Finnish")

##Proportion of Tweets in Swedish
p.svP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.f.sv*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.f.sv*100/demog.dat1
    $nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Swedish")

##Proportion of Tweets in Russian
p.ruP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.f.ru*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.f.ru*100/demog.dat1
    $nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
```

```
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Russian")

##Proportion of Tweets in other, p for en,fi,sv,ru >.6
p.othP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill = nrtweets.f.oth*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label =round(demog.dat1$nrtweets.f.oth*100/demog.dat1$nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Other Language,\n Language Tag Probability > 0.6")

##Proportion of Tweets in other, p for en,fi,sv,ru >0
p.oth2P <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill = nrtweets.f.oth2*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.f.oth2*100/demog.dat1$nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + theme_bw() + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Other Language,\n Language Tag Probability > 0")

p.gdpP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill = GDPperCap)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$GDPperCap,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + theme_bw() + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("GDP per Capita")

######Multiplot for ggplot2 from  http://stackoverflow.com/questions/24387376/r-wired-error-could-not-find-function-multiplot

multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  require(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

```

```
481   numPlots = length(plots)
482
483   # If layout is NULL, then use 'cols' to determine layout
484   if (is.null(layout)) {
485     # Make the panel
486     # ncol: Number of columns of plots
487     # nrow: Number of rows needed, calculated from # of cols
488     layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
489                      ncol = cols, nrow = ceiling(numPlots/cols))
490   }
491
492  if (numPlots==1) {
493     print(plots[[1]])
494
495   } else {
496     # Set up the page
497     grid.newpage()
498     pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))
499
500     # Make each plot, in the correct location
501     for (i in 1:numPlots) {
502       # Get the i,j matrix positions of the regions that contain this subplot
503       matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))
504
505       print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
506                                       layout.pos.col = matchidx$col))
507     }
508   }
509 }
510
511 scaleA<-scale_fill_gradient("In Percent",limits=c(0,6),low = "#ffffcc", high = "#ff4444",space =
       "Lab", na.value = "grey80", guide = "colourbar")
512 scaleB<-scale_fill_gradient("In Percent",limits=c(0,3),low = "#ffffcc", high = "#ff4444",space =
       "Lab", na.value = "grey80", guide = "colourbar")
513 scaleC<-scale_fill_gradient("In Percent",limits=c(0,.5),low = "#ffffcc", high = "#ff4444",space =
        "Lab", na.value = "grey80", guide = "colourbar")
514 scaleD<-scale_fill_gradient("In Percent",limits=c(0,1),low = "#ffffcc", high = "#ff4444",space =
       "Lab", na.value = "grey80", guide = "colourbar")
515
```

```
library(gridExtra)
grid_arrange_shared_legend <- function(...) {
    plots <- list(...)
    g <- ggplotGrob(plots[[1]] + theme(legend.position="right"))$grobs
    legend <- g[[which(sapply(g, function(x) x$name) == "guide-box")]]
    lheight <- sum(legend$height)
    grid.arrange(
        do.call(arrangeGrob, lapply(plots, function(x)
            x + theme(legend.position="none"))),
        legend,
        ncol = 1,
        heights = unit.c(unit(1, "npc") - lheight, lheight))
}
grid_arrange_shared_legend(p.emot20.names.1P, p.emot20.names.2P, p.emot20.names.3P, p.emot20.
    names.4P)
multiplot(p.emot20.names.1P + scaleA, p.emot20.names.2P + scaleA, p.emot20.names.3P + scaleA, p.
    emot20.names.4P + scaleA,cols=4)


##Proportion of Tweets in English
p.enP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill=
    nrtweets.en*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.en*100/demog.dat1$
    nrtweets.all,digits=1), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in English")

##Proportion of Tweets in Finnish
p.fiP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.fi*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.fi*100/demog.dat1$
    nrtweets.all,digits=1), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Finnish")

##Proportion of Tweets in Swedish
```

```
p.svP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.f.sv*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.f.sv*100/demog.dat1
    $nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Swedish")

##Proportion of Tweets in Russian
p.ruP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.f.ru*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.f.ru*100/demog.dat1
    $nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Russian")

##Proportion of Tweets in other, p for en,fi,sv,ru >.6
p.othP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.f.oth*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.f.oth*100/demog.
    dat1$nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Other Language,\n Language Tag Probability > 0.6")

##Proportion of Tweets in other, p for en,fi,sv,ru >0
p.oth2P <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.f.oth2*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat1$nrtweets.f.oth2*100/demog.
    dat1$nrtweets.all,digits=3), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2) +
labs(x=" ", y=" ") + theme_bw() + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#
    ff4444",space = "Lab", na.value = "grey80", guide = "colourbar") + coord_map() +
ggtitle("Proportion of Tweets in Other Language,\n Language Tag Probability > 0")

scaleE<-scale_fill_gradient("In Percent",limits=c(0,70),low = "#ffffcc", high = "#ff4444",space =
     "Lab", na.value = "grey80", guide = "colourbar")
multiplot(p.fiP + scaleE+ theme(plot.margin=NULL),p.enP + scaleE+ theme(plot.margin=NULL), p.svP
    + scaleE+ theme(plot.margin=NULL), p.ruP + scaleE+ theme(plot.margin=NULL),cols=4)
```

```
571
572 ############ FINDING THE GENDER OF FINNISH USERS OF TWITTER
573 ############
574
575 femalenames<-readLines(<path to list of female names>)
576 femalenames1<-femalenames[femalenames!=""]
577 femalenames2<-unlist(strsplit(femalenames1," "))
578 femalenames3<-tolower(femalenames2)
579
580 malenames<-readLines(<path to list of male names>)
581 malenames1<-malenames[malenames!=""]
582 malenames2<-unlist(strsplit(malenames1," "))
583 malenames3<-tolower(malenames2)
584
585 finnnames<-unique(df11[,2])   #Usernames of Finland English Corpus
586
587 finnmales<-sapply(malenames3,function(y) grep(y,finnnames))
588 finnmales1<-unlist(finnmales)
589 finnmales2<-finnnames[finnmales1]
590
591 finnfemales<-sapply(femalenames3,function(y) grep(y,finnnames))
592 finnfemales1<-unlist(finnfemales)
593 finnfemales2<-finnnames[finnfemales1]
594
595 df12<-df11[df11[,2] %in% c(finnmales2,finnfemales2),]
596 maletweets<-which(df11[,2] %in% finnmales2)
597 femaletweets<-which(df11[,2] %in% finnfemales2)
598
599 maletweets12<-which(df12[,2] %in% finnmales2)
600 femaletweets12<-which(df12[,2] %in% finnfemales2)
601
602 df11$gender<-""
603 df12$gender[maletweets12]<-"m"                    #all tweets with gender
604 df12$gender[femaletweets12]<-"f"
605
606 df11$gender[maletweets]<-"m"                      #all English tweets
607 df11$gender[femaletweets]<-"f"
608
609 df11females<-df11[df11[,56]=="f",]
```

```
df11males<-df11[which(df11[,56]=="m"),]

############## Tweet length
###############

flength<-table(nchar(df11[,1]))           # Finland English
flengthm<-table(nchar(df11boys[,1]))      # Finland English male
flengthf<-table(nchar(df11girls[,1]))     # Finland English female
clength<-table(nchar(dfc11[,1]))          #Comparison English

lnames<-intersect(names(flength),names(clength))
lnamesmf<-intersect(names(flengthm),names(flengthf))
fln<-flength[lnames]
cln<-clength[lnames];cln[141]<-12
flnm<-flengthm[lnamesmf]
flnf<-flengthf[lnamesmf]
tweetlength<-rbind((fln/length(df11[,1])),(cln/length(dfc11[,1])))
tweetlengthmf<-rbind((flnm/length(df11boys[,1])),(flnf/length(df11girls[,1])))

ks.test((fln/length(df11[,1])),(cln/length(dfc11[,1]))) #two-sample kolmogorov-smirnov
t.test(nchar(df11[,1]),nchar(dfc11[,1]))    #two-sample pop. means (Welch's t)

ks.test((flnm/length(df11boys[,1])),(flnf/length(df11girls[,1]))) #two-sample kolmogorov-smirnov
t.test(nchar(df11boys[,1]),nchar(df11girls[,1]))

fncharm<-round(mean(nchar(df11[,1])),digits=3)
fncharsd<-round(sd(nchar(df11[,1])),digits=3)
fncharmm<-round(mean(nchar(df11boys[,1])),digits=3)   #males
fncharmsd<-round(sd(nchar(df11boys[,1])),digits=3)
fncharfm<-round(mean(nchar(df11girls[,1])),digits=3)   #females
fncharfsd<-round(sd(nchar(df11girls[,1])),digits=3)
cncharm<-round(mean(nchar(dfc11[,1])),digits=3)
cncharsd<-round(sd(nchar(dfc11[,1])),digits=3)

tweetlength.mp <- barplot(tweetlength,beside=T,xlab="Number of characters",ylab="Proportion of
    Corpus")
legend("top",legend=c(expression(paste("Finland English Corpus, ",mu," = 71.10, ",sigma," = 36.05
    ")),expression(paste("Comparison English Corpus, ",mu," = 78.99, ",sigma," = 36.41"))),cex
    =0.8,col=gray.colors(2),fill=gray.colors(2))
```

```
title("Tweet Length in Characters")

tweetlengthmf.mp <- barplot(tweetlengthmf,beside=T,xlab="Number of characters",ylab="Proportion
    of Corpus")
legend("top",legend=c(expression(paste("Finland English Corpus, males, ",mu," = 76.31, ",sigma,"
    = 36.14")),expression(paste("Finland English Corpus, females, ",mu," = 65.78, ",sigma," =
    34.75"))),cex=0.8,col=gray.colors(2),fill=gray.colors(2))
title("Tweet Length in Characters")

fwordlength<-table(df11[,55])
cwordlength<-table(dfc11[,45])
wordlnames<-intersect(names(fwordlength),names(cwordlength))
fwordln<-fwordlength[wordlnames]
cwordln<-cwordlength[wordlnames]
tweetwordlength<-rbind((fwordln/length(df11[,55])),(cwordln/length(dfc11[,45])))

fwordm<-round(mean(df11[,55]),digits=2)
fwordsd<-round(sd(df11[,55]),digits=2)
cwordm<-round(mean(dfc11[,45]),digits=2)
cwordsd<-round(sd(dfc11[,45]),digits=2)

#word length with nopunctuation
df11a<-strsplit(df11[,13]," ")
df11b<-sapply(df11a,function(x) gsub("[[:punct:]]","",x))
df11b.1<-sapply(df11b,function(x) x<-x[-which(x=="")])
df11c<-sapply(df11b.1,function(x) length(x))
df11d<-round(mean(df11c),digits=2)

dfc11a<-strsplit(dfc11[,3]," ")
dfc11b<-sapply(dfc11a,function(x) gsub("[[:punct:]]","",x))
dfc11b.1<-sapply(dfc11b,function(x) x<-x[-which(x=="")])
dfc11c<-sapply(dfc11b.1,function(x) length(x))
dfc11d<-round(mean(dfc11c),digits=2)

ks.test((fln/length(df11[,55])),(cln/length(dfc11[,45]))) #two-sample kolmogorov-smirnov
t.test((df11[,55]),(dfc11[,45]))    #two-sample pop. means (Welch's t)

tweetwordlength.mp <- barplot(tweetwordlength,beside=T,xlab="Number of words",ylab="Proportion of
    Corpus")
```

```
legend("top",legend=c(expression(paste("Finland English Corpus, ",mu," = 13.27, ",sigma," = 7.39")),expression(paste("Comparison English Corpus, ",mu," = 15.73, ",sigma," = 7.62"))),cex=0.8,col=gray.colors(2),fill=gray.colors(2))
title("Tweet Length in Tokens")

#########Word length
#########

mean(nchar(unlist(strsplit(df11[,13]," ")))) #mean word length, finland engl.
mean(nchar(unlist(strsplit(dfc11[,3]," ")))) #mean word length, comp. engl.

t.test(nchar(unlist(strsplit(df11[,13]," "))),nchar(unlist(strsplit(dfc11[,3]," "))))

length(which(nchar(unlist(strsplit(df11[,13]," ")))>6))/length(unlist(strsplit(df11[,13]," "))) #proportion of tokens longer than 6 characters, finland english
length(which(nchar(unlist(strsplit(dfc11[,3]," ")))>6))/length(unlist(strsplit(dfc11[,3]," "))) #proportion of tokens longer than 6 characters, comp english
mean(nchar(unlist(strsplit(df11boys[,13]," ")))) #mean word length, finland engl. male
mean(nchar(unlist(strsplit(df11girls[,13]," ")))) #mean word length, finland engl. female

t.test(nchar(unlist(strsplit(df11boys[,13]," "))),nchar(unlist(strsplit(df11girls[,13]," "))))

length(which(nchar(boysw)>6))/length(boysw)#proportion of tokens longer than 6 characters,FI english male
length(which(nchar(girlsw)>6))/length(girlsw)#proportion of tokens longer than 6 characters,FI english female
chisq.test(matrix(c(length(which(nchar(boysw)>6)),length(boysw)-length(which(nchar(boysw)>6)),length(which(nchar(girlsw)>6)),length(girlsw)-length(which(nchar(girlsw)>6))),ncol=2))

#Zipf plot rank frequency from Kosice presentation
library(ggplot2)
library(zipfR)
df11Zipf<-tolower(df11[,13])
df11Zipf1.1<-unlist(strsplit(df11Zipf," "))
df11Zipf2<-gsub("[[:digit:]]{1,}","NUM",df11Zipf1.1)
df11Zipf3<-sort(table(df11Zipf2),decreasing=T)
df11Zipf.tfl<-tfl(df11Zipf3,type=names(df11Zipf3))
dfc11Zipf<-tolower(dfc11[,3])
dfc11Zipf1.1<-unlist(strsplit(dfc11Zipf," "))
```

```
dfc11Zipf2<-gsub("[[:digit:]]{1,}","NUM",dfc11Zipf1.1)
dfc11Zipf3<-sort(table(dfc11Zipf2),decreasing=T)
dfc11Zipf.tfl<-tfl(dfc11Zipf3,type=names(dfc11Zipf3))

ZipfF<-df11Zipf.tfl[1:100,]
ZipfC<-dfc11Zipf.tfl[1:100,]

Zipf.P<-ggplot(ZipfF,aes(log10(k),log10(f)))+geom_point(color="darkblue",size=3,alpha=1/2)+geom_
    point(data=ZipfC,color="red",size=3,alpha=1/2)
Zipf.P+geom_text(hjust=ifelse(1:nrow(ZipfF) %% 2 == 1, 0, 1),label=sprintf(" %s ",ZipfF$type),
    size=3,colour="darkblue")+
  geom_text(data=ZipfC,hjust=ifelse(1:nrow(ZipfC) %% 2 == 1, 0, 1),label=sprintf(" %s ",ZipfC$
      type),size=3,colour="red") +
    labs(title="Zipf profile, top 100 types, Comparison English and Finland English Corpora",
        lineheight=.8, face="bold")   +
    ylab("log frequency") + xlab("log rank")

