Swearing on Twitter: Examining tweeted profanities from the United States and the Nordic countries

Tapio Anttila
682285A Bachelor's Seminar and Thesis
English Philology
Faculty of Humanities
University of Oulu
Autumn 2017

Table of contents

Abstract

Communicating in English and using the social media platform Twitter have both become increasingly common all around the world, for example the Nordic countries. As a site of authentic language material, Twitter is very popular among researchers. Moreover, the Twitter Application Programming Interface (API) is very accessible, and it is relatively easy to write programming scripts for collecting large amounts of Twitter data. However, the topic of swearing on Twitter has not attracted much attention. This thesis investigates a limited set of swear words and their derivatives, and how they are used on Twitter in the United States and the Nordic countries of Sweden, Norway, Denmark, and Finland. Two corpora of English tweets, originating from the US and the Nordic countries, were collected through the Twitter Streaming API, utilizing the programming language Python. The frequencies and relative frequencies of swear words in both corpora were calculated with the computer software WordSmith Tools, and compared with each other. The results from the relatively small samples indicate that people from the US swear more on Twitter, and also use more offensive swear words. Descriptions of numerous future research opportunities conclude this thesis.

Keywords: Corpus linguistics, swearing, Twitter, computer-mediated communication.

Tiivistelmä

Englannin käyttö kommunikaation kielenä ja Twitterin käyttö sosiaalisen median alustana ovat yleistyneet huomattavasti kaikkialla maailmassa, esimerkiksi Pohjoismaissa. Tutkijoiden keskuudessa Twitter on hyvin suosittu autenttisen kielimateriaalin lähteenä. Tämän lisäksi Twitterin ohjelmointirajapinta (engl. API) on helposti lähestyttävä, ja tietokoneohjelmien laatiminen suurien datamäärien keräämiseksi on suhteellisen vaivatonta. Silti, kiroileminen Twitterissä on todella vähän tutkittu aihe. Tämä tutkielma keskittyy rajoitettuun joukkoon kirosanoja ja niiden johdannaisia, ja tutkii niiden käyttöä Twitterissä Yhdysvalloissa ja neljässä Pohjoismaassa: Ruotsissa, Norjassa, Tanskassa ja Suomessa. Näissä maissa kirjoitetuista englanninkielisistä tviiteistä koottiin kaksi korpusta, hyödyntäen Python-ohjelmointikieltä ja Twitterin Streaming API:tä. Kirosanojen esiintymistaajuudet ja suhteelliset esiintymistaajuudet molemmissa korpuksissa laskettiin tietokoneohjelma WordSmith Tools -tietokoneohjelman avulla, ja saatuja lukuja verrattiin keskenään. Suhteellisen pienestä otannasta saatujen tulosten mukaan yhdysvaltalaiset ihmiset kiroilevat enemmän Twitterissä, ja käyttävät myös loukkaavampia kirosanoja. Lukuisten mahdollisten jatkotutkimusaiheiden kuvaukset päättävät tämän tutkielman.

Avainsanat: Korpuslingvistiikka, kiroilu, Twitter, tietokonevälitteinen viestintä

1. Introduction

A lecture I attended in 2016 introduced a unique perspective into corpus linguistics. Collecting a corpus from Twitter and performing quantitative analyses on the material seemed intriguing, especially since the approach incorporated programming into linguistics. I have the information processing science as a minor subject, so I valued the chance to use those skills as well. I have no personal experience of Twitter, and I am not an active user of social media, but social media research nevertheless seemed like a relative and applicable topic. Moreover, collecting a corpus from Twitter is, as it turned out, very quick and easy, even without profound knowledge concerning the platform. Swearing was one of the first things that occurred to me when thinking about possible units of analysis. I found no prior research regarding swear word usage on Twitter. That is, there was not much to refer to, which made the process feel difficult at times. On the other hand, it also implied that I have an original viewpoint on the subject. Swearing is an interesting phenomenon, and the multifunctionality of swear words makes them a very peculiar set of speech tokens. After starting this project, I heard about the Nordic research network SwiSca (Swearing in Scandinavia) and their annual Symposiums on swearing. This lead to a realization that swearing is also, as a research topic, current and applicable. As Beers Fägersten (2007) puts it, "profanity is a legitimate research area within psychology, philology and linguistics" (p. 15). It is also relevant to acknowledge Jay's (1992) bibliography with nearly 400 entries on swearing.

In the three subsections that follow, I will discuss previous research on Twitter and swearing, and outline the hypotheses of this thesis. Then, in section two, the data and methods of this study are described, and ethical considerations are discussed. The results are presented and discussed in section three. Section four, the conclusion, will present opportunities for further research.

1.1. Previous research on Twitter

Over the last decade, Twitter has become one of the most popular social media platforms on a global scale. Twitter users post status updates, up to 140 characters long, which are called tweets. In addition to text, tweets may contain e.g. images, links, or references to other tweets and Twitter users. As McCormick et al (2015) define, users can project their thoughts independently, communicate with each other through private messaging, re-tweet other tweets, and contribute to broader conversations by using # (hashtag) identifiers. "Twitter allows users to articulate "following" relationships, such that the tweets from those whom the user follows are displayed as a sequential feed that is updated in real time" (McCormick et al, 2015, p. 6). McCormick et al (2015) also call attention to the fact that on Twitter, unlike on various other social media platforms, users

most often receive a mix of public and private (user-approved followers') attention. This is likely to affect the form and content of the user's tweets, at least for users who are aware of the mix of attention. Tweets are by default public to all.

By accessing the Twitter API (Application Programming Interface) using programming scripts, it is relatively easy to collect tweets to be treated as a corpus. The API is a way of guiding the communication between two (or more) web platforms (McCormick et al, 2015). "When applied to web-based data collection, the API allows the researcher to specify which elements of information he or she wishes to retrieve from the primary platform" (McCormick et al, 2015, p. 7). Twitter, like many other platforms, has released its API for research use.

Although social media language is usually written, the register is commonly informal. Thus, the style of the language is closer to speech than in other written registers (Thelwall, 2008). Social media is therefore a very applicable source for authentic interactional language, strengthened further by the fact that large amounts of data can be collected easily. However, it should be mentioned that not all content on social media is informal, and the average nature of tweets vary by country, for example, as discussed later.

A tweet also contains metadata, such as a username, the location of the user, and the language of the text of the tweet. Thanks to these parameters, sociolinguists have found an interest in Twitter. Some studies focus on tweet language differences between genders (Coats, 2016b; Coats, 2017), others on the locations of Twitter users (Coats, 2016a; Coats, 2016c; Eisenstein & Pavalanathan, 2015; Mocanu et al., 2013). Twitter does not provide means to supply gender information, but gender disambiguation is possible through other means (see e.g. Coats, 2016b). This study takes advantage of the 'geotags' within the tweet data structure, and only tweets from certain areas are analyzed.

In this study, tweets from the following countries are under scrutiny: Finland, Sweden, Norway, Denmark, and the United States (although Iceland, traditionally considered a part of the Nordic, is absent in this study, the four North European countries are occasionally referred to as the Nordic countries later on). The North European countries were chosen simply because of proximity to myself. It is also worth pointing out that Sweden is one of the leading countries in Twitter adoption: in a listing of countries with the most Twitter users per capita, Sweden is number nine (Mocanu et al., 2013). Although "most content produced within each country is written in its own dominant language" (Mocanu et al., 2013, p. 4), English is the most common tweet language on a global scale (probably because of attempts to reach a broader audience). This fact is strongly present in Nordic Twitter as well. In fact, Twitter users from Denmark, for example, send more tweets in English than in their native language (Coats, 2017). Therefore, analyzing only English

tweets from North European countries is not a big issue. The US tweets are analyzed to explore possible differences caused by geographical distance and by the fact that English is the dominant language of the area.

The book chapter "Grammatical feature frequencies of English on Twitter in Finland", written by Coats (2016a), is especially relevant as a reference study. The study characterizes the nature of Twitter English in Finland, comparing it to the Twitter English of the whole world. The findings state that Finns use more emoticons and expressive lengthening (e.g. *nooooo*, *goood*), and that Finnish tweets are shorter overall and contain fewer long words. This suggests that English on Twitter in Finland is less informative and more interactive. That is, the style tends to be more casual and conversational. Coats (2016a) also states that Twitter English in Finland is "a variety in which expression of affective stance plays an important role" (p. 203). Of course, the study's findings cannot be generalized to apply to all the Nordic countries, but a degree of likelihood in Twitter English within areas with little geographical distance is probable. Unfortunately, no study of this kind has been made concerning English on Twitter in the US, at least to my knowledge. However, grades of Twitter adoption are very high in the US (Mocanu et al., 2013), and therefore a large portion of the world's English tweets are likely to originate from North America. Consequently, while presenting the Comparison English Corpus he used in the study, Coats states that "an examination of the messages suggests that a relatively high proportion of the tweets originate from the United States" (2016a, p. 186). Therefore, the study more or less compares the Twitter Englishes of Finland and the US. It can then be deduced that tweets from the US, as opposed to Nordic tweets, are more informational, and with narrative objectives.