############ Lexical features
############

bothfc<-intersect(finw1,compw1)
finw2<-table(factor(finw1,levels=bothfc))
compw2<-table(factor(compw1,levels=bothfc))
LEXfc<-data.frame(1:length(finw2),names(finw2),unname(finw2),unname(compw2))          #observed
    frequencies for lexical items by corpus
rownames(LEXfc)<-NULL
LEXfc[,3]<-NULL;LEXfc[,4]<-NULL                  #remove R artifact columns
names(LEXfc)<-c("id","word","o11","o12")      #setting up as contingency table cells
LEXfc$o21<-length(finw)-LEXfc$o11               #non-occurrences
LEXfc$o22<-length(compw)-LEXfc$o12              #non-occurrences
LEXfc1<-transform(LEXfc,o11=as.numeric(o11),o12=as.numeric(o12),o21=as.numeric(o21),o22=as.
    numeric(o22))                                #to allow vector calculations
LEXfc1<-transform(LEXfc1,r1=o11+o12,r2=o21+o22,c1=o11+o21,c2=o12+o22,N=o11+o12+o21+o22)  #row,
    column and total sums
LEXfc1<-transform(LEXfc1,e11=(r1*c1)/N,e12=(r1*c2)/N,e21=(r2*c1)/N,e22=(r2*c2/N))   #expected
    values from contingency table
LEXfc1<-transform(LEXfc1,chisq = N * (abs(o11*o22 - o12*o21)-N/2)^2/(r1*r2*c1*c2)) #This is a
    vectorized version of the Chisq.test with Yates' continuity correction
```

```
LEXfc1<-transform(LEXfc1,logl=2*(ifelse(o11>0,o11*log(o11/e11),0)+ifelse(o12>0,o12*log(o12/e12)
    ,0)+ifelse(o21>0,o21*log(o21/e21),0)+ifelse(o22>0,o22*log(o22/e22),0)))     #Dunning's G (log
    -likelihood ratio)
LEXfc1<-transform(LEXfc1,mi=log(o11/e11))            #mutual information
LEXfc1<-transform(LEXfc1,oddsratio=log(((o11+.5)*(o22+.5))/((o12+.5)*(o21+.5))))   #odds ratio (
    theta)
LEXfc1<-transform(LEXfc1,dice=(2*o11)/(r1+c1))        #Dice coefficient
LEXfc1<-transform(LEXfc1,fc=(o11/e11)/(o12/e12))      #Finland-Comparison ratio
LEXfc1<-transform(LEXfc1,chisqp.value=pchisq(chisq,1,lower.tail=F))
LEXfc1$signi<-""
LEXfc1$signi[which(LEXfc1$chisqp.value<.05)]<-"*"
LEXfc1$signi[which(LEXfc1$chisqp.value<.01)]<-"**"
LEXfc1$signi[which(LEXfc1$chisqp.value<.001)]<-"***"
LEXfc1.1<-LEXfc1[which(LEXfc1[,3]>0),]
LEXfc1.2<-LEXfc1.1[which(LEXfc1.1[,4]>0),]   #subset of data; only tokens with at least one
    occurrence in both Finland English and Comparison English
length(LEXfc1.2[which(LEXfc1.2$fc<=.8333333333),1])+length(LEXfc1.2[which(LEXfc1.2$fc>=1.2),1]) #
    number of types with substantial difference in Finland/Comparison use
LEXfcfc<-LEXfc1.2[order(-LEXfc1.2$fc),]
row.names(LEXfcfc)<-1:length(LEXfcfc[,1])
xtable(LEXfcfc[1:20,c(2,21,16,22,23)],digits=c(0,0,0,1,-2,0))
LEXfccf<-LEXfc1.2[order(LEXfc1.2$fc),]
LEXfccf$fc<-1/LEXfccf$fc
row.names(LEXfccf)<-1:length(LEXfccf[,1])
xtable(LEXfccf[1:21,c(2,21,16,22,23)],digits=c(0,0,0,1,-2,0))
LEXtotal<-data.frame(LEXfcfc[1:20,c(2,21)],LEXfccf[c(1:19,21),c(2,21)])
xtable(LEXtotal,digits=c(0,0,0,0,0))

#Lexical items by gender
LEX<-data.frame(1:length(malesbothw1),names(malesbothw1),unname(malesbothw1),unname(femalesbothw1
    ))          #observed frequencies for lexical items by gender
rownames(LEX)<-NULL
LEX[,3]<-NULL;LEX[,4]<-NULL                 #remove R artifact columns
names(LEX)<-c("id","word","o11","o12")      #setting up as contingency table cells
LEX$o21<-length(malesw)-LEX$o11              #non-occurrences
LEX$o22<-length(femalesw)-LEX$o12             #non-occurrences
LEX1<-transform(LEX,o11=as.numeric(o11),o12=as.numeric(o12),o21=as.numeric(o21),o22=as.numeric(
    o22))                                 #to allow vector calculations
```

```
LEX1<-transform(LEX1,r1=o11+o12,r2=o21+o22,c1=o11+o21,c2=o12+o22,N=o11+o12+o21+o22)  #row, column and total sums
LEX1<-transform(LEX1,e11=(r1*c1)/N,e12=(r1*c2)/N,e21=(r2*c1)/N,e22=(r2*c2/N))   #expected values from contingency table
LEX1<-transform(LEX1,chisq = N * (abs(o11*o22 - o12*o21)-N/2)^2/(r1*r2*c1*c2)) #This is a vectorized version of the Chisq.test with Yates' continuity correction
LEX1<-transform(LEX1,logl=2*(ifelse(o11>0,o11*log(o11/e11),0)+ifelse(o12>0,o12*log(o12/e12),0)+ifelse(o21>0,o21*log(o21/e21),0)+ifelse(o22>0,o22*log(o22/e22),0)))     #Dunning's G (log-likelihood ratio)
LEX1<-transform(LEX1,mi=log(o11/e11))            #mutual information
LEX1<-transform(LEX1,oddsratio=log(((o11+.5)*(o22+.5))/((o12+.5)*(o21+.5))))   #odds ratio (theta)
LEX1<-transform(LEX1,dice=(2*o11)/(r1+c1))        #Dice coefficient
LEX1<-transform(LEX1,mf=(o11/e11)/(o12/e12))      #Male-female ratio
LEX1<-transform(LEX1,chisqp.value=pchisq(chisq,1,lower.tail=F))
LEX1$signi<-""
LEX1$signi[which(LEX1$chisqp.value<.05)]<-"*"
LEX1$signi[which(LEX1$chisqp.value<.01)]<-"**"
LEX1$signi[which(LEX1$chisqp.value<.001)]<-"***"
LEX1.1<-LEX1[which(LEX1[,3]>0),]
LEX1.2<-LEX1.1[which(LEX1.1[,4]>0),]                #subset of data; only tokens with at least one occurrence by both males and females
length(LEX1.2[which(LEX1.2$mf<=.8333333333),1])+length(LEX1.2[which(LEX1.2$mf>=1.2),1]) #number of types with substantial difference in male/female use

LEXmf<-LEX1.2[order(-LEX1.2$mf),]
row.names(LEXmf)<-1:length(LEXmf[,1])
xtable(LEXmf[1:20,c(2,21,16,22,23)],digits=c(0,0,0,1,-2,0))
LEXfm<-LEX1.2[order(LEX1.2$mf),]

LEX.gend<-data.frame(LEXmf[c(1:3,5:21),c(2,21)],LEXfm[c(1:19,21),c(2,21)])

LEXtotal<-data.frame(LEXfcfc[c(1:3,5:21),c(2,21)],LEXfccf[c(1:19,21),c(2,21)])
xtable(LEXtotal,digits=c(0,0,0,0,0))

bt<-lapply(df11.3, function(x) grep(paste("^",names(malesbothw1[1:20]),"$",sep="",collapse="|"),x)
df11males.1<-tolower(df11males[,13])
sds<-sapply(df11males.1, function(x) unlist(strsplit(x," ")))
```

```
df11males.2<-strsplit(df11males.1," ")

sharedw<-intersect(malesw1,femalesw1)
sharedw1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1",sharedw)
maleperc<-lapply(sharedw1, function(x) length(unique(df11males[grep(paste("^",x,"$",sep=""),
    df11males.2),2]))/length(unique(df11males[,2])))

###################PoS Finland vs. Comparison
###################
###################

library(stringr)
finlp<-unlist(strsplit(tolower(df11[,12])," "))
compp<-unlist(strsplit(tolower(dfc11[,2])," "))
sort(table(finlp),decreasing=T)[1:20]
sort(table(compp),decreasing=T)[1:20]
finlbothfcp<-table(factor(finlp,levels=intersect(unique(finlp),unique(compp))))
compbothfcp<-table(factor(compp,levels=intersect(unique(finlp),unique(compp))))
(length(compp)/length(finlp))     #tokens comp eng. vs. tokens fin eng.
fincompratio<-finlbothfcp*(length(compp)/length(finlp))/compbothfcp      #fin/comp ratio for PoS,
     controlled for number of male and female tweets

poslist<-readLines(<path to list of Penn Treebank tags with corresponding descriptions>,encoding=
    "UTF-8")
poslist1<-strsplit(poslist,"\t")
poslistpos<-sapply(poslist1,"[",1);poslistpos<-str_trim(poslistpos)
poslistdesc<-sapply(poslist1,"[",2);poslistdesc<-str_trim(poslistdesc)
poslistpos1<-tolower(poslistpos)
names(fincompratio)<-poslistdesc[match(names(fincompratio),poslistpos1)]
names(fincompratio)[24]<-"Quotation mark"

fincompratio.df<-as.data.frame(fincompratio,stringsAsFactors=F)
names(fincompratio.df)<-c("Feature","Finland-Comparison Ratio")

fincompratio.df1<-fincompratio.df[c(1:29,31,33:36,38:39,41),]
fincompratio.df2<-fincompratio.df1[order(-fincompratio.df1[,2]),]
row.names(fincompratio.df2)<-1:37

matchnames<-poslistpos[match(fincompratio.df2[,1],poslistdesc)]
```

```
matchnames[18]<-"'";matchnames[16]<-"PRP."
testcols.m<-match(matchnames,names(df11))
flength.m<-length(unlist(strsplit(df11[,13]," ")))#total no. PoS tags by males
clength.m<-length(unlist(strsplit(dfc11[,2]," ")))#total no. PoS tags by males
propChi.m<-lapply(testcols.m,function(x) chisq.test(matrix(c(sum(df11[,x]),flength.m-sum(df11[,x
    ]),sum(dfc11[,x-10]),clength.m-sum(dfc11[,x-10])),ncol=2)))    #This is a proportions (i.e.
    Chi-Squared adjusted for proportions) test for independent samples using the Yates correction

propChi.m.xs<-sapply(propChi.m,'[[','statistic')
propChi.m.df<-sapply(propChi.m,'[[','parameter')
propChi.m.pv<-sapply(propChi.m,'[[','p.value')


fincompratio.df2$propChi.m.xs<-propChi.m.xs
fincompratio.df2$propChi.m.df<-propChi.m.df
fincompratio.df2$propChi.m.pv<-propChi.m.pv
signi<-which(fincompratio.df2$propChi.m.pv<.05)
vsigni<-which(fincompratio.df2$propChi.m.pv<.01)
hsigni<-which(fincompratio.df2$propChi.m.pv<.001)

fincompratio.df2$signi<-""
fincompratio.df2$signi[which(fincompratio.df2$propChi.m.pv<.05)]<-"*"
fincompratio.df2$signi[which(fincompratio.df2$propChi.m.pv<.01)]<-"**"
fincompratio.df2$signi[which(fincompratio.df2$propChi.m.pv<.001)]<-"***"

xtab.fincompratio.df<-xtable(fincompratio.df)
xtab.fincompratio.df1<-xtable(fincompratio.df1)     #with a few less interesting features removed
xtab.fincompratio.df2<-xtable(fincompratio.df2[,c(1:3,5:6)],digits=c(0,0,2,1,-2,0))

############################PoS by gender
############################
############################
library(stringr)
malesp<-unlist(strsplit(tolower(df12[df12[,56]=="m",12])," "))
femalesp<-unlist(strsplit(tolower(df12[df12[,56]=="f",12])," "))
sort(table(malesp),decreasing=T)[1:20]
sort(table(femalesp),decreasing=T)[1:20]
bothp<-unlist(strsplit(tolower(df12[df12[,56]!="",12])," "))
bothp.sorted<-sort(table(bothp),decreasing=T)
```

```
malesbothp<-table(factor(malesp,levels=names(bothp.sorted)))
femalesbothp<-table(factor(femalesp,levels=names(bothp.sorted)))
(length(femalesp)/length(malesp))     #tokens by females vs. tokens by males

genderratiop<-malesbothp*(length(femalesp)/length(malesp))/femalesbothp      #male/female ratio
    for PoS, controlled for number of male and female tweets

poslist<-readLines(<path to list of Penn Treebank tags with corresponding descriptions>,encoding=
    "UTF-8")
poslist1<-strsplit(poslist,"\t")
poslistpos<-sapply(poslist1,"[",1);poslistpos<-str_trim(poslistpos)
poslistdesc<-sapply(poslist1,"[",2);poslistdesc<-str_trim(poslistdesc)
poslistpos1<-tolower(poslistpos)
names(genderratiop)<-poslistdesc[match(names(genderratiop),poslistpos1)]
names(genderratiop)[27]<-"Quotation mark"

genderratiop.df<-as.data.frame(genderratiop,stringsAsFactors=F)
names(genderratiop.df)<-c("Feature","Male-Female Ratio")

genderratiop.df1<-genderratiop.df[c(1:28,30:32,34:39),]
genderratiop.df2<-genderratiop.df1[order(-genderratiop.df1[,2]),]
row.names(genderratiop.df2)<-1:37

testcols<-c(23, 44, 32, 19, 25, 21, 52, 26, 33, 28, 18, 49, 31, 14, 20, 38, 27, 39, 17, 48, 28,
    22, 51, 42, 29, 46, 36, 47, 30, 45, 53, 54, 43, 50, 35, 40, 34)

mlength<-length(unlist(strsplit(df11males[,12]," ")))#total no. PoS tags by males
flength<-length(unlist(strsplit(df11females[,12]," ")))#total no. PoS tags by males
propChi<-lapply(testcols,function(x) chisq.test(matrix(c(sum(df11males[,x]),mlength-sum(df11males
    [,x]),sum(df11females[,x]),flength-sum(df11females[,x])),ncol=2)))    #This is a proportions
    (i.e. Chi-Squared adjusted for proportions) test for independent samples using the Yates
    correction

chisq.test(matrix(c(sum(df11males[,32]),mlength-sum(df11males[,32]),sum(df11females[,32]),flength
    -sum(df11females[,32])),ncol=2))

matrix(c(sum(df11males[,40]),2211-sum(df11males[,40]),sum(df11females[,40]),5047-sum(df11females
    [,40])),ncol=2)
```

```
propChi.xs<-sapply(propChi,'[[','statistic')
propChi.df<-sapply(propChi,'[[','parameter')
propChi.pv<-sapply(propChi,'[[','p.value')


genderratiop.df2$propChi.xs<-propChi.xs
genderratiop.df2$propChi.df<-propChi.df
genderratiop.df2$propChi.pv<-propChi.pv
signi<-which(genderratiop.df2$propChi.pv<.05)
vsigni<-which(genderratiop.df2$propChi.pv<.01)
hsigni<-which(genderratiop.df2$propChi.pv<.001)

genderratiop.df2$signi<-""
genderratiop.df2$signi[which(genderratiop.df2$propChi.pv<.05)]<-"*"
genderratiop.df2$signi[which(genderratiop.df2$propChi.pv<.01)]<-"**"
genderratiop.df2$signi[which(genderratiop.df2$propChi.pv<.001)]<-"***"

xtab.genderratiop.df<-xtable(genderratiop.df)
xtab.genderratiop.df1<-xtable(genderratiop.df1)     #with a few less interesting features removed
xtab.genderratiop.df2<-xtable(genderratiop.df2[,c(1:3,5:6)],digits=c(0,0,2,1,-2,0))

####################PROFANITY

prof<-readLines(<path to list of profanity>)   #from www.noswearing.com/dictionary/
prof1<-prof[-(which(prof==""))];library(stringr)
prof2<-str_trim(prof1)
prof3<-paste(prof2,"s",sep="")
prof4<-c(prof2,prof3)
df11g<-unlist(strsplit(df11girls[,13]));df11g<-tolower(df11g)
df11b<-unlist(strsplit(df11boys[,13]));df11b<-tolower(df11b)
length(which(df11g %in% prof4))*1000/length(df11g)     #female Finland English profanity
length(which(df11b %in% prof4))*1000/length(df11b)        #male Finland English profanity
df11.3<-unlist(strsplit(df11[,3]," "));df11.3<-tolower(df11.3)  #Comparison English profanity
dfc11.3<-unlist(strsplit(dfc11[,3]," "));dfc11.3<-tolower(dfc11.3)     #Finland English profanity
chisq.test(matrix(c(length(which(df11b %in% prof4)),length(df11b)-length(which(df11b %in% prof4))
    ,length(which(df11g %in% prof4)),length(df11g)-length(which(df11g %in% prof4))),ncol=2))
chisq.test(matrix(c(length(which(df11.3 %in% prof4)),length(df11.3)-length(which(df11.3 %in%
    prof4)),length(which(dfc11.3 %in% prof4)),length(dfc11.3)-length(which(dfc11.3 %in% prof4))),
    ncol=2))
```

```
943
944
945 ############## Emoticons
946 ###############
947
948 finwcap<-unlist(strsplit(df11[,13]," "))
949 fPoS2u<-finwcap[which(finp=="UH")]
950 neg.emot<-grep("[A-Ca-cE-Ne-nQ-Wq-wY-Zy-z]",fPoS2u)     ## all UH tags that contain letters
      except for d and x, which are common in :D and XD smileys
951 fPoS2u.smileys<-sort(table(fPoS2u[-neg.emot]),decreasing=T)
952 nma.ffe<-names(fPoS2u.smileys[c(1   , 2   , 4   , 5   , 6   , 9   , 11 , 16 , 19 , 20 , 21 , 22 , 24 ,
      27 , 29 , 31 , 33 , 35 , 38 , 39   ,43 , 44 ,
953 45 , 46 , 47 , 49 , 50 , 51 , 53 , 55 , 56 , 58 , 59 , 60 , 61 , 64 , 66 , 68 , 69 , 72 , 73 , 77
       , 78 , 80 ,
954 82 , 83 , 84 , 85 , 89 , 90 , 92 , 95 , 101, 107, 108, 109, 110, 111, 113, 114, 117, 119, 120,
      121 ,122, 123,
955 124, 125, 126, 129, 130, 135, 136, 137, 138, 139, 140, 141, 142, 143, 145, 147, 148, 151, 152,
      153 ,154, 155,
956 156, 158, 160, 161, 162, 163, 164, 165, 166, 169, 175, 179, 180, 185, 188, 190, 191, 192, 193,
      196 ,197, 198,
957 202, 203, 204, 209, 210, 211, 214, 215, 216, 217, 227, 233, 234, 248, 254, 259, 260, 261 ,265,
      269,
958 273, 274, 276, 277, 279, 284, 286, 287, 288, 289, 290, 291, 292, 295, 296, 297, 298, 306, 307,
      308 ,309, 310,
959 311, 312, 314, 316, 319, 320, 322, 323, 324, 332, 334, 335, 336, 337, 340, 345, 348, 349, 352,
      370 ,371, 372,
960 373, 374, 375, 376, 377, 378, 379, 382, 383, 384, 385, 387, 388, 389, 429, 440, 3, 15, 17, 30,
      32, 34, 42, 63, 65, 76, 91, 100, 112,128,133,157,199,220,228,229,230,231, 232, 303, 313, 338,
       381, 415, 417, 418, 419,
      420,421,422,423,424,434,439,441,442,443,444,445,446,18,57,81,321,450,451)])
961 nma.ffe1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", nma.ffe) #escape regex so control
      characters can be used in grep
962
963 uC<-which(compp=="UH")
964 compwcap<-unlist(strsplit(dfc11[,3]," "))
965 cPoS2u<-compwcap[which(compp=="UH")]
966 #cPoSD2u<-cPoSD2[which(cPoSD2[,2]=="E"),]        #note that default tagger has very different
      output than Penn tagger
967
```

```
neg.emotC<-grep("[A-Ca-cE-Ne-nQ-Wq-wY-Zy-z]",cPoS2u)
emot.smileys.C<-cPoS2u[-neg.emotC]
emot.smileys.C.nma.e<-cPoS2u[(grep(paste("^",nma.e1,"$",collapse="|"),cPoS2u))]
emot.smileys.C.u<-unique(emot.smileys.C)
emot.smileys.C.u1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", emot.smileys.C.u)
cPoS2u.smileys<-sort(table(cPoS2u[-neg.emotC]),decreasing=T)

##### Getting coordinates of all tweets with at least one emoticon/smiley
lon.emot.smileys<-df11[-neg.emot,9];lat.emot.smileys<-df11[-neg.emot,10]
lonlat.emot.smileys<-data.frame(lon.emot.smileys,lat.emot.smileys)
coordinates(lonlat.emot.smileys)<-c(1,2) #creates Spatial Points Frame
proj4string(lonlat.emot.smileys)<-CRS(" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84
    =0,0,0") #use the proj4string of the gadm projection
lonlat.emot.smileys<-spTransform(lonlat.emot.smileys,CRS=CRS(" +proj=longlat +ellps=WGS84 +datum=
    WGS84 +no_defs +towgs84=0,0,0"))#as above
over.emot.smileys<-over(lonlat.emot.smileys,as(fin.adm2.spdf,"SpatialPolygons"))
over.emot.smileys.f<-factor(over.emot.smileys,levels=as.character(c(seq(1:21))))       #Necessary
     because some polygons zero frequency
over.emot.smileys.f.df<-data.frame(table(over.emot.smileys.f))
nrtweets.emot.smileys.f = over.emot.smileys.f.df$Freq
nrtweets.emot.smileys.f.df<-data.frame(fin.adm2.spdf$NAME_2,nrtweets.emot.smileys.f)
names(nrtweets.emot.smileys.f.df)[1]<-"id"

##### Emoticons in Finland English Corpus
fPoS2u.s<-fPoS2u.smileys[intersect(names(fPoS2u.smileys),nma.ffe)][1:20]
nm4<-names(fPoS2u.s)
nm4.1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", nm4)
for(i in 1:20){assign(paste("emot20.names.", i, sep = ""),grep(paste("^", nm4.1[i], "$",sep=""),
    fPoS2u))}  #freqs of top 20 smileys to variables
for(i in 1:406){assign(paste("emot.smileys.u.", i, sep = ""),grep(paste("^", emot.smileys.u1[i],
    "$",sep=""),fPoS2u))}  #freqs of top 406 smileys to variables

#Getting coordinates of each emoticon
for(i in 1:20){assign(paste("lon.emot20.names.", i, sep = ""),df11[eval(parse(text=paste("emot20.
    names.",i,sep=""))),9])}
for(i in 1:20){assign(paste("lat.emot20.names.", i, sep = ""),df11[eval(parse(text=paste("emot20.
    names.",i,sep=""))),10])}
for(i in 1:20){assign(paste("lonlat.emot20.names.", i, sep = ""),data.frame(eval(parse(text=paste
    ("lon.emot20.names.",i,sep="")),eval(parse(text=paste("lat.emot20.names.",i,sep=""))))))}
```

```
for(i in 1:20){eval(parse(text=paste("coordinates(lonlat.emot20.names.",i,")",sep="","<-c(1,2)"))
    )}
#Setting the proj4string parameter (map projection) to the same as that of the Finland map data
    gadm
for(i in 1:20){eval(parse(text=paste("proj4string(lonlat.emot20.names.",i,")",sep="","<-","CRS(\"
     +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84=0,0,0\")")))}
#putting the points into the map coordinates
for(i in 1:20){eval(parse(text=paste("lonlat.emot20.names.",i,sep="","<-","spTransform(lonlat.
    emot20.names.",i,",CRS=CRS(\" +proj=longlat +ellps=WGS84 +datum=WGS84 +no_defs +towgs84
    =0,0,0\"))")))}
#Overlay of the points within the map polygons
for(i in 1:20){eval(parse(text=paste("over.emot20.names.", i, sep = "","<-","over(lonlat.emot20.
    names.",i,",as(fin.adm2.spdf,\"SpatialPolygons\"))")))}
#Necessary because some polygons zero frequency
for(i in 1:20){eval(parse(text=paste("over.emot20.names.", i,".f",sep = "","<-","factor(over.
    emot20.names.",i,",levels=as.character(c(seq(1:21))))")))}
for(i in 1:20){eval(parse(text=paste("over.emot20.names.", i,".f.df",sep = "","<-","data.frame(
    table(over.emot20.names.",i,".f))")))}
for(i in 1:20){eval(parse(text=paste("nrtweets.emot20.names.",i,".f",sep="","<-","over.emot20.
    names.",i,".f.df$Freq")))}
for(i in 1:20){eval(parse(text=paste("nrtweets.emot20.names.",i,".f.df",sep="","<-","data.frame(
    fin.adm2.spdf$NAME_2,nrtweets.emot20.names.",i,".f)")))}
for(i in 1:20){eval(parse(text=paste("names(nrtweets.emot20.names.",i,".f.df)[1]",sep="","<-","\"
    id\"")))}

####### Merging emoticon frequency per region with other regional frequencies
demog.dat2<-data.frame(demog.dat1,nrtweets.emot20.names.1.f,nrtweets.emot20.names.2.f,nrtweets.
    emot20.names.3.f,nrtweets.emot20.names.4.f,nrtweets.emot20.names.5.f,
nrtweets.emot20.names.6.f,nrtweets.emot20.names.7.f,nrtweets.emot20.names.8.f,nrtweets.emot20.
    names.9.f,nrtweets.emot20.names.10.f,nrtweets.emot20.names.11.f,
nrtweets.emot20.names.12.f,nrtweets.emot20.names.13.f,nrtweets.emot20.names.14.f,nrtweets.emot20.
    names.15.f,nrtweets.emot20.names.16.f,nrtweets.emot20.names.17.f,
nrtweets.emot20.names.18.f,nrtweets.emot20.names.19.f,nrtweets.emot20.names.20.f,nrtweets.emot.
    smileys.f)
demog.dat2$emot20<-rowSums(demog.dat2[1:21,27:46])
frame.all<-merge(fin.adm2.df,demog.dat2,by.y="id",all.x=T)

##Proportion of Tweets with at least one of top 20 most common smilies
```

```
p.smileysP <- ggplot(frame.all, aes(x = long, y = lat, group = group)) + geom_polygon(aes(fill =
    nrtweets.emot.smileys.f*100/nrtweets.all)) +
geom_text(data = fin.adm2.centroids.df, aes(label = round(demog.dat2$nrtweets.emot.smileys.f*100/
    demog.dat2$nrtweets.all,digits=2), x = long, y = lat, group = NAME_2),dig.lab=3, size = 2.2)
    +
labs(x=" ", y=" ") + scale_fill_gradient("In Percent",low = "#ffffcc", high = "#ff4444",space = "
    Lab", na.value = "grey80", guide = "colourbar") + coord_map() + ggtitle("Proportion of tweets
     with at least one\n of the 20 most frequent emoticons")


#################CORRELATION WITH SOME DEMOGRAPHIC DATA
pctweets<-demog.dat2$nrtweets.emot.smileys.f*100/demog.dat1$nrtweets.all
region.names<-as.character(demog.dat2[,1])
demog.dat4<-demog.dat2[,c(2,3,4,5,8,9,10,11,12,19,20,21,22,23,24,25,26,27,28,29,30,55)]
demog.dat5<-sapply(demog.dat4[,1:22],function(x) as.numeric(x))
demog.dat6<-data.frame(region.names,demog.dat5,stringsAsFactors=F)
#demog.dat7<-demog.dat6[2/demog.dat6[,3]
smiley.cor<-sapply(demog.dat6[,2:23],function(x) cor((demog.dat6[,23]/demog.dat6[,3]),x))
smiley.cor.test<-sapply(demog.dat6[,2:23],function(x) cor.test((demog.dat6[,23]/demog.dat6[,3]),x
    ))

smiley.cor.df<-data.frame(unlist(smiley.cor.test[4,]),unlist(smiley.cor.test[1,]),unlist(smiley.
    cor.test[3,]))
names(smiley.cor.df)<-c("Spearman's rho","test statistic","p-value")
smiley.cor.df1<-smiley.cor.df[order(-smiley.cor.df[,1]),]
smiley.cor.df2<-smiley.cor.df1[c(1,3:8,10:14,21),]
rownames(smiley.cor.df2)<-c("Russian tweets","English tweets","All tweets","Other language tweets
    ","Finnish tweets","University degrees 2012","GDP 2012","Number of Pupils","Polytechnical
    degrees","Number of high school leavers","Swedish tweets","Vocational qualifications","Land
    area")
smiley.cor.df2$signi<-""
smiley.cor.df2$signi[which(smiley.cor.df2[,3]<.05)]<-"*"
smiley.cor.df2$signi[which(smiley.cor.df2[,3]<.01)]<-"**"
smiley.cor.df2$signi[which(smiley.cor.df2[,3]<.001)]<-"***"
xtable(smiley.cor.df2,digits=c(0,2,2,-2,0))

##Proportion of tweets that include top 20 smileys
for(i in 1:20){eval(parse(text=paste("p.emot20.names.",i,"P",sep="","<-","ggplot(frame.all, aes(x
     = long, y = lat, group = group)) + geom_polygon(aes(fill =nrtweets.emot20.names.",i,".f*100/
```

```
nrtweets.all))+geom_text(data=fin.adm2.centroids.df,aes(label=round(demog.dat2$nrtweets.
emot20.names.",i,".f*100/demog.dat2$nrtweets.all,digits=2),x=long,y=lat,group=NAME_2),dig.lab
=3,size=2.2)")))}
for(i in 1:20){eval(parse(text=paste("p.emot20.names.",i,"P",sep="","<-","p.emot20.names.",i,"P+
labs(x=NULL, y=NULL)+scale_fill_gradient(\"In Percent\",low=\"#ffffcc\",high=\"#ff4444\",
space=\"Lab\",na.value=\"grey80\",guide=\"colourbar\")")))}
for(i in 1:20){eval(parse(text=paste("p.emot20.names.",i,"P",sep="","<-","p.emot20.names.",i,"P+
coord_map()")))}
for(i in 1:20){assign(paste("lon.emot20.names.", i, sep = ""),df11[eval(parse(text=paste("emot20.
names.",i,sep="")))),9])}
for(i in 1:20){assign(paste("p.emot20.names.s",i,sep=""),nm4[i])}
for(i in 1:20){eval(parse(text=paste("p.emot20.names.",i,"P",sep="","<-","p.emot20.names.",i,"P+
theme(legend.key.width=unit(.5, \"cm\"),legend.text=element_text(size=12),legend.title =
element_text(size=12),axis.text.y = element_blank(),
axis.text.x = element_blank(),axis.ticks =element_blank(),legend.margin = unit(0, \"cm\"),plot.
margin = unit(c(0,0,0,0), \"cm\"),plot.title=element_text(size=18))+ggtitle(p.emot20.names.s"
,i,")")))}

matchesffe<-df11.uhs[(grep(paste(nma.ffe1,collapse="|"),df11.uhs[,1])),] #finding tweets with
emoticons
matchesffea<-df11[(grep(paste(nma.ffe1,collapse="|"),df11[,1])),]   #there are many words with xp
 or xo in them
matches.allfe<-df9.17[(grep(paste(nma.ffe1,"\\s|",nma.ffe1,"$",sep="",collapse="|"),df9.17[,1]))
,]

length(unique(matches.allfe[,2]))/length(unique(df9.17[,2]))  #prop. of all Finland Tweeters who
used at least one emoticon
length(matches.allfe[,1])/length(df9.17[,1])  #prop. of all Finland tweets with an emoticon

fPoS2u.smileys[intersect(nma.ffe,names(fPoS2u.smileys))][1:20] #Fin.En. emoticons top 20 most
freq

sort(cPoS2u.smileys[intersect(nma.ffe,names(cPoS2u.smileys))],decreasing=T)[1:20]

sum(fPoS2u.smileys[intersect(names(fPoS2u.smileys),nma.ffe)][1:20])/sum(fPoS2u.smileys[intersect(
nma.ffe,names(fPoS2u.smileys))])    ### top 20 as percent of all FE emoticons

topsFin<-sort(fPoS2u.smileys[intersect(names(fPoS2u.smileys),nma.ffe)],decreasing=T)[1:20]
```