In summary, the studies by Mocanu et al (2013) and Coats (2016a) are especially relevant for this thesis. It is essential to remember how widespread English is on Twitter, and that English on Twitter in Finland is relatively casual and interactive.

1.2. Previous research on swearing

As Dewaele (2004) defines,

> ST-words [swear- and taboo words] are multifunctional, pragmatic units which assume, in addition to the expression of emotional attitudes, various discourse functions. They contribute, for instance, to the coordination of turn-taking between the interlocutors, the organization of the interaction, and the structuring of verbal exchange; in that, they are similar to discourse markers ... The use of ST-words is also a linguistic device used to affirm in-group membership and establish boundaries and social norms for language. (p. 84-85)

Swear words are commonly used for emotional release in response to sudden pain or bad news, and to frustration and anger (Thelwall, 2008). ST-words will be referred to as swear words from now on. To a large degree, swear words have lost their original meaning, since they are in such frequent use (Stenström, 2006). Moreover, swear words do not function solely as interjections, but also as, for example, nouns and verbs, integrated into the clause structure (Stenström, 2006).

Instructed language learners have little knowledge of swear words, since swears are considered too offensive for the classroom (Dewaele, 2004). While the people in the Scandinavian region acquire their strong base knowledge of English in school, this knowledge is strengthened for example through consumption of films, TV-series, books, online content etc. in the English language. Therefore, Nordic learners of English are likely to become familiar with the most common English swear words, at the least. Still, swear word usage on an individual level is typically different in different languages. As Dewaele (2004) characterizes, advanced speakers use more emotion words, and swear words in the L1 have more emotional force. Therefore, L1 is a common choice for swearing. On the other hand, some people feel that L1 swear words are too powerful, and thus avoid using them. In cases like this, other-than-first languages are used, since the swear words have smaller emotional impact and can be used more freely. Escaping the "restrictive social conventions of [one's] native culture" (Dewaele, 2004, p. 96) is a possible motivation. Overall, as Dewaele states, swearing is most common in the speaker's most dominant language.

Thelwall's (2008) study on swearing in MySpace is an interesting recourse, even though its validity can be questioned because of the study's age. Developments on the web are known to be very rapid, and MySpace has lost millions of users since 2008, mainly to Facebook (see e.g. Chandler, 2011). Nevertheless, Thelwall's general findings about swearing on social media should not be deemed irrelevant. Firstly, the informal context supposedly leads to more swearing. Social media is a very fitting environment for discussions between friends about subjects that cause anger and frustration. Secondly, social media can be an important site for identity expression, and a part of it is, at least among young people, using swear words. Then again, "[a] member may take care when writing comments in the knowledge that shared friends are likely to read them and judge them" (Thelwall, 2008, p. 91).

Beer Fägersten (2007) has studied swear word offensiveness by surveying University students in Florida. The survey used in the study listed twelve frequently used swear or taboo words, and the results revealed that *nigger* is considered the most offensive, followed by *cunt*, *motherfucker*, *bitch*, and *fuck*. The participants also listened to six recorded conversations which included swear words. In the recordings, *fuck* was by far the most frequent swear word, specifically in its form *fucking*, used as an adjectival or adverbial intensifier. Beers Fägersten points out that the

perceived offensiveness of swear words largely depends on the context of usage. Swears listed on paper, without context, are deemed more offensive than the same words occurring in conversation, and "literal uses of the swear words are more offensive than the non-literal uses of the same words" (Beers Fägersten, 2007, p. 28).

The hypotheses, described in the next section, rely on four observations from Dewaele (2004) and Thelwall (2008). Instructed language learners lack the swear word knowledge of L1 speakers, and advanced speakers use more emotion words (i.e. swear). The context of social media is often informal, and it can be an important site for identity expression.