```
topsFin.df<-data.frame(topsFin,round(100*unname(topsFin)/sum(fPoS2u.smileys[intersect(names(
    fPoS2u.smileys),nma.ffe)]),digits=1))
names(topsFin.df)<-c("Type","Percent of all emoticons")

topsComp<-sort(cPoS2u.smileys[intersect(nma.ffe,names(cPoS2u.smileys))],decreasing=T)[1:20]
topsComp.df<-data.frame(topsComp,round(100*unname(topsComp)/sum(cPoS2u.smileys[intersect(nma.ffe,
    names(cPoS2u.smileys))]),digits=1))
names(topsFin)<-c("Type","Percent of all emoticons")
names(topsComp)<-c("Type","Percent of all emoticons")

dfc11.uhs<-dfc11[grep("UH",dfc11[,2]),]
matchesC<-dfc11.uhs[(grep(paste(nma.ffe1,collapse="|"),dfc11.uhs[,1])),]  # This makes a frame
    with those with an "UH" tag that were emoticons
matchesCa<-dfc11[(grep(paste(nma.ffe1,collapse="|"),dfc11[,1])),]
matches.allC<-dfc6[(grep(paste(nma.ffe1,collapse="|"),dfc6[,1])),]

length(matchesC[,1])/length(dfc11[,1])   #This provides prop. of English tweets in comparison
    corpus with at least one emoticon
length(matches.allC[,1])/length(dfc6[,1])  #prop. of all Comparison tweets with an emoticon
#sum(matchesC$UH)/length(dfc11$V1)
length(unique(matchesC[,2]))/length(unique(dfc11[,2]))
sum(sort(cPoS2u.smileys[intersect(nma.ffe,names(cPoS2u.smileys))],decreasing=T)[1:20])/sum(cPoS2u
    .smileys[intersect(nma.ffe,names(cPoS2u.smileys))])    ### top 20 as percent of all CE
    emoticons


fPoS2u.s<-fPoS2u.smileys[intersect(names(fPoS2u.smileys),nma.ffe)][1:20]          #Top 20
    emoticons using Penn Treebank tags and own filtering steps
for(b in 1:20){cat(b,names(fPoS2u.s[b]),fPoS2u.s[b],fPoS2u.s[b]*100/sum(fPoS2u.smileys),"\n",sep=
    "\t")}
cPoS2u.s<-cPoS2u.smileys[intersect(names(cPoS2u.smileys),nma.ffe)][1:20]           #Top 20
    emoticons using Penn Treebank tags and own filtering steps
#Using Schnoebelen's (2012) set of 28 emoticons
schnoeb.names<-c(":)",";)",":(",":D",":-)",":P","(:",":/","XD","=)",";-)",":O",";D",":))",":-(","
    :')","D:",":]","=D",":-/","=]",";P",":-P",":'(",":-D","=/","=P","=(")
schnoeb.prop<c(.396, .105, .083, .075, .049, .045, .041, .034, .03, .021, .019, .016, .009, .008,
     .007, .006, .006, .006, .006, .005, .005, .005, .004, .004, .004, .004, .004, .004)
#schnoeb.names<-tolower(schnoeb.names)
schnoeb.fin<-intersect(fPoS2u,schnoeb.names)
```

```
schnoeb.fin1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1",schnoeb.fin)
schnoeb.fin.freq<-fPoS2u[grep(paste("^",schnoeb.fin1,"$",sep="",collapse="|"),fPoS2u)]
fPoS2u.s1<-sort(table(factor(schnoeb.fin.freq,levels=schnoeb.names)),decreasing=T)[1:28]
schnoeb.cm<-intersect(cPoS2u,schnoeb.names)
schnoeb.cm1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1",schnoeb.cm)
schnoeb.cm.freq<-cPoS2u[grep(paste("^",schnoeb.cm1,"$",sep="",collapse="|"),cPoS2u)]
cPoS2u.s28<-cPoS2u.smileys[intersect(names(cPoS2u.smileys),nma.ffe)][1:28]      #Using the
    default model for the CMU PoS Tagger
for(b in 1:28){cat(b,names(cPoS2u.s28[b]),cPoS2u.s28[b],cPoS2u.s28[b]*100/sum(cPoS2u.s28),"\n",
    sep="\t")}    #"
cPoS2u.s1<-sort(table(factor(schnoeb.cm.freq,levels=schnoeb.names)),decreasing=T)[1:28]     #
    Using Schnoebelen's (2012) set of 28 emoticons
for(b in 1:28){cat(b,names(cPoS2u.s1[b]),cPoS2u.s1[b]*100/sum(cPoS2u.s1),"\n",sep="\t")}
for(i in 1:20){assign(paste("emot20.names.", i, sep = ""),grep(paste("^", nm3[i], "$",sep=""),
    fPoS2u))}
fPoS2u.s3<-table(factor(schnoeb.fin.freq,levels=schnoeb.names))/sum(table(factor(schnoeb.fin.freq
    ,levels=schnoeb.names)))
cPoS2u.s3<-table(factor(schnoeb.cm.freq,levels=schnoeb.names))/sum(table(factor(schnoeb.cm.freq,
    levels=schnoeb.names)))

####### Significance testing
ks.test(jitter(fPoS2u.s3),jitter(cPoS2u.s3))       #Kolmogorov-Smirnov for Fininsh English vs.
    Comparison English
ks.test(jitter(fPoS2u.s3),jitter(prop.table(schnoeb.prop))) #Fininsh English vs. Schnoebelen 2012
ks.test(jitter(cPoS2u.s3),jitter(prop.table(schnoeb.prop)))#Comparison English vs. Schnoebelen
    2012
wilcox.test(jitter(fPoS2u.s3),jitter(cPoS2u.s3),paired=T)       #Wilcoxon ranked sums test for
    Fininsh English vs. Comparison English
wilcox.test(jitter(fPoS2u.s3),jitter(prop.table(schnoeb.prop)),paired=T) #Fininsh English vs.
    Schnoebelen 2012
wilcox.test(jitter(cPoS2u.s3),jitter(prop.table(schnoeb.prop)),paired=T)#Comparison English vs.
    Schnoebelen 2012

smiley.props<-rbind(schnoeb.prop,fPoS2u.s3,cPoS2u.s3)
smiley.propsP<-barplot(smiley.props,beside=T,ylab="Frequency per Thousand Words",cex.names=.7)
legend("topright",legend=c("Schnoebelen 2012","Finland English Corpus","Comparison Corpus"),col=
    gray.colors(3),fill=gray.colors(3))
title("Relative Frequencies of Most Common Emoticons in Schnoebelen 2012")
####Nose correlation
```

```
nma.ff.allnoses<-nma.ffe[c
    (5,12:14,17,64,86,89,102,104:106,109,130,132,137:143,148,150:151,157,161,194,227,237:238)]
nma.ff.allnoses1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", nma.ff.allnoses)            #Regex
     to handle escape characters in R; top ten noses
nma.ff.dashnoses<-nma.ff.allnoses[c(1,3,5:7,10:12,18:22,26,30:31)]
nma.ff.dashnoses1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", nma.ff.dashnoses)       ##top ten
     dashnoses
nma.ff.aponoses<-nma.ff.allnoses[c(2,4,13:14,16:17,29)]                                  #top six
     aponoses
nma.ff.aponoses1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", nma.ff.aponoses)
nma.ff.asian<-nma.ffe[c(8, 20, 34, 42:44, 49:50, 54, 58:59, 61:63, 65, 68:69, 72, 80, 91, 95,
    100:101, 110:113, 116, 121, 124:125, 127:129, 135, 158, 162:164, 167:171, 182, 186:189)]
                #top ten Asian smileys
nma.ff.asian1<-gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", nma.ff.asian)
FnosetweetsfPoS<-lapply(nma.ff.allnoses1,function(y) grep(paste("^",y,"$",sep=""),finwcap))
FnosetweetsfPoS1<-unlist(FnosetweetsfPoS)


FdashnosetweetsfPoS<-lapply(nma.ff.dashnoses1,function(y) grep(paste("^",y,"$",sep=""),finwcap))
FdashnosetweetsfPoS1<-unlist(FdashnosetweetsfPoS)

FaponosetweetsfPoS<-lapply(nma.ff.aponoses1,function(y) grep(paste("^",y,"$",sep=""),finwcap))
FaponosetweetsfPoS1<-unlist(FaponosetweetsfPoS)

FasiantweetsfPoS<-lapply(nma.ff.asian1,function(y) grep(paste("^",y,"$",sep=""),finwcap))
FasiantweetsfPoS1<-unlist(FasiantweetsfPoS)


Fnosetweets.tab<-sort(table(finwcap[FnosetweetsfPoS1]),decreasing=T)
Fdashnosetweets.tab<-sort(table(finwcap[FdashnosetweetsfPoS1]),decreasing=T)
Faponosetweets.tab<-sort(table(finwcap[FaponosetweetsfPoS1]),decreasing=T)
Fasiantweets.tab<-sort(table(finwcap[FasiantweetsfPoS1]),decreasing=T)


CnosetweetscPoS<-lapply(nma.ff.allnoses1,function(y) grep(paste("^",y,"$",sep=""),compwcap))
CnosetweetscPoS1<-unlist(CnosetweetscPoS)

CdashnosetweetscPoS<-lapply(nma.ff.dashnoses1,function(y) grep(paste("^",y,"$",sep=""),compwcap))
CdashnosetweetscPoS1<-unlist(CdashnosetweetscPoS)
```

```
1159
1160 CaponosetweetscPoS<-lapply(nma.ff.aponoses1,function(y) grep(paste("^",y,"$",sep=""),compwcap))
1161 CaponosetweetscPoS1<-unlist(CaponosetweetscPoS)
1162
1163 CasiantweetscPoS<-lapply(nma.ff.asian1,function(y) grep(paste("^",y,"$",sep=""),compwcap))
1164 CasiantweetscPoS1<-unlist(CasiantweetscPoS)
1165
1166 Cnosetweets.tab<-sort(table(compwcap[CnosetweetscPoS1]),decreasing=T)
1167 Cdashnosetweets.tab<-sort(table(compwcap[CdashnosetweetscPoS1]),decreasing=T)
1168 Caponosetweets.tab<-sort(table(compwcap[CaponosetweetscPoS1]),decreasing=T)
1169 Casiantweets.tab<-sort(table(compwcap[CasiantweetscPoS1]),decreasing=T)
1170
1171 allnosenames<-intersect(names(Cnosetweets.tab),names(Fnosetweets.tab))
1172 dashnosenames<-intersect(names(Cdashnosetweets.tab),names(Fdashnosetweets.tab))
1173 aponosenames<-intersect(names(Caponosetweets.tab),names(Faponosetweets.tab))
1174 asiannames<-intersect(names(Casiantweets.tab),names(Fasiantweets.tab))
1175
1176
1177 f.allnosenames<-table(factor(finwcap[FnosetweetsfPoS1],levels=allnosenames))/sum(fPoS2u.smileys[
        intersect(names(fPoS2u.smileys),nma.ffe)])
1178 f.dashnosenames<-table(factor(finwcap[FdashnosetweetsfPoS1],levels=dashnosenames))/sum(fPoS2u.
        smileys[intersect(names(fPoS2u.smileys),nma.ffe)])
1179 f.aponosenames<-table(factor(finwcap[FaponosetweetsfPoS1],levels=aponosenames))/sum(fPoS2u.
        smileys[intersect(names(fPoS2u.smileys),nma.ffe)])
1180 f.asiannames<-table(factor(finwcap[FasiantweetsfPoS1],levels=asiannames))/sum(fPoS2u.smileys[
        intersect(names(fPoS2u.smileys),nma.ffe)])
1181
1182 c.allnosenames<-table(factor(compwcap[CnosetweetscPoS1],levels=allnosenames))/sum(cPoS2u.smileys[
        intersect(names(cPoS2u.smileys),nma.ffe)])
1183 c.dashnosenames<-table(factor(compwcap[CdashnosetweetscPoS1],levels=dashnosenames))/sum(cPoS2u.
        smileys[intersect(names(cPoS2u.smileys),nma.ffe)])
1184 c.aponosenames<-table(factor(compwcap[CaponosetweetscPoS1],levels=aponosenames))/sum(cPoS2u.
        smileys[intersect(names(cPoS2u.smileys),nma.ffe)])
1185 c.asiannames<-table(factor(compwcap[CasiantweetscPoS1],levels=asiannames))/sum(cPoS2u.smileys[
        intersect(names(cPoS2u.smileys),nma.ffe)])
1186
1187 sum(c.dashnosenames)/sum(f.dashnosenames)   ### difference ratio in rate per thousand words
1188 sum(f.aponosenames)/sum(c.aponosenames)
1189 sum(f.asiannames)/sum(c.asiannames)
```

```
1190
1191 wilcox.test(jitter(f.allnosenames),jitter(c.allnosenames),paired=T)       #Wilcoxon ranked sums
       test for Fininsh English vs. Comparison English
1192 wilcox.test(jitter(f.dashnosenames),jitter(c.dashnosenames),paired=T) #Fininsh English vs.
       Comparison English
1193 wilcox.test(jitter(f.aponosenames),jitter(c.aponosenames),paired=T)#Fininsh English vs.
       Comparison English
1194 wilcox.test(jitter(f.asiannames),jitter(c.asiannames),paired=T)#Fininsh English vs. Comparison
       English
1195
1196 ks.test(jitter(f.dashnosenames),jitter(c.dashnosenames))
1197 f.dashnosenames[1]
1198
1199 dnosetest<-matrix(c(length(finwcap[FdashnosetweetsfPoS1]),sum(fPoS2u.smileys[intersect(names(
       fPoS2u.smileys),nma.ffe)])-length(finwcap[FdashnosetweetsfPoS1]),length(compwcap[
       CdashnosetweetscPoS1]),sum(cPoS2u.smileys[intersect(names(cPoS2u.smileys),nma.ffe)])-length(
       compwcap[CdashnosetweetscPoS1])),ncol=2)
1200 chisq.test(dnosetest)
1201
1202 aponosetest<-matrix(c(length(finwcap[FaponosetweetsfPoS1]),sum(fPoS2u.smileys[intersect(names(
       fPoS2u.smileys),nma.ffe)])-length(finwcap[FaponosetweetsfPoS1]),length(compwcap[
       CaponosetweetscPoS1]),sum(cPoS2u.smileys[intersect(names(cPoS2u.smileys),nma.ffe)])-length(
       compwcap[CaponosetweetscPoS1])),ncol=2)
1203 chisq.test(aponosetest)
1204
1205 asiannamestest<-matrix(c(length(finwcap[FasiantweetsfPoS1]),sum(fPoS2u.smileys[intersect(names(
       fPoS2u.smileys),nma.ffe)])-length(finwcap[FasiantweetsfPoS1]),length(compwcap[
       CasiantweetscPoS1]),sum(cPoS2u.smileys[intersect(names(cPoS2u.smileys),nma.ffe)])-length(
       compwcap[CasiantweetscPoS1])),ncol=2)
1206 chisq.test(asiannamestest)
1207
1208 emot.gend<-lapply(nma.ffe1,function(y) grep(y,df12[,1]))         #which of the 201 top emoticons
       are in tweets where m and f are tagged
1209 emot.gend1<-unlist(emot.gend)
1210 (length(which(df12[emot.gend1,56]=="f"))/length(which(df12[emot.gend1,56]=="m")))/(length(which(
       df11[,56]=="f"))/length(which(df11[,56]=="m")))        #ratio emoticon tweets female/male
1211
```

```
chisq.test(matrix(c(length(which(df12[emot.gend1,56]=="f")),length(df11girls[,1])-length(which(df12[emot.gend1,56]=="f")),length(which(df12[emot.gend1,56]=="m")),length(df11boys[,1])-length(which(df12[emot.gend1,56]=="m"))),ncol=2))
    #Chi-squared observed-expected test for female and male use of emoticons

(length(grep("'",df11girls[,1]))/length(grep("'",df11boys[,1])))/(length(which(df11[,56]=="f"))/length(which(df11[,56]=="m")))    #ratio apostrophes tweets female/male

chisq.test(matrix(c(length(grep("'",df11girls[,1])),length(df11girls[,1])-length(grep("'",df11girls[,1])),length(grep("'",df11boys[,1])),length(df11boys[,1])-length(grep("'",df11boys[,1]))),ncol=2))
df12f<-strsplit(df12[which(df12[,56]=="f"),13]," ")
#lapply(df12[which(df12[,56]=="f"),13],function(x) strsplit(x," "))
df12m<-strsplit(df12[which(df12[,56]=="m"),13]," ")
df12f.1<-lapply(df12f,function(x) gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", x))
df12m.1<-lapply(df12m,function(x) gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", x))

liv<-nma.ffe1
fems<-do.call(rbind,lapply(df12f.1,function(x,liv){table(factor(x,levels = nma.ffe1,ordered = TRUE))},liv = liv))
mals<-do.call(rbind,lapply(df12m.1,function(x,liv){table(factor(x,levels = nma.ffe1,ordered = TRUE))},liv = liv))

sum(fems)*1000/length(unlist(df12f))  #female emoticons per thousand words
sum(mals)*1000/length(unlist(df12m))  #male emoticons per thousand words
t.test(rowSums(fems),rowSums(mals))     #avg. emoticons per tweet, males and females, t-test

df12a<-strsplit(df12[,13]," ")
df12a.1<-lapply(df12a,function(x) gsub("([.|()\\^{}+$*?]|\\[|\\])", "\\\\\\1", x))
gend.dash<-do.call(rbind,lapply(df12a.1,function(x,liv){table(factor(x,levels = nma.ff.dashnoses1,ordered = TRUE))},liv = liv))
gend.apo<-do.call(rbind,lapply(df12a.1,function(x,liv){table(factor(x,levels = nma.ff.aponoses1,ordered = TRUE))},liv = liv))
gend.asian<-do.call(rbind,lapply(df12a.1,function(x,liv){table(factor(x,levels = nma.ff.asian1,ordered = TRUE))},liv = liv))

gend.dashnosetweets<-lapply(nma.ff.dashnoses1,function(y) grep(y,df12[,1]))
gend.dashnosetweets1<-unlist(gend.dashnosetweets)
```

```
gend.aponosetweets<-lapply(nma.ff.aponoses1,function(y) grep(y,df12[,1]))
gend.aponosetweets1<-unlist(gend.aponosetweets)

gend.asiantweets<-lapply(nma.ff.asian1,function(y) grep(y,df12[,1]))
gend.asiantweets1<-unlist(gend.asiantweets)

(length(which(df12[gend.dashnosetweets1,56]=="f"))/length(which(df12[gend.dashnosetweets1,56]=="m")))/(length(which(df11[,56]=="f"))/length(which(df11[,56]=="m")))    #adjusted female to male dash nose emoticons

dashnose.gender.test<-matrix(c(length(which(df12[gend.dashnosetweets1,56]=="m")),length(df12[df12[,56]=="m",1])-length(which(df12[gend.dashnosetweets1,56]=="m")),length(which(df12[gend.dashnosetweets1,56]=="f")),length(df12[df12[,56]=="f",1])-length(which(df12[gend.dashnosetweets1,56]=="f"))),ncol=2)
chisq.test(dashnose.gender.test)

(length(which(df12[gend.aponosetweets1,56]=="f"))/length(which(df12[gend.aponosetweets1,56]=="m")))/(length(which(df11[,56]=="f"))/length(which(df11[,56]=="m")))      #adjusted female to male apostrophe nose emoticon ratio

aponose.gender.test<-matrix(c(length(which(df12[gend.aponosetweets1,56]=="m")),length(df12[df12[,56]=="m",1]),length(which(df12[gend.aponosetweets1,56]=="f")),length(df12[df12[,56]=="f",1])),ncol=2,nrow=2)
chisq.test(aponose.gender.test)

(length(which(df12[gend.asiantweets1,56]=="f"))/length(which(df12[gend.asiantweets1,56]=="m")))/(length(which(df11[,56]=="f"))/length(which(df11[,56]=="m")))

asian.gender.test<-matrix(c(length(which(df12[gend.asiantweets1,56]=="m")),length(df12[df12[,56]=="m",1]),length(which(df12[gend.asiantweets1,56]=="f")),length(df12[df12[,56]=="f",1])),ncol=2)
chisq.test(asian.gender.test)


################## Non-standard Orthography
##################
mis3<-readLines(<path to list of non-standard orthography items>)
correct3<-readLines(<path to list of corresponding standard orthography items>)

```

```
errorsF<-intersect(finw,mis3)       #error types present in Finland English data
errorsC<-intersect(compw,mis3)       #error types present in Comparison English data
correctF<-intersect(finw,correct3)
correctC<-intersect(compw,correct3)

commonEr.w<-union(errorsF,errorsC)
commonCor.w<-union(correctF,correctC)
finEr<-finw[which(finw %in% commonEr.w)]
compEr<-compw[which(compw %in% commonEr.w)]
finCor<-finw[which(finw %in% commonCor.w)]
compCor<-compw[which(compw %in% commonCor.w)]

finErT<-sort(table(factor(finEr,levels=commonEr.w)),decreasing=T)
compErT<-sort(table(factor(compEr,levels=commonEr.w)),decreasing=T)
finCorT<-sort(table(factor(finCor,levels=commonCor.w)),decreasing=T)
compCorT<-sort(table(factor(compCor,levels=commonCor.w)),decreasing=T)

sum(finErT)/sum(finCorT)       #Errors per 1000 words of the wikipedia list  Finland English
sum(compErT)/sum(compCorT)

sum(finErT)*1000/length(finw)       #Errors per 1000 words of the wikipedia list  Finland English
sum(compErT)*1000/length(compw)      #Errors per 1000 words of the wikipedia list  Comparison
    English
chisq.test(matrix(c(sum(finErT),length(finw)-sum(finErT),sum(compErT),length(compw)-sum(compErT))
    ,ncol=2))  #Significance test

commonEr<-rbind(finErT[1:20]*1000/length(finw),compErT[1:20]*1000/length(compw))

commonErP<-barplot(commonEr,beside=T,las=2,ylim=c(0,.7),ylab="Frequency per Thousand Words")
legend("topright",legend=c("Finland English Corpus","Comparison Corpus"),cex=1,col=gray.colors(2)
    ,fill=gray.colors(2))
title("Frequency of Commonly Misspelled Words in the Material")
text(commonErP,commonEr+.02,labels=round(commonEr,digits=2),cex=0.3,col="black")


#####Misspellings by Gender (Finland)
#####################################

df12.1m<-df12[which(df12[,56]=="m"),]
```

```
df12.1f<-df12[which(df12[,56]=="f"),]
#df12.10<-data.frame(df12.1,df12[2:53])
df12.2m<-strsplit(df12.1m[,13]," ")
df12.2f<-strsplit(df12.1f[,13]," ")
df12.3m<-tolower(unlist(df12.2m))
df12.3f<-tolower(unlist(df12.2f))

df12.2mPoS<-strsplit(df12.1m[,12]," ")
df12.2fPoS<-strsplit(df12.1f[,12]," ")
df12.3mPoS<-tolower(unlist(df12.2mPoS))
df12.3fPoS<-tolower(unlist(df12.2fPoS))
#Fwords<-sort(unique(unlist(df12.2)))
errorsFl<-tolower(errorsF)
df12.4m<-sort(table(df12.3m[which(df12.3m %in% errorsFl)]),decreasing=T)
df12.4f<-sort(table(df12.3f[which(df12.3f %in% errorsFl)]),decreasing=T)

length(df12.4m)*length(df11girls[,1])/length(df12.4f)*length(df11boys[,1])
sum(df12.4m)*1000/length(df12.3m)       ##Male frequency  misspellings per 1000 words
sum(df12.4f)*1000/length(df12.3f)    ##Female  frequency  misspellings per 1000 words

chisq.test(matrix(c(sum(df12.4m),length(df12.3m),sum(df12.4f),length(df12.3f)),ncol=2))
#Significance test, misspellings by gender

################## Expressive Lengthening
##################

fPoS2l<-finw
cPoS2l<-compw
for(i in 1:26){assign(paste("F.letters.3.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{3}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}       ##assigns every
    instance of 3 or more letters in a sequence to a variable
#regex gets anything (.*), then negative lookbehind (?<!",letters[i],") which is the condition
    that the letter i does not follow another letter i, the the letter itself 3 times, then the
    condition that it is not followed by the letter i, then followed by anything (.*)
F.letters.3.w<-F.letters.3.w[-(grep("www.",fPoS2l[F.letters.3.w]))]
for(i in 1:26){assign(paste("F.letters.4.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{4}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}       ##assigns every
    instance of 4 letters in a sequence to a variable
```

```
for(i in 1:26){assign(paste("F.letters.5.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{5}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}      ##assigns every instance of 5 letters in a sequence to a variable

for(i in 1:26){assign(paste("F.letters.6.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{6}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}      ##assigns every instance of 3  letters in a sequence to a variable

for(i in 1:26){assign(paste("F.letters.7.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{7}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}      ##assigns every instance of 3  letters in a sequence to a variable

for(i in 1:26){assign(paste("F.letters.8.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{8}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}      ##assigns every instance of 3 letters in a sequence to a variable

for(i in 1:26){assign(paste("F.letters.9.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{9}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}      ##assigns every instance of 3 letters in a sequence to a variable

for(i in 1:26){assign(paste("F.letters.10.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{10}(?!",letters[i],")(.*)",sep=""),fPoS2l,perl=T))}      ##assigns every instance of 10 letters in a sequence to a variable

for(i in 1:26){print(length(eval(parse(text=paste("F.letters.3.",letters[i],sep=""))))))}   ### prints 3-letter expressive lengthening

for(t in 1:8){assign(paste("length.f.",t+2,sep=""),vector("integer",26))}
for(t in 1:8){
for(b in 1:26){eval(parse(text=paste("length.f.",t+2,"[[",b,"]]","<-","length(F.letters.",t+2,"."
,letters[b],")",sep="")))
}
 }

tot.f.leng<-lapply(1:8,function(t){f.leng<-lapply(1:26,function(i){T<-eval(parse(text=paste("F.letters.",t+2,".",letters[i],sep = ""))))})})           #this gets the row numbers of all tokens in fPoS2l that contain lengthenings
tot.f.leng<-unlist(tot.f.leng)

```

```
names(length.f.3)<-letters
names(length.f.4)<-letters
names(length.f.5)<-letters
names(length.f.6)<-letters
names(length.f.7)<-letters
names(length.f.8)<-letters
names(length.f.9)<-letters
names(length.f.10)<-letters

lengthenedF<-rbind(length.f.3,length.f.4,length.f.5,length.f.6,length.f.7,length.f.8,length.f.9,
    length.f.10)
lengthenedF[,order(-length.f.3)]

for(i in 1:26){assign(paste("c.letters.3.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{3}(?!",letters[i],")(.*)",sep=""),cPoS21,perl=T))}         ##assigns every
    instance of 3 letters in a sequence to a variable
c.letters.3.w<-c.letters.3.w[-(grep("www.",cPoS21[c.letters.3.w]))]     #to remove the url
    addresses

for(i in 1:26){assign(paste("c.letters.4.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{4}(?!",letters[i],")(.*)",sep=""),cPoS21,perl=T))}           ##assigns every
     instance of 4 letters in a sequence to a variable

for(i in 1:26){assign(paste("c.letters.5.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{5}(?!",letters[i],")(.*)",sep=""),cPoS21,perl=T))}          ##assigns every
    instance of 5 letters in a sequence to a variable

for(i in 1:26){assign(paste("c.letters.6.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{6}(?!",letters[i],")(.*)",sep=""),cPoS21,perl=T))}           ##assigns every
     instance of 6 letters in a sequence to a variable

for(i in 1:26){assign(paste("c.letters.7.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{7}(?!",letters[i],")(.*)",sep=""),cPoS21,perl=T))}          ##assigns every
    instance of 7 letters in a sequence to a variable

for(i in 1:26){assign(paste("c.letters.8.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],
    ")",letters[i],"{8}(?!",letters[i],")(.*)",sep=""),cPoS21,perl=T))}           ##assigns every
     instance of 8 letters in a sequence to a variable
```

```
for(i in 1:26){assign(paste("c.letters.9.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{9}(?!",letters[i],")(.*)",sep=""),cPoS2l,perl=T))}          ##assigns every instance of 9 letters in a sequence to a variable

for(i in 1:26){assign(paste("c.letters.10.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{10}(?!",letters[i],")(.*)",sep=""),cPoS2l,perl=T))}           ##assigns every instance of 10 letters in a sequence to a variable

for(i in 1:26){print(length(eval(parse(text=paste("c.letters.3.",letters[i],sep=""))))))}   ### prints 3-letter expressive lengthening

for(t in 1:8){assign(paste("length.c.",t+2,sep=""),vector("integer",26))}
for(t in 1:8){
for(b in 1:26){eval(parse(text=paste("length.c.",t+2,"[[",b,"]]","<-","length(c.letters.",t+2,"."
,letters[b],")",sep="")))
}
 }

tot.c.leng<-lapply(1:8,function(t){lapply(1:26,function(i){eval(parse(text=paste("c.letters.",t
+2,".",letters[i],sep = "")))})})          #this gets the row numbers of all tokens in cPoS2l
 that contain lengthenings
tot.c.leng<-unlist(tot.c.leng)

finlengthenings<-sort(table(fPoS2l[tot.f.leng]),decreasing=T)
clengthenings<-sort(table(cPoS2l[tot.c.leng]),decreasing=T)

names(length.c.3)<-letters
names(length.c.4)<-letters
names(length.c.5)<-letters
names(length.c.6)<-letters
names(length.c.7)<-letters
names(length.c.8)<-letters
names(length.c.9)<-letters
names(length.c.10)<-letters

lengthenedC<-rbind(length.c.3,length.c.4,length.c.5,length.c.6,length.c.7,length.c.8,length.c.9,
length.c.10)
lengthenedC[,order(-length.c.3)]
```