1.3. Hypotheses

A relevant consideration is the fact that the length of tweets most likely affects swear word usage. It is possible that many people do not want to 'waste' the very limited amount of allowed characters on foul language. Moreover, the United States might at first seem intolerant of swearing, mainly because the Federal Communications Commission (FCC) has rather strict rules regarding indecent and profane content on broadcast television. Albeit, "[t]he relatively prudish nature of US television does not, however, imply that the attitude of the population is prudish[…] Moreover, swearing seems to be common in the US" (Thelwall, 2008, p. 90). Many American media networks publish tweets through their Twitter accounts, and possibly follow the FCC guidelines, but most tweets are written by ordinary, swearing individuals. Swearing and usage of slang is characteristic of teenage talk (Stenström, 2006) and Twitter has a slightly younger user base than e.g. Facebook or YouTube (Hutchinson, 2017). It is then very possible that young users publish a large portion of the tweeted profanities, and have a large effect on the total number of swear word occurrences. As a summary of the previous subsections, it is important to remember the following aspects.

Firstly, English on Twitter in Finland is interactive and informal. Again, this statement cannot be, without reserve, generalized to affect all the Nordic countries, but it is not far-fetched to suppose so. Finnish Twitter users use relatively large amounts of emoticons and structures with expressive lengthening. This fortifies the notion that the Twitter English in question is conversational and casual in nature. One could then deduce that swear words would probably be frequent, due to the informal context.

Secondly, people most commonly swear in their dominant language. L1 swears are considered to have more emotional impact, and emotional release is just what swear words are typically used for. This would in turn suggest that swear words would appear more frequently among the US Twitter users, since the swearing language is their mother tongue. Moreover, the

Americans have more knowledge, experience, and fluency in English swear word usage. Larger amounts of confidence probably lead to larger amounts of swear words.

This study tries to find proof for these points. Does the supposed interactive nature of Nordic Twitter Englishes prompt increased swear word occurrences? Does writing tweets in one's dominant language increase or decrease swear word occurrences? I hypothesize that the supposed informal context of Nordic Twitter English and the usage of the dominant language in the American Twitter English balance the number of swear word occurrences, so that the relative amounts of swear words are close to each other between the corpora. However, Twitter English in the US apparently has a narrative composition, which could affect swear word occurrences negatively. Still, larger amounts on either geographical area would not be surprising. The subject of swear word offensiveness, based on Beers Fägersten (2007), will be touched upon. However, it is difficult to predict which geographical group uses more offensive swear words, since people from different cultural and linguistical backgrounds might think differently about swear word offensiveness. Again, the confidence of the American Twitter users could imply more offensive swearing. On the other hand, the Nordic Twitter users may appear more tolerant of offensive language, since the language is not their own, and they lack the amount of language knowledge the American users have.

2. Data and methods

The Twitter data was collected with the programming language Python and its module Tweepy (Roesslein, 2015). The Twitter Streaming API, which provides the user with tweets as they are broadcast in real time, was accessed on the end of October. The default Streaming API access is limited to 1% of the traffic on the platform (Twitter offers commercial alternatives for better access), but since the volume of traffic is so extensive, the Streaming API still provides a way of collecting a considerable corpus relatively quickly. Two Python scripts selected geo-tagged tweets originating from within two different geographical 'bounding boxes', one of which circumscribed the borders of the United States (24 - 50 N° and 66 - 125 W°), and the other the borders of the four Nordic countries (54 - 71 N° and 5 - 32 E°), which were selected mostly due to proximity. The tweets were initially stored in JSON format (JSON - Introduction, n.d.), with a single tweet on a single row, and with the numerous metadata attribute-value pairs one after another. The output consisted of 56 813 tweets, with 27 661 originating from North America and 29 152 from Northern Europe. The data required about 0.2 gigabytes of storage space.

The next pair of Python scripts manipulated the tweet data and filtered the dozens of metadata attributes which are not relevant to this study. The remaining three fields, *lang* (the