```
sum(lengthenedF[,order(-length.f.3)]*1000/length(fPoS2l))      ###total expressive lengthenings per 1000 words
sum(lengthenedC[,order(-length.c.3)]*1000/length(cPoS2l))
chisq.test(matrix(c(sum(lengthenedF),length(fPoS2l)-sum(lengthenedF),sum(lengthenedC),length(cPoS2l)-sum(lengthenedC)),ncol=2))    #goodness-of-fit for no. of expressive lengthenings in FEC and CEC

tot.f.leng.100<-sort(table(fPoS2l[sort(tot.f.leng)]),decreasing=T)[1:100]     ## 100 most frequent types with lengthening
tot.c.leng.100<-sort(table(cPoS2l[sort(tot.c.leng)]),decreasing=T)[1:100]     ## 100 most frequent types with lengthening
tot.f.leng.100[grep('([[:alpha:]])\\1{4}(?![[:alpha:]])(.*)',names(tot.f.leng.100),perl=T)]

ln.F<-as.data.frame(lengthenedF)
ln.F<-ln.F[order(colSums(lengthenedF),decreasing=T)]
ln.C<-as.data.frame(lengthenedC)
cor.test(match(names(ln.C),letters),match(names(ln.F),letters), method="spearman") #Spearman's rho to compare the two rankings of letters by frequency
ln.C<-ln.C[,(match(names(ln.F),names(ln.C)))] #lengthenings in CEC in same letter order as FEC

finlengtheningsP<-barplot((unname(finlengthenings[c(3:4,6:16,18:23,25)]))*1000/length(fPoS2l),names.arg=c(names(finlengthenings[c(3:4,6)]),"XDDD",names(finlengthenings[c(8:15)]),":DDD",names(finlengthenings[c(18:23,25)])),las=2,ylim=c(0,.1),ylab="Frequency per Thousand Words")
title("Most frequent expressive lengthening types, Finland English Corpus")
text(finlengtheningsP,unname(finlengthenings[c(3:4,6:16,18:23,25)])*1000/length(fPoS2l)+.005,labels=round(unname(finlengthenings[c(3:4,6:16,18:23,25)])*1000/length(fPoS2l),digits=3),cex=0.6,col="black")
clengtheningsP<-barplot((unname(clengthenings[c(1:11,13:21)]))*1000/length(cPoS2l),names.arg=names(clengthenings[c(1:11,13:21)]),las=2,ylim=c(0,.11),ylab="Frequency per Thousand Words")
title("Most frequent expressive lengthening types, Comparison English Corpus")
text(clengtheningsP,unname(clengthenings[c(1:12,14:21)])*1000/length(cPoS2l)+.005,labels=round(unname(clengthenings[c(1:12,14:21)])*1000/length(cPoS2l),digits=3),cex=0.6,col="black")

letters3<-rbind(as.numeric(ln.F[1,1:26])*1000/length(finw),as.numeric(ln.C[1,1:26])*1000/length(compw))                              ##### 3-char sequences per thousand words
letters4<-rbind(as.numeric(ln.F[2,1:26])*1000/length(finw),as.numeric(ln.C[2,1:26])*1000/length(compw))
letters5<-rbind(as.numeric(ln.F[3,1:26])*1000/length(finw),as.numeric(ln.C[3,1:26])*1000/length(compw))
```

```
letters6<-rbind(as.numeric(ln.F[4,1:26])*1000/length(finw),as.numeric(ln.C[4,1:26])*1000/length(
    compw))
letters7<-rbind(as.numeric(ln.F[5,1:26])*1000/length(finw),as.numeric(ln.C[5,1:26])*1000/length(
    compw))
letters8<-rbind(as.numeric(ln.F[6,1:26])*1000/length(finw),as.numeric(ln.C[6,1:26])*1000/length(
    compw))
letters9<-rbind(as.numeric(ln.F[7,1:26])*1000/length(finw),as.numeric(ln.C[7,1:26])*1000/length(
    compw))
letters10<-rbind(as.numeric(ln.F[8,1:26])*1000/length(finw),as.numeric(ln.C[8,1:26])*1000/length(
    compw))

letters3P<-barplot(letters3,beside=T,ylim=c(0,.5))
legend("topright",title="Frequency of Types with 3-Character Sequences",legend=c("Finland English
     Corpus","Comparison Corpus"),cex=1,col=gray.colors(2),fill=gray.colors(2))
title("Frequency of expressive lengthening sequences per letter sequence and thousand words, 3-6
    repetitions")
text(letters3P,letters3+.02,labels=round(letters3,digits=2),cex=0.5,col="black")
letters4P<-barplot(letters4,beside=T,ylim=c(0,.5))
legend("topright",legend="Frequency of Types with 4-Character Sequences")
text(letters4P,letters4+.02,labels=round(letters4,digits=2),cex=0.5,col="black")
letters5P<-barplot(letters5,beside=T,ylim=c(0,.5))
legend("topright",legend="Frequency of Types with 5-Character Sequences")
text(letters5P,letters5+.02,labels=round(letters5,digits=2),cex=0.5,col="black")

letters6P<-barplot(letters6,beside=T,ylim=c(0,.5),names.arg=names(ln.F),cex.names=1.7)
legend("topright",legend="Frequency of Types with 6-Character Sequences")
text(letters6P,letters6+.02,labels=round(letters7,digits=2),cex=0.5,col="black")

letters7P<-barplot(letters7,beside=T,ylim=c(0,.1))
title("Frequency of expressive lengthening sequences per letter sequence and thousand words, 7-10
     repetitions")
legend("topright",title="Frequency of Types with 7-Character Sequences",legend=c("Finland English
     Corpus","Comparison Corpus"),cex=1,col=gray.colors(2),fill=gray.colors(2))
text(letters7P,letters7+.01,labels=round(letters7,digits=2),cex=0.5,col="black")
letters8P<-barplot(letters8,beside=T,ylim=c(0,.1))
legend("topright",legend="Frequency of Types with 8-Character Sequences")
text(letters8P,letters8+.01,labels=round(letters8,digits=2),cex=0.5,col="black")
letters9P<-barplot(letters9,beside=T,ylim=c(0,.1))
legend("topright",legend="Frequency of Types with 9-Character Sequences")
```

```
text(letters9P,letters9+.01,labels=round(letters5,digits=2),cex=0.5,col="black")
letters10P<-barplot(letters10,beside=T,ylim=c(0,.1),names.arg=names(ln.F),cex.names=1.7)
legend("topright",legend="Frequency of Types with 10-Character Sequences")
text(letters10P,letters10+.01,labels=round(letters10,digits=2),cex=0.5,col="black")

ln.C<-as.data.frame(lengthenedC)                        #re-order the data
ln.C<-ln.C[order(colSums(lengthenedC),decreasing=T)]  #re-order the data
letters.all.c.P<-barplot(colSums(ln.C)*1000/length(compw),ylim=c(0,.45))
title("Overall frequency of expressive lengthening sequences per letter sequence\n and thousand
    words, Comparison English Corpus")
text(letters.all.c.P,colSums(ln.C)*1000/length(compw)+.02,labels=round(colSums(ln.C)*1000/length(
    compw),digits=2),cex=0.5,col="black")

################Expressive lengthening by gender
################

femaleslow<-tolower(unlist(strsplit(df11females[,13]," ")))
maleslow<-tolower(unlist(strsplit(df11males[,13]," ")))

for(i in 1:26){assign(paste("F.letters.females.3.",letters[i],sep = ""),grep(paste("(.*)(?<!",
    letters[i],")",letters[i],"{3}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}       ##
    assigns every instance of 3 or more letters in a sequence to a variable
#regex gets anything (.*), then negative lookbehind (?<!",letters[i],") which is the condition
    that the letter i does not follow another letter i, then the letter itself 3 times, then the
    condition that it is not followed by the letter i, then followed by anything (.*)

F.letters.females.3.w<-F.letters.females.3.w[-(grep("www.",femaleslow[F.letters.females.3.w]))]

for(i in 1:26){assign(paste("F.letters.females.4.",letters[i],sep = ""),grep(paste("(.*)(?<!",
    letters[i],")",letters[i],"{4}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}       ##
    assigns every instance of 4 or more letters in a sequence to a variable

for(i in 1:26){assign(paste("F.letters.females.5.",letters[i],sep = ""),grep(paste("(.*)(?<!",
    letters[i],")",letters[i],"{5}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}      ##
    assigns every instance of 5 or more letters in a sequence to a variable

for(i in 1:26){assign(paste("F.letters.females.6.",letters[i],sep = ""),grep(paste("(.*)(?<!",
    letters[i],")",letters[i],"{6}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}      ##
    assigns every instance of 3 or more letters in a sequence to a variable
```

```
1496
1497 for(i in 1:26){assign(paste("F.letters.females.7.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{7}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}      ##
       assigns every instance of 3 or more letters in a sequence to a variable
1498
1499 for(i in 1:26){assign(paste("F.letters.females.8.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{8}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}      ##
       assigns every instance of 3 or more letters in a sequence to a variable
1500
1501 for(i in 1:26){assign(paste("F.letters.females.9.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{9}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}      ##
       assigns every instance of 3 or more letters in a sequence to a variable
1502
1503 for(i in 1:26){assign(paste("F.letters.females.10.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{10}(?!",letters[i],")(.*)",sep=""),femaleslow,perl=T))}      ##
       assigns every instance of 10 or more letters in a sequence to a variable
1504
1505 for(i in 1:26){print(length(eval(parse(text=paste("F.letters.females.3.",letters[i],sep=""))))))}
       ### prints 3-letter expressive lengthening
1506
1507 for(t in 1:8){assign(paste("length.f.females.",t+2,sep=""),vector("integer",26))}
1508 for(t in 1:8){
1509 for(b in 1:26){eval(parse(text=paste("length.f.females.",t+2,"[[",b,"]]","<-","length(F.letters.
       females.",t+2,".",letters[b],")",sep="")))
1510 }
1511  }
1512
1513 tot.female.leng<-lapply(1:8,function(t){lapply(1:26,function(i){eval(parse(text=paste("F.letters.
       females.",t+2,".",letters[i],sep = "")))})})          #this gets the row numbers of all
       tokens in cPoS21 that contain lengthenings
1514 tot.female.leng<-unlist(tot.female.leng)
1515 female.lengthenings<-sort(table(femaleslow[tot.female.leng]),decreasing=T)
1516
1517 names(length.f.females.3)<-letters
1518 names(length.f.females.4)<-letters
1519 names(length.f.females.5)<-letters
1520 names(length.f.females.6)<-letters
1521 names(length.f.females.7)<-letters
1522 names(length.f.females.8)<-letters
```

```
1523 names(length.f.females.9)<-letters
1524 names(length.f.females.10)<-letters
1525
1526 lengthenedF.females<-rbind(length.f.females.3,length.f.females.4,length.f.females.5,length.f.
       females.6,length.f.females.7,length.f.females.8,length.f.females.9,length.f.females.10)
1527 lengthenedF.females[,order(-length.f.females.3)]
1528
1529 for(i in 1:26){assign(paste("F.letters.males.3.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{3}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}       ##
       assigns every instance of 3 or more letters in a sequence to a variable
1530 #regex gets anything (.*), then negative lookbehind (?<!",letters[i],") which is the condition
       that the letter i does not follow another letter i, the the letter itself 3 times, then the
       condition that it is not followed by the letter i, then followed by anything (.*)
1531 F.letters.males.3.w<-F.letters.males.3.w[-(grep("www.",maleslow[F.letters.males.3.w]))]
1532 for(i in 1:26){assign(paste("F.letters.males.4.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{4}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}       ##
       assigns every instance of 4 or more letters in a sequence to a variable
1533
1534 for(i in 1:26){assign(paste("F.letters.males.5.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{5}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}      ##
       assigns every instance of 5 or more letters in a sequence to a variable
1535
1536 for(i in 1:26){assign(paste("F.letters.males.6.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{6}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}      ##
       assigns every instance of 3 or more letters in a sequence to a variable
1537
1538 for(i in 1:26){assign(paste("F.letters.males.7.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{7}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}      ##
       assigns every instance of 3 or more letters in a sequence to a variable
1539
1540 for(i in 1:26){assign(paste("F.letters.males.8.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{8}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}      ##
       assigns every instance of 3 or more letters in a sequence to a variable
1541
1542 for(i in 1:26){assign(paste("F.letters.males.9.",letters[i],sep = ""),grep(paste("(.*)(?<!",
       letters[i],")",letters[i],"{9}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}      ##
       assigns every instance of 3 or more letters in a sequence to a variable
1543
```

```
for(i in 1:26){assign(paste("F.letters.males.10.",letters[i],sep = ""),grep(paste("(.*)(?<!",letters[i],")",letters[i],"{10}(?!",letters[i],")(.*)",sep=""),maleslow,perl=T))}      ## assigns every instance of 10 or more letters in a sequence to a variable

for(i in 1:26){print(length(eval(parse(text=paste("F.letters.males.3.",letters[i],sep=""))))))}
    ### prints 3-letter expressive lengthening

for(t in 1:8){assign(paste("length.f.males.",t+2,sep=""),vector("integer",26))}
for(t in 1:8){
for(b in 1:26){eval(parse(text=paste("length.f.males.",t+2,"[[",b,"]]","<-","length(F.letters.males.",t+2,".",letters[b],")",sep="")))
}
 }

tot.male.leng<-lapply(1:8,function(t){lapply(1:26,function(i){eval(parse(text=paste("F.letters.males.",t+2,".",letters[i],sep = "")))})})          #this gets the row numbers of all tokens in cPoS2l that contain lengthenings
tot.male.leng<-unlist(tot.male.leng)

male.lengthenings<-sort(table(maleslow[tot.male.leng]),decreasing=T)

both.lengthenings<-intersect(names(female.lengthenings),names(male.lengthenings))
female.lengthenings[both.lengthenigns]

female.lengthenings[match(both.lengthenings,names(female.lengthenings))]/male.lengthenings[match(both.lengthenings,names(male.lengthenings))]          #FM ratio for lengthening types in both female and male FinEng subsections


names(length.f.males.3)<-letters
names(length.f.males.4)<-letters
names(length.f.males.5)<-letters
names(length.f.males.6)<-letters
names(length.f.males.7)<-letters
names(length.f.males.8)<-letters
names(length.f.males.9)<-letters
names(length.f.males.10)<-letters
```

```
lengthenedF.males<-rbind(length.f.males.3,length.f.males.4,length.f.males.5,length.f.males.6,
    length.f.males.7,length.f.males.8,length.f.males.9,length.f.males.10)
lengthenedF.males[,order(-length.f.males.3)]

sum(lengthenedF.males[,order(-length.f.males.3)])*1000/length(malesw1)     ###total expressive
    lengthenings per 1000 words for males
sum(lengthenedF.females[,order(-length.f.females.3)])*1000/length(femalesw1)  ###total expressive
     lengthenings per 1000 words for females

chisq.test(matrix(c(sum(lengthenedF.males),length(malesw1),sum(lengthenedF.females),length(
    femalesw1)),ncol=2))

###### Major Word Class Frequencies
f.nouns<-length(which(finp=="NN"|finp=="NNP"|finp=="NNS"))
c.nouns<-length(which(compp=="NN"|compp=="NNP"|compp=="NNS"))
f.verbs<-length(which(finp=="VB"|finp=="VBD"|finp=="VBG"|finp=="VBN"|finp=="VBP"|finp=="VBZ"))
c.verbs<-length(which(compp=="VB"|compp=="VBD"|compp=="VBG"|compp=="VBN"|compp=="VBP"|compp=="VBZ
    "))
f.pronouns<-length(which(finp=="PRP"|finp=="PRP."))
c.pronouns<-length(which(compp=="PRP"|compp=="PRP."))
f.major<-c(f.nouns,f.verbs,f.pronouns)*1000/length(finp)
c.major<-c(c.nouns,c.verbs,c.pronouns)*1000/length(compp)
conv.major<-c(143,140,140);class.major<-c(186,153,135);text.major<-c(302,86,26);acad.major<-c
    (295,83,17)
major<-cbind(f.major,c.major,conv.major,class.major,text.major,acad.major)
major.P<-barplot(major,ylab="Frequency per 1000 words",ylim=c(0,350), beside = TRUE, names.arg =
    c("Finland English Corpus","Comparison English Corpus","Conversation","Classroon Teaching","
    Textbooks","Academic Prose"),,cex.names=.8,las=0)
legend("top",legend=c("Nouns","Verbs","Pronouns"),col=gray.colors(3),fill=gray.colors(3))
title("Frequency of Major Word Classes")