language of the tweet), *country_code* (the country of the tweet based on user input), and *text* (the actual written message to be analyzed) were written to a text file. The *lang* field is based on BCP 47 language identifier (Davis & Phillips, 2009), so it is not necessary to doubt the value of the attribute. However, since the *country_code* depends on user input, the reported country might be incorrect. The tweets originate from within a specified geographical area, though. Moreover, tweets that had a *country_code* value other than 'US', 'FI', 'SE' 'DE' or 'NO' were removed during the filtering process. These aspects then imply that the reported country cannot be too far. In fact, the difference between the GPS-tag (coordinates automatically attached to individual messages) data and the user input-based geolocation data relates more to demographics: GPS-tagged tweets are usually written by women and young people, and the text-based location is more accurate for men and for older people (Eisenstein & Pavalanathan, 2015). The third script pair accepted only English tweets from the five specified countries, and these tweets were written to a new text file. Filtering by country only reduced the number of tweets to 26 283 for the US and 17 523 for the four Nordic countries. When the non-English tweets were removed, the numbers for the US and the Nordic countries reduced to 22 910 and 5 673, respectively. This data sets will be referred to as US corpus and Nordic corpus.

Especially when compared to other Twitter studies, the data amounts of this study seem very small. The reason for this simple: I do not have the time or resources to let the data pile up for months, weeks nor even days. As a comparison, Coats (2016a) collected the Finnish tweet data for over a month and received 32 916 English tweets. For this study, the Nordic tweets were collected in less than 20 hours, and only 5 673 English tweets were received. The amount is rather small at least in comparison to the US corpus. Therefore, the Nordic tweets are not further divided into even smaller subsets. That is, the four countries within the Nordic corpus are not analyzed separately, but rather as a single unit. The relatively small corpora directly affect the credibility of my research, and the results presented later are not nearly as reliable as the results from studies with larger amounts of data. Nevertheless, the results bring about some insight to the topic of swearing on Twitter. Generalizations cannot be made, but at the least this study lays foundations for further research.

Based on calculations performed with the programming language R, the English Twitter data contains 368 129 word tokens, 294 761 of which are in the US corpus and 73 368 in the Nordic corpus. It should be noted that all character strings separated by blank spaces are considered as words. These also include emojis, URLs, and other data that would not usually be treated as lexical items. Treating these items as words is debatable, but since the focus of this study is essentially the amount of swear words among all other material, such debates are not addressed

here. In possible further research, the corpus could be prepared more thoroughly, meaning that non-lexical items are removed or clearly marked.

The next task is finding and counting the occurrences of swear words. The relative frequency distribution of swear words for the US and the Nordic countries will be presented and compared with each other in section 3. The swear words under scrutiny are limited to seven words, including the following: *ass*, *bastard*, *bitch*, *damn*, *fuck*, *hell*, and *shit*. These words were intuitively selected based on Beers Fägersten's (2007) list and on comedian George Carlin's monologue "Seven Words You Can Never Say on Television" from 1972. The set is not exhaustive, of course, but rather represents examples of the most frequent swear words. Forthcoming uses of the term 'swear word(s)' refer to these seven-words and their derivatives.

The WordSmith Tools software was used for the swear word counting. Numerous searches were done in the software's 'Concord' function, which finds all the instances of a given word or phrase. In addition to the base forms of the swear words, numerous inflected forms and derivatives, such as *fucking* and *Goddamn*, were also searched, along with a few examples of non-standard spellings, such as *f**k* and *fuuuck*. On occasion, the context of the occurrence (i.e. surrounding words) was inspected to make sure that the occurrence really is a swear word. Of course, it was not possible to search for every possible variation. There are countless ways of masking and spelling swear words, and individuals can easily invent new methods. Consequently, the swear word amounts are most likely underestimates.

As mentioned above, everything but the language, country code and the tweet text are removed from the data before any analysis is done. Therefore, metadata about the tweet author cannot be inspected. Even if identification information was preserved, tweets are by default public, and the Streaming API naturally collects only public tweets. Asking the authors for a permission to use their tweets in this study would then be unnecessary for two different reasons, and probably impossible as well. Moreover, this study only focuses on the amount of swear words related to the amount of all words. This indicates that the context and the content of the tweets are left unconsidered. On some occasions, the collocates of a found swear word are inspected, but only to the extent of verifying the said word's nature as a swear word. Recognizing authors, even by reading the full tweet texts, is not possible, and even if it was, it is good to remember that all the tweets under scrutiny are published as public.

3. Results

In this section, the search queries and their results will be presented. The relative frequency distribution of swear words in the US corpus and Nordic corpus are presented and discussed in

sections 3.1. and 3.2., respectively. Then the relative frequencies are presented side by side in a single table, and comparisons are made and discussed in section 3.3.

A total of 26 different search words were used in WordSmith Tools -software's 'Concord' function in order to find the plural forms and other derivatives of the swear words in addition to the basic forms. For *fuck*, a few non-standard spellings were also searched. The reason for this is that the swear word is very common, although it is at the same time considered relatively offensive. Therefore, the likelihood of this particular word appearing in an 'masked' or 'censored' form is considerably high. Instances of non-standard spellings, along with instances of *ass* and *bitch*, were confirmed as intended swear words with quick concordance inspections. No occurrences of *ass* and *bitch* referred to donkeys or female dogs, respectively. Compounds which include a swear word were not searched for (excepted for the surprising discovery, *facefuck*, which did not yield considerable results), since the amount of possible combinations is in principle limitless. There is no way of compiling a totally representative collection of compound swear words. As an aside, an interesting point, found during the concordance inspection, was made within the tweet material: "'F*ck' is not less offensive than 'Fuck'." The amounts of swear word occurrences are presented in Table 1.

Table 1. Frequencies of all found swear words in both corpora.

Swear word	**# of occurrences in US corpus**	**# of occurrences in Nordic corpus**
fuck	218	42
fucked	42	7
fucking	164	34
fucker (*)	11	2
f**k (**)	0	0
f*ck (**)	1	1
fuuuck	3	0
facefuck	1	0
f—k	0	0
ass	201	12
asses	7	0
dumbass	9	1
bastard	1(***)	0
bastards	2	1
bitch	137	9
bitches	39	1

son of a bitch	0	0
damn	125	21
damned	6	0
darn	2	2
goddamn	5	1
hell	89	12
hellish	0	0
shit	331	44
shits	8	1
shitted	1	0
shitting	1	0
shat	0	0
shite	0	0
Total	1 404	191

(*) A few cases of *fucker* were found within the word *motherfucker*.

(**) These words as search queries provided every character string that began with *f* and ended in *k* or *ck*. Words such as *Facebook* and *feedback* also turned up, but the query also returned an unpredicted swear word *facefuck*.

(***) *Bastard*, as it turned out, occurred within the beer brand name "Backwoods Bastard", and could not then be considered a real swear word.

As the occurrence amounts show, the common swear words in their singular form (except for *bastard*) and standard spelling are the most frequent. The words that did not occur in the corpora are naturally left out of further analysis. Furthermore, words that did not occur in the Nordic corpus are also not considered. The reason behind this choice is the fact that it is impossible to know whether the non-occurrence happened simply due to the small sample size. Since the 'reality' of the non-occurrence cannot be confirmed, it is the safest to drop the associated words. The preserved swear words, and their frequencies, can then be seen in Table 2.

Table 2. Frequencies of selected swear words in both corpora.

Swear word	**# of occurrences in US corpus**	**# of occurrences in Nordic corpus**
fuck	218	42
fucked	42	7
fucking	164	34
fucker	11	2
f*ck	1	1
ass	201	12

dumbass	9	1
bastards	2	1
bitch	137	9
bitches	39	1
damn	125	21
darn	2	2
goddamn	5	1
hell	89	12
shit	331	44
shits	8	1
Total	1 384	191

The relative frequency distribution of swear words in the US corpus and in the Nordic corpus will be presented and discussed in sections 3.1. and 3.2., respectively. Comparisons of results from the two corpora are performed in section 3.3.

3.1. The relative swear word frequencies in the US corpus

The relative frequencies for the swear words within the American tweets are presented in Table 3.

Table 3. Frequencies and relative frequencies of swear words in the US corpus.

Swear word	**Frequency**	**Relative frequency (%)**
fuck	218	0,07396
fucked	42	0,00238
fucking	164	0,05564
fucker	11	0,00373
f*ck	1	0,00034
ass	201	0,06819
dumbass	9	0,00305
bastards	2	0,00068
bitch	137	0,04648
bitches	39	0,01323
damn	125	0,04241
darn	2	0,00068
goddamn	5	0,00170

hell	89	0,03019
shit	331	0,11229
shits	8	0,00271
Total	1 384	0,46953

Almost half a percent of all words in the American tweet corpus are swear words, as can be seen in the 'Total' row. This roughly means that every 200th word is a swear word. It could be concluded that swearing is not a salient habit with Twitter users from the United States. However, comparing the result with the frequency of the most common word in the English language gives a different perspective. The word *the* has a frequency of five percent in all English, i.e. it occurs five times in hundred words and ten times in 200 words. Ten versus (roughly) one is not a great difference, especially in light of the fact that the 0,47 percent is likely an underestimate. There are many more swear words, and several other spellings for the words included in the half a percent. It is also probable that swear words occur more frequently in e.g. speech than on Twitter.