##############DETERMINERS: Demonstrative

f.dt<-finw[which(finp=="DT")]
c.dt<-compw[which(compp=="DT")]

f.that<-grep("that",tolower(f.dt))
c.that<-grep("that",tolower(c.dt))
```

```
1605
1606 f.this<-grep("this",tolower(f.dt))
1607 c.this<-grep("this",tolower(c.dt))
1608
1609 f.these<-grep("these",tolower(f.dt))
1610 c.these<-grep("these",tolower(c.dt))
1611
1612 f.those<-grep("those",tolower(f.dt))
1613 c.those<-grep("those",tolower(c.dt))
1614
1615 f.demonstr<-c(length(f.this),length(f.that),length(f.these),length(f.those),sum(c(length(f.this),
       length(f.that),length(f.these),length(f.those))))*1000/length(finp)
1616 c.demonstr<-c(length(c.this),length(c.that),length(c.these),length(c.those),sum(c(length(c.this),
       length(c.that),length(c.these),length(c.those))))*1000/length(compp)
1617
1618 demonstr<-rbind(f.demonstr,c.demonstr)
1619 demonstrnames<-c(expression(italic("this")),expression(italic("that")),expression(italic("these")
       ),expression(italic("those")),"all demonstrative\n determiners")
1620 demonstr.P<- barplot(demonstr,ylab="Frequency per 1000 words",ylim=c(0,14), beside = TRUE, names.
       arg =c("this","that","these","those","(all demonstrative\ndeterminers)"))
1621
1622 legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
       colors(2))
1623 title("Frequency of Demonstrative Determiners in the Material")
1624 text(demonstr.P,demonstr+0.6,labels=round(demonstr,digits=2),cex=0.7,col="black")
1625
1626 sum(f.demonstr[c(1,3)])/sum(f.demonstr[c(2,4)])  #Finland proximal/distal ratio
1627 sum(c.demonstr[c(1,3)])/sum(c.demonstr[c(2,4)])  #Comparison proximal/distal ratio
1628
1629 ######################DETERMINERS
1630 f.deter<-length(f.dt)*1000/length(finp)
1631 c.deter<-length(c.dt)*1000/length(compp)
1632 deter<-rbind(f.deter,c.deter)
1633
1634 deter.P<- barplot(deter,ylab="Frequency per 1000 words",ylim=c(0,100), beside = TRUE)
1635 legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
       colors(2))
1636 title("Frequency of Determiners in the Material")
1637 text(deter.P,deter+5,labels=round(deter,digits=2),cex=0.9,col="black")
```

```
1638
1639 ######################ARTICLES
1640 f.a<-grep("a",f.dt)
1641 f.an<-grep("an",f.dt)
1642 f.the<-grep("the",f.dt)
1643 c.a<-grep("a",c.dt)
1644 c.an<-grep("an",c.dt)
1645 c.the<-grep("the",c.dt)
1646
1647 f.artic<-c(length(f.a),length(f.an),length(f.the))*1000/length(finp)
1648 c.artic<-c(length(c.a),length(c.an),length(c.the))*1000/length(compp)
1649 artic<-rbind(f.artic,c.artic)
1650
1651 artic.P<- barplot(artic,ylab="Frequency per 1000 words",ylim=c(0,50), beside = TRUE,names.arg=c("
        a","an","the"))
1652 legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
        colors(2))
1653 title("Frequency of Articles in the Material")
1654 text(artic.P,artic+2,labels=round(artic,digits=2),cex=0.9,col="black")
1655
1656 #########Quantifying determiners
1657 f.all<-grep("all",tolower(f.dt))
1658 c.all<-grep("all",tolower(c.dt))
1659
1660 f.both<-grep("both",tolower(f.dt))
1661 c.both<-grep("both",tolower(c.dt))
1662
1663 f.another<-grep("another",tolower(f.dt))
1664 c.another<-grep("another",tolower(c.dt))
1665
1666 f.each<-grep("each",tolower(f.dt))
1667 c.each<-grep("each",tolower(c.dt))
1668
1669 f.every<-grep("every",tolower(f.dt))
1670 c.every<-grep("every",tolower(c.dt))
1671
1672 f.many<-grep("many",tolower(f.dt))
1673 c.many<-grep("many",tolower(c.dt))
1674
```

```
#f.much<-grep("much",tolower(f.dt[,1]))  #is tagged as adverb or adjective
#c.much<-grep("much",tolower(c.dt[,1]))

f.some<-grep("some",tolower(f.dt))
c.some<-grep("some",tolower(c.dt))

f.few<-grep("few",tolower(f.dt))
c.few<-grep("few",tolower(c.dt))

f.little<-grep("little",tolower(f.dt))
c.little<-grep("little",tolower(c.dt))

f.any<-grep("any",tolower(f.dt))
c.any<-grep("any",tolower(c.dt))

f.either<-grep("either",tolower(f.dt))
c.either<-grep("either",tolower(c.dt))

f.neither<-grep("neither",tolower(f.dt))
c.neither<-grep("neither",tolower(c.dt))

f.quantif<-c(length(f.all),length(f.both),length(f.another),length(f.every),length(f.many),length
    (f.some),length(f.any),length(f.either),length(f.neither),sum(c(length(f.all),length(f.both),
    length(f.another),length(f.every),length(f.many),length(f.some),length(f.any),length(f.either
    ),length(f.neither))))*1000/length(finp)
c.quantif<-c(length(c.all),length(c.both),length(c.another),length(c.every),length(c.many),length
    (c.some),length(c.any),length(c.either),length(c.neither),sum(c(length(c.all),length(c.both),
    length(c.another),length(c.every),length(c.many),length(c.some),length(c.any),length(c.either
    ),length(c.neither))))*1000/length(compp)

quantif<-rbind(f.quantif,c.quantif)
quantif.P<- barplot(quantif,ylab="Frequency per 1000 words",ylim=c(0,8), beside = TRUE, names.arg
     =c("all","both","another","every","many","some","any","either","neither","all quant.\
    ndeterminers"))
gend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
    colors(2))
title("Frequency of Quantifying Determiners in the Material")
text(quantif.P,quantif+0.2,labels=round(quantif,digits=3),cex=0.7,col="black")

```

```
############NUM DIGITS
f.digits<-grep("^[[:digit:]]+$",finw)
c.digits<-grep("^[[:digit:]]+$",compw)
digits<-rbind(length(f.digits)*1000/length(finw),length(c.digits)*1000/length(compw))
digits.P<-barplot(digits,ylim=c(0,10),ylab="Frequency per 1000 words",beside=T)
legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
    colors(2))
title("Frequency of Numerical Digits")
text(digits.P,digits+0.2,labels=round(digits,digits=3),cex=0.9,col="black")

############MOST FREQUENT NUM DIGITS
f.digitsmf<-sort(table(finw[f.digits]),decreasing=T)
c.digitsmf<-sort(table(factor(compw[c.digits],levels=names(f.digitsmf))),decreasing=T)
digitsmf<-rbind(f.digitsmf[c(1:11,13:21)]*1000/length(finw),c.digitsmf[c(1:11,13:21)]*1000/length
    (compw))
digitsmf.P<-barplot(digitsmf,ylim=c(0,1.2),ylab="Frequency per 1000 words",beside=T)
legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
    colors(2))
title("Most Frequent Numerical Digits")
text(digitsmf.P,digitsmf+0.02,labels=round(digitsmf,digits=2),cex=0.4,col="black")


############ORDINALS as DIGITS WITH ORDINAL INDICATORS
f.ordin<-grep("^[[:digit:]]+(st|nd|rd|th)$",finw)
c.ordin<-grep("^[[:digit:]]+(st|nd|rd|th)$",compw)
f.xst<-grep("^[[:digit:]]+(st)$",finw)
f.nd<-grep("^[[:digit:]]+(nd)$",finw)
f.rd<-grep("^[[:digit:]]+(rd)$",finw)
f.th<-grep("^[[:digit:]]+(th)$",finw)
c.xst<-grep("^[[:digit:]]+(st)$",compw)
c.nd<-grep("^[[:digit:]]+(nd)$",compw)
c.rd<-grep("^[[:digit:]]+(rd)$",compw)
c.th<-grep("^[[:digit:]]+(th)$",compw)

f.ordin1<-c(length(f.xst)*1000/length(finw),length(f.nd)*1000/length(finw),length(f.rd)*1000/
    length(finw),length(f.th)*1000/length(finw));c.ordin1<-c(length(c.xst)*1000/length(compw),
    length(c.nd)*1000/length(compw),length(c.rd)*1000/length(compw),length(c.th)*1000/length(
    compw))

```

```
ordin<-rbind(length(f.ordin)*1000/length(finw),length(c.ordin)*1000/length(compw))
ordin1<-rbind(f.ordin1,c.ordin1)

ordin1.P<-barplot(ordin1,ylim=c(0,.5),ylab="Frequency per 1000 words",beside=T,names.arg=c("-st",
    "-nd","-rd","-th"))
legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
    colors(2))
title("Frequency of Ordinal Suffixes with Numerical Digits")
text(ordin1.P,ordin1+0.02,labels=round(ordin1,digits=2),cex=0.8,col="black")

#################NUMERALS AS WORDS
f.numas<-finw[finp=="CD"]
c.numas<-compw[compp=="CD"]
f.numword<-grep("^[a-z]+$",finw[which(finp=="CD")])
c.numword<-grep("^[a-z]+$",compw[which(compp=="CD")])
f.numwords<-sort(table(finw[which(finp=="CD")][f.numword]),decreasing=T)
c.numwords<-sort(table(factor((compw[which(compp=="CD")][c.numword]),levels=names(f.numwords))),
    decreasing=T)
numwords<-rbind(f.numwords[1:20]*1000/length(finw),c.numwords[1:20]*1000/length(compw))
dimnames(numwords)[[2]][13]<-"couple"
numwords.P<-barplot(numwords,ylim=c(0,2),ylab="Frequency per 1000 words",beside=T,las=2)
legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.
    colors(2))
title("Frequency of Numerals as Words")
text(numwords.P,numwords+0.05,labels=round(numwords,digits=2),cex=0.45,col="black")

#############ORDINALS as WORDS
fw.word<-finw[which(finp=="CD")][f.numword]
cw.word<-compw[which(compp=="CD")][c.numword]

fw.xst<-grep("?.*first$",finw)
fw.nd<-grep("?.*second$",finw)
fw.rd<-grep("?.*third$",finw)
fw.th<-grep("?.*fourth$|?.*fifth$|?.*sixth$|?.*seventh|?.*eighth$|?.*ninth$|?.*tenth$|?.*eleventh
    $|?.*twelfth$|?.*teenth|?.*ieth$|?.*edth$|?.*andth$|?.*ionth$",finw)
cw.xst<-grep("?.*first$",compw)
cw.nd<-grep("?.*second$",compw)
cw.rd<-grep("?.*third$",compw)
```

```
cw.th<-grep("?.*fourth$|?.*fifth$|?.*sixth$|?.*seventh|?.*eighth$|?.*ninth$|?.*tenth$|?.*eleventh$|?.*twelfth$|?.*teenth|?.*ieth$|?.*edth$|?.*andth$|?.*ionth$",compw)

fw.ordin1<-c(length(fw.xst)*1000/length(finw),length(fw.nd)*1000/length(finw),length(fw.rd)*1000/length(finw),length(fw.th)*1000/length(finw))
cw.ordin1<-c(length(cw.xst)*1000/length(compw),length(cw.nd)*1000/length(compw),length(cw.rd)*1000/length(compw),length(cw.th)*1000/length(compw))
w.ordin1<-rbind(fw.ordin1,cw.ordin1)

w.ordin1.P<-barplot(w.ordin1,ylim=c(0,1),ylab="Frequency per 1000 words",beside=T,names.arg=c("-st","-nd","-rd","-th"))
legend("top",legend=c("Finland English Corpus","Comparison Corpus"),col=gray.colors(2),fill=gray.colors(2))
title("Frequency of Ordinal Words by Suffix")
text(w.ordin1.P,w.ordin1+0.02,labels=round(w.ordin1,digits=2),cex=0.8,col="black")


################ Multidimensional Analysis
################

fPoSd<-split(finp, ceiling(seq_along(finp)/((length(finp)/100))))
cPoSd<-split(compp, ceiling(seq_along(compp)/((length(compp)/100))))
CHlev <- c(sort(unique(unlist(fPoSd))))
chunkF <- do.call(rbind,lapply(fPoSd,function(x,lev){ table(factor(x,levels = CHlev,ordered = TRUE))},lev = CHlev))
chunkF2<-as.data.frame(chunkF)
chunkFperThous<-(chunkF2[,1:(length(CHlev))])*1000/(unlist(lapply(fPoSd, function(x) sum(table(factor(x,levels=CHlev))))))
names(chunkFperThous)<-sort(unique(unlist(fPoSd)))
CHlevC <- c(sort(unique(unlist(cPoSd))))
chunkC <- do.call(rbind,lapply(cPoSd,function(x,lev){ table(factor(x,levels = CHlevC,ordered = TRUE))},lev = CHlevC))
chunkC2<-as.data.frame(chunkC)
chunkCperThous<-(chunkC2[,1:(length(CHlevC))])*1000/(unlist(lapply(cPoSd, function(x) sum(table(factor(x,levels=CHlevC))))))
names(chunkCperThous)<-sort(unique(unlist(cPoSd)))
corFC<-cor(chunkFperThous,chunkCperThous)

drops<-c("FW","-LRB-","-RRB-","SYM","NNPS")
```

```
DF<-chunkFperThous
DF1<-DF[,!(names(DF) %in% drops)]
fit <- factanal(DF1, 7, rotation="varimax")
print(fit, digits=2, cutoff=.3, sort=TRUE)
# plot factor 1 by factor 2
load <- fit$loadings[,1:7]
plot(load,type="n") # set up plot
text(load,labels=names(DF1),cex=.8,col=1) # add variable names; plot of Finland Corpus Factors 1
    and 2
plot(load,type="n") # set up plot
text(load,labels=names(DF1),cex=.8,col=1)

DFc<-chunkCperThous
DFc1<-DFc[,!(names(DFc) %in% drops)]
fitc <- factanal(DFc1, 7, rotation="varimax")
print(fitc, digits=2, cutoff=.3, sort=TRUE)
# plot factor 1 by factor 2
loadc <- fitc$loadings[,1:7]
plot(loadc,type="n") # set up plot
text(loadc,labels=names(DFc1),cex=.8,col=1) # add variable names; plot of Comparison Corpus
    Factors 1 and 2
plot(loadc,type="n") # set up plot
text(loadc,labels=names(DFc1),cex=.8,col=1)


######Example tweets with and without factor loading tags x<.3 or x>.3.
fin.factor1.ex<-grep("(?=.*MD)(?=.*PRP)(?=.*VB)(?=.*VBP)(?=.*USR)(?!.*:)(?!.*NNP)(?!.*URL)(?!.*NN
    )(?!.*HT)(?!.*IN)",df11[,12],perl=T)
fin.factor2.ex<-grep("(?=.*VBN)(?=.*VBG)(?=.*TO)(?=.*NN)(?=.*IN)(?=.*PRP\\$)(?!.*'')(?!.*RT)(?!.*
    USR)(?!.*UH)",df11[,12],perl=T)  #doesnt work for negs
fin.factor3.ex<-grep("(?=.*DT)(?=.*NN)(?=.*WRB)",df11[,12],perl=T)
fin.factor4.ex<-grep("(?=.*RB)(?=.*CC)(?=.*JJR)(?=.*VBD)(?=.*WRB)",df11[,12],perl=T)

G2.1<-data.frame(scale(G2))      #33000 tweets; scale converts the scores in the vectors to z-
    scores
names(G2.1)<-names(G2)
G2.2<-data.frame(scale(chunkF2)) #100 chunks, not individual tweets
names(G2.2)<-names(chunkF2)
G2c.1<-data.frame(scale(G2c))    #192000 tweets
```