The popularity of swear words is then the following (in decreasing order): *fuck* and its derivatives (436 occurrences, relative frequency 0,15%), *shit* and *shits* (339; 0,12%) *ass* and *dumbass* (210; 0,07%), *bitch* and *bitches* (176; 0,06%), *damn* and its derivatives (132; 0,04%), *hell* (89; 0,03%), and *bastards* (2; 0,0007%). Reflecting Beers Fägersten's (2007) study, the offensiveness of these swear words would be the following (in decreasing order): *fuck* and *bitch* (equally offensive), *bastard*, *ass*, *shit*, and *damn* and *hell* (equally offensive). There does not seem to be a strong correlation between the frequency and the offensiveness. The most offensive word is also the most frequent, whereas the second to most offensive is the least frequent. These two lists do have similarities, e.g. that the least offensive swear words are also low on the frequency scale. However, it could be expected that the most offensive words would be the least used, but this is clearly not the case.

3.2. The relative swear word frequencies in the Nordic corpus

The relative frequencies for the swear words within the Nordic tweets are presented in Table 4.

Table 4. Frequencies and relative frequencies of swear words in the Nordic corpus.

Swear word	**Frequency**	**Relative frequency (%)**
fuck	42	0,05725
fucked	7	0,00954
fucking	34	0,04634

fucker	2	0,00273
f*ck	1	0,00136
ass	12	0,01636
dumbass	1	0,00136
bastards	1	0,00136
bitch	9	0,01227
bitches	1	0,00136
damn	21	0,02862
darn	2	0,00273
goddamn	1	0,00136
hell	12	0,01636
shit	44	0,05997
shits	1	0,00136
Total	191	0,26033

Based on these numbers, every 400th word is a swear word in the Nordic corpus. Again, swearing does not seem to be very common, though it is good to remember that the frequencies are most likely underestimates. This is increasingly true in the case of the Nordic corpus, since the corpus is significantly smaller. Swear word popularity in decreasing order is the following: *fuck* and its derivatives (86 occurrences, relative frequency 0,12%), *shit* and *shits* (45; 0,06%), *damn* and its derivatives (24; 0,03%), *ass* and *dumbass* (13; 0,02%), *hell* (12; 0,02%), *bitch* and *bitches* (10; 0,01%), and *bastards* (1; 0,001%). The list is similar to the one from the US corpus, except for *damn* with its derivatives and *hell*. These mildly offensive words close to the top, and relatively offensive words (*bitch*, *bastard*) are the least used. The Nordic people seem to use less offensive swear words on Twitter, an observation which is of course breached by the high frequency of *fuck*.

3.3. Comparison of swearing on Twitter in the US and the Nordic

The easiest observation to make between the swearing data from the US and the Nordic countries is the difference in total relative frequencies. The 0,47% of the US corpus is much higher, almost twice as high, than the 0,26% of the Nordic corpus. Of course, the US corpus is over four times as large as the Nordic corpus, which might easily explain the difference. However, it could be supposed that a quadruple increase in corpus size should result in a quadruple increase in swear word frequency. Nevertheless, based on the data and results at hand, the conclusion is that Americans swear more on Twitter than Nordic people do. This most likely relates to the effect language dominance has on swearing (Dewaele, 2004). Americans use the language in question

most of the time, which entails that they are exposed to more swearing, and swear more in English. Knowledge about and confidence regarding swear words usage surely has an impact. This result could be surprising to some, since the apparent prudish nature of American television might call to question the commonness of swearing in the US. Apparently, this cultural feature is trumped by the context of social media.

The (supposed) informal context of Nordic Twitter English does not weight as much as language dominance. Nordic Twitter English may be more casual, and have e.g. more emoticons and cases of expressive lengthening (Coats, 2016a), or other features associated with conversational texts. Still, swear words might not be one of those features. Another view is that swear word frequencies do not reflect the nature of Twitter English, be it conversational or informational. The hypothesis of the amounts of swear words being balanced between the two corpora has to be rejected.

The relative frequencies of swear words in the US corpus and Nordic corpus are presented side by side in Table 5.

Table 5. Relative frequencies of swear words in the US and the Nordic.

Swear word	**Relative frequency, US (%)**	**Relative frequency, Nordic (%)**
fuck	0,07396	0,05725
fucked	0,00238	0,00954
fucking	0,05564	0,04634
fucker	0,00373	0,00273
f*ck	0,00034	0,00136
ass	0,06819	0,01636
dumbass	0,00305	0,00136
bastards	0,00068	0,00136
bitch	0,04648	0,01227
bitches	0,01323	0,00136
damn	0,04241	0,02862
darn	0,00068	0,00273
goddamn	0,00170	0,00136
hell	0,03019	0,01636
shit	0,11229	0,05997
shits	0,00271	0,00136

Total	0,46953	0,26033

The number of occurrences of *fuck* and *shit* (in their base form) are distinct in both corpora. *Shit* is clearly the most common in both corpora, when the frequencies are inspected individually. However, in the Nordic corpus, the difference between *fuck* and *shit* is only about 0,003 percentage points, whereas the same difference is about 0,04 percentage points in the US corpus. This feature is curious. Perhaps the reason is that *fuck*, which is, as a swear word, immensely common, might be the only swear word non-native speakers know and recognize. Of course, education and being exposed to English leads to wider vocabulary. Yet, if one asks a Nordic person with average English skills to list English swear words, the list would arguably be formed mostly of *fuck* and its derivatives. The Nordic person might not consider words like *ass* and *damn* as swear words, but this of course depends a lot on the person. In the US corpus, the relative frequency of *fuck* and its derivatives is 0,15%, and in the Nordic corpus it is 0,12%. The frequencies are remarkably close. Apparently, the different versions of *fuck* are common in all varieties of English.

The popularity lists of both corpora are exactly alike for the first, the second and the last swear words. The third place in the US corpus is taken by *ass* and *dumbass*, the third most offensive of these swear words. However, the third place of the Nordic popularity list is reserved for *damn* (and its derivatives), the least offensive swear word. *Ass* is found in the Nordic list at rank four, which is the rank of *bitch* and *bitches* in the US list. *Bitch*, *bitches*, and *bastards*, the most offensive swear words along with *fuck*, are at the very bottom of the Nordic popularity list. It is then possible to conclude that the US corpus retains more offensive swearing than the Nordic corpus.

The confidence of the American Twitter users is most likely visible in this aspect too. People from the United States have more fluency in articulating their emotions (i.e. frustration) through swearing, which is likely to affect the swear word choices. Moreover, they possibly also have a deeper understanding of the emotional strength behind the various swear words. The ideas of using swear words with *less* emotional force and escaping the social restrictions of one's native culture (Dewaele, 2004) do not hold ground here. There may be many reasons for the Nordic people to choose milder swear words. Perhaps the lower levels of English fluency make the Nordic people avoid swear words that carry a meaning of personal insult. In other words, Nordic Twitter users choose general swear words just to be safe. In-depth studies about the correlation (or non-correlation) between swearing frequency and swearing offensiveness would be very useful and interesting.

4. Conclusion

The results of this study answered the research questions from section 1.3. Although the context of Nordic Twitter English is supposedly informal, swear words are considerably more frequent in the US corpus. In addition, Americans use offensive swear words more frequently. Therefore, the answer to the question 'does the supposed interactive nature of Nordic Twitter Englishes prompt increased swear word occurrences?' is 'no'. Further, the answer to 'does writing tweets in one's dominant language increase or decrease swear word occurrences?' is 'increases'.

This study could function as a starting point for numerous different approaches towards the topic of swearing on Twitter. The first aspect that should be considered in future research is naturally the corpus size. Larger sample sizes would yield more reliable results, and even provide grounds for generalizations. Moreover, larger amounts of Nordic tweet data would make it worthwhile to inspect the swearing (or other features) on the level of individual countries. The United States could even be left out of the comparison, and the researcher could only study swearing on Twitter in Sweden, Norway, Denmark and Finland. Iceland should also be included in the study, if the researcher is interested in the Nordic countries. Naturally it is also possible to investigate the differences in Twitter English inside a single country, and see, for example, whether swearing is more common in the Northern or in the Southern regions. Alternatively, the geographical areas could be expanded even more, to include the whole of Europe and the whole of the Americas, or even the whole world. Nevertheless, if comparisons of two (or more) geographical areas are made, the amounts of English tweets should be close to each other.

Another expandable