```
names(G2c.1)<-names(G2c)
G2c.2<-data.frame(scale(chunkC2))#100 chunks, not tweets
names(G2c.2)<-names(chunkC2)
drops<-c("FW","-LRB-","-RRB-","SYM","NNPS","TOK")

G2.1<-G2.1[,!(names(G2.1) %in% drops)]
G2c.1<-G2c.1[,!(names(G2c.1) %in% drops)]

fit1<-princomp(G2[,1:41])     #unscaled, 33000 tweets
fit2<-princomp(G2.1,cor=T)   #scaled, 33000 tweets
fit3<-princomp(G2c[,1:41])    #unscaled, 192000 tweets
fit4<-princomp(G2c.1)  #scaled, 192000 tweets
fit5<-princomp(chunkF2)#unscaled, Finland chunks
fit6<-princomp(G2.2)          #scaled, Finland chunks
fit7<-princomp(chunkC2)#unscaled, Comparison chunks
fit8<-princomp(G2c.2)         #scaled, Comparison chunks

summary(fit2)

biplot(fit2,cex=c(.5,1.2),xlabs=c(rep("\u25cb",,color="grey50",alpha=.1,length(G2[,1]))))      #
    ## "?" is U+25CB
biplot(fit4,ylim=c(-.01,.015),xlim=c(-.01,.0075),cex=c(.5,1),xlabs=c(rep("\u25cb",length(G2c[,1])
    )),alpha=.1)
biplot(fit7,cex=c(.5,1.2),xlabs=c(rep("\u25cb",length(G2.2[,1]))),alpha=.1)
biplot(fit8,cex=c(.5,1.2),xlabs=c(rep("\u25cb",length(G2c.2[,1]))),alpha=.1)
biplot(fit2,cex=c(.5,1.2),xlabs=c(rep("o",color="grey50",alpha=.1,length(G2[,1]))))      ### "?"
    is U+25CB

library(MASS)
fMASS<-persp(kde2d(fit4$scores[,1],fit4$scores[,2],n=200),                  #for a 3d plot of the
     tweets in the first 2 dimensions
phi=30,theta=20,d=10,col="lightblue",shade=.75,ltheta=-100,
border=NA,expand=.5,xlab="Component 1",ylab="Component 2",zlab="density")
persp(kde2d(fit4$scores[,1],fit4$scores[,2],n=200),                 #for a 3d plot of the tweets
     in the first 2 dimensions
phi=30,theta=20,d=10,col="lightblue",shade=.75,ltheta=-100,
border=NA,expand=.5,xlab="Component 1",ylab="Component 2",zlab="density")
```

```
1869
1870 ################ N-grams: Lexical and Grammatical Bundles
1871 ################
1872
1873 fPoS2np<-unname(sapply(finw, function(x) gsub("[[:punct:]]","",x)))
1874 fPoSneg<-which(fPoS2np=="")
1875 fPoS2npW<-fPoS2np[-fPoSneg]
1876 fPoS2npPOS<-finp[-fPoSneg]
1877 fPoS2np<-data.frame(fPoS2npW,fPoS2npPOS,stringsAsFactors=F)
1878
1879 fVnp1<-fPoS2np$fPoS2npW
1880
1881 fVnp1.2<-c(fVnp1[-1], "END")
1882 fVnp2<-fPoS2np$fPoS2npPOS
1883 fVnp2.2<-c(fVnp2[-1], "END")
1884
1885 fVnp1.3<-c(fVnp1.2[-1], "END")
1886 fVnp2.3<-c(fVnp2.2[-1], "END")
1887
1888 fVnp1.4<-c(fVnp1.3[-1], "END")
1889 fVnp2.4<-c(fVnp2.3[-1], "END")
1890
1891 fVnp1qg<-paste(fVnp1,fVnp1.2,fVnp1.3,fVnp1.4)
1892 fVnp2qg<-paste(fVnp2,fVnp2.2,fVnp2.3,fVnp2.4)
1893 fnpQuadgrams<-data.frame(fVnp1qg,fVnp2qg,stringsAsFactors=F)
1894
1895 cPoS2np<-unname(sapply(compw, function(x) gsub("[[:punct:]]","",x)))
1896 cPoSneg<-which(cPoS2np=="")
1897 cPoS2npW<-cPoS2np[-cPoSneg]
1898 cPoS2npPOS<-compp[-cPoSneg]
1899 cPoS2np<-data.frame(cPoS2npW,cPoS2npPOS,stringsAsFactors=F)
1900
1901 cVnp1<-cPoS2np$cPoS2npW
1902
1903 cVnp1.2<-c(cVnp1[-1], "END")
1904 cVnp2<-cPoS2np$cPoS2npPOS
1905 cVnp2.2<-c(cVnp2[-1], "END")
1906
1907 cVnp1.3<-c(cVnp1.2[-1], "END")
```

```
cVnp2.3<-c(cVnp2.2[-1], "END")

cVnp1.4<-c(cVnp1.3[-1], "END")
cVnp2.4<-c(cVnp2.3[-1], "END")
cVnp1qg<-paste(cVnp1,cVnp1.2,cVnp1.3,cVnp1.4)
cVnp2qg<-paste(cVnp2,cVnp2.2,cVnp2.3,cVnp2.4)
cnpQuadgrams<-data.frame(cVnp1qg,cVnp2qg,stringsAsFactors=F)

fnpQ<-sort(table(fnpQuadgrams[,1]),decreasing=T)
nptest5<-match(names(fnpQ),fnpQuadgrams[,1])
fnpQ.tab<-data.frame(fnpQ,fnpQuadgrams[nptest5,2])

cnpQ<-sort(table(cnpQuadgrams[,1]),decreasing=T)
nptest6<-match(names(cnpQ),cnpQuadgrams[,1])
cnpQ.tab<-data.frame(cnpQ,cnpQuadgrams[nptest6,2])


fnpQp<-sort(table(fnpQuadgrams[,2]),decreasing=T)

hthththts.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="HT HT HT HT"),1]
hthththts.f20<-sort(table(hthththts.f),decreasing=T)[1:20]

nnpnnpnnpnnps.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="NNP NNP NNP NNP"),1]
nnpnnpnnpnnps.f20<-sort(table(nnpnnpnnpnnps.f),decreasing=T)[1:20]

nnindtnns.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="NN IN DT NN"),1]
nnindtnns.f20<-sort(table(nnindtnns.f),decreasing=T)[1:20]

prpvbptovbs.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="PRP VBP TO VB"),1]
prpvbptovbs.f20<-sort(table(prpvbptovbs.f),decreasing=T)[1:20]

indtjjnns.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="IN DT JJ NN"),1]
indtjjnns.f20<-sort(table(indtjjnns.f),decreasing=T)[1:20]

prpvbpdtnns.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="PRP VBP DT NN"),1]
prpvbpdtnns.f20<-sort(table(prpvbpdtnns.f),decreasing=T)[1:20]

nnpincdnnss.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="NNP IN CD NNS"),1]
nnpincdnnss.f20<-sort(table(nnpincdnnss.f),decreasing=T)[1:20]
```

```
1947
1948 indtnnins.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="IN DT NN IN"),1]
1949 indtnnins.f20<-sort(table(indtnnins.f),decreasing=T)[1:20]
1950
1951 nnpnnpnnpurls.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="NNP NNP NNP URL"),1]
1952 nnpnnpnnpurls.f20<-sort(table(nnpnnpnnpurls.f),decreasing=T)[1:20]
1953
1954 incdnnsurls.f<-fnpQuadgrams[which(fnpQuadgrams[,2]=="IN CD NNS URL"),1]
1955 incdnnsurls.f20<-sort(table(incdnnsurls.f),decreasing=T)[1:20]
1956
1957 fQuadgramsTypes<-data.frame(names(hthththts.f20),names(nnpnnpnnpnnps.f20),names(nnindtnns.f20),
       names(prpvbptovbs.f20),names(indtjjnns.f20),names(prpvbpdtnns.f20),names(nnpincdnnss.f20),
       names(indtnnins.f20),names(nnpnnpnnpurls.f20),names(incdnnsurls.f20))
1958
1959 cnpQp<-sort(table(cnpQuadgrams[,2]),decreasing=T)
1960
1961 nnpnnpnnpnnps.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="NNP NNP NNP NNP"),1]
1962 nnpnnpnnpnnps.c20<-sort(table(nnpnnpnnpnnps.c),decreasing=T)[1:20]
1963
1964 nnindtnns.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="NN IN DT NN"),1]
1965 nnindtnns.c20<-sort(table(nnindtnns.c),decreasing=T)[1:20]
1966
1967 indtnnnns.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="IN DT NN NN"),1]
1968 indtnnnns.c20<-sort(table(indtnnnns.c),decreasing=T)[1:20]
1969
1970 indtnnins.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="IN DT NN IN"),1]
1971 indtnnins.c20<-sort(table(indtnnins.c),decreasing=T)[1:20]
1972
1973 indtjjnns.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="IN DT JJ NN"),1]
1974 indtjjnns.c20<-sort(table(indtjjnns.c),decreasing=T)[1:20]
1975
1976 vbgtonnpnnps.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="VBG TO NNP NNP"),1]
1977 vbgtonnpnnps.c20<-sort(table(vbgtonnpnnps.c),decreasing=T)[1:20]
1978
1979 dtnnindts.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="DT NN IN DT"),1]
1980 dtnnindts.c20<-sort(table(dtnnindts.c),decreasing=T)[1:20]
1981
1982 dtjjnnins.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="DT JJ NN IN"),1]
1983 dtjjnnins.c20<-sort(table(dtjjnnins.c),decreasing=T)[1:20]
```

```
1984
1985 tovbtnns.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="TO VB DT NN"),1]
1986 tovbtnns.c20<-sort(table(tovbtnns.c),decreasing=T)[1:20]
1987
1988 nninnnpnnps.c<-cnpQuadgrams[which(cnpQuadgrams[,2]=="NN IN NNP NNP"),1]
1989 nninnnpnnps.c20<-sort(table(nninnnpnnps.c),decreasing=T)[1:20]
1990
1991 cQuadgramsTypes<-data.frame(names(nnpnnpnnpnnps.c20),names(nnindtnns.c20),names(indtnnnns.c20),
       names(indtnnins.c20),names(indtjjnns.c20),names(vbgtonnpnnps.c20),names(dtnnindts.c20),names(
       dtjjnnins.c20),names(tovbtnns.c20),names(nninnnpnnps.c20))
```

Appendix B

List of Finnish Names

TABLE B.1: List of Male and Female Names Used to Fetermine Gender in the Finland English Corpus (from http://www.sci.fi/~kajun/finns/)

Male Names				Female Names			
Aamos	Ismo	Lauri	Robin	Aamu	Heljä	Maire	Sanna-Leena
Aapo	Isto	Leevi	Roope	Aija	Helmi	Mari	Sanni
Aarne	Jaakko	Luukas	Sakari	Aila	Helvi	Maria	Sara
Aatos	Jali	Magnus	Saku	Aili	Henna	Marja	Sari
Ahti	Jan	Manu	Sami	Aino	Henriikka	Marjo	Satu
Aki	Jan-Erik	Marco	Samppa	Aira	Ida	Mathilda	Seija
Aki-Petteri	Jani	Marcus	Sampsa	Aliisa	Iida	Meeri	Selma
Akseli	Janne	Markku	Samsa	Amanda	Iines	Merja	Senja
Aleksi	Jari	Marko	Samuli	Anette	Ilse	Miia	Siiri
Anssi	Jarkko	Markus	Santeri	Anita	Ilta	Mikaela	Sini
Antero	Jarmo	Martti	Sauli	Anja	Impi	Milla	Sinikka
Antti	Jarno	Matias	Sebastian	Anna	Irene	Minna	Sirja
Ari	Jaska	Matti	Seppo	Anna-Liisa	Jaana	Mira	Sirkka
Ari-Pekka	Jean	Mattiesko	Severi	Anne	Jasmin	Monica	Sirpa
Armas	Jere	Mauno	Stefan	Anneli	Jenna	Natalia	Sisko
Arsi	Jesse	Maunu	Stig	Annemari	Jenni	Nea	Sofia
Arto	Joel	Mauri	Tahvo	Anni	Johanna	Nelma	Sointu
Arttu	Johan	Miika	Taneli	Anniina	Jonna	Niina	Sonja
Arvi	Johannes	Miikka	Tapani	Annika	Josefiina	Noora	Suoma
Arvid	Jonatan	Mika	Tapio	Annikki	Julia	Oili	Susanna
Atso	Joni	Mika-Matti	Tauno	Annukka	Justiina	Olga	Suvi
Atte	Jonne	Mikael	Teemu	Anu	Kaari	Oona	Säde
August	Joona	Mikki	Teppo	Arja	Kaarina	Outi	Taija
Aulis	Joonas	Mikko	Tero	Armi	Kaija	Paula	Taimi
Bo	Jorma	Miska	Teuvo	Auli	Kaiju	Pauliina	Taina
Christian	Jouko	Niklas	Timo	Aune	Kaisa	Petra	Tanja
Daavid	Jouni	Niko	Toini	Aurora	Karita	Pia	Tarja
Eemeli	Juha	Nils	Toivo	Carita	Karoliina	Piia	Teija
Eemil	Juhana	Olavi	Tom	Carola	Katariina	Piia-Noora	Tellervo
Eerik	Juhani	Olle	Tomi	Eeva	Kati	Pinja	Terhi
Eero	Juho	Olli	Tommi	Eija	Katja	Pirjo	Terttu
Eetu	Jukka	Olli-Pekka	Tommy	Eija-Riitta	Katri	Pirkko	Tiia
Eino	Jukka-Pekka	Onni	Toni	Eila	Kerttu	Päivi	Tiina
Einojuhani	Jussi	Oskar	Tony	Eliisa	Kia	Päivä	Tove
Elias	Juuso	Oskari	Topi	Elina	Kirsi	Raakel	Tuija
Emppu	Jyri	Otto	Topias	Elisa	Kirsti	Raija	Tuula
Ensio	Jyrki	Paavo	Tuomas	Elisabeth	Kristiina	Reeta	Tuuli
Erkki	Kaarle	Panu	Tuomo	Ella	Kyllikki	Reija	Tuulia
Erno	Kaarlo	Pasi	Tuukka	Elsa	Laila	Riia	Tuulikki
Esa	Kai	Pauli	Tyko	Emilia	Laura	Riikka	Tytti
Esa-Pekka	Kaj	Pekka	Urho	Emma	Leea	Riitta	Tyyne
Esko	Kalervo	Pentti	Valentin	Emmi	Leena	Rita	Ulla
Frans	Kalevi	Pertti	Valtteri	Essi	Leila	Ritva	Ulla-Maj
Fredrik	Kalle	Peter	Veijo	Eveliina	Lempi	Ronja	Ulpu
Hannes	Kari	Petri	Veikko	Hanna	Liisa	Roosa	Vappu
Hanno	Karri	Petteri	Veli	Hanna-Maria	Liisi	Saara	Veera
Hannu	Kauko	Pirkka	Veli-Matti	Hannele	Lotta	Saimi	Venla
Harri	Keijo	Pontus	Veli-Pekka	Heidi	Lyyli	Salla	Vilma
Harry	Keke	Raimo	Vesa	Helena	Maarit	Sanelma	Virpi
Heikki	Kim	Raine	Vihtori	Heli	Maija	Sanna	Åsa
Henri	Kimi	Raino	Vilho				
Henrik	Kimmo	Rasmus	Viljo Ville				
Hessu	Konsta	Reijo	Ville-Veikko				
Hugo	Kristian	Reima	Väinö				
Iiro	Kyösti	Reino	Yrjö				
Iivari	Lalli	Retu	Åke				
Ilkka	Lasse	Riku					
Ilmari	Lassi	Risto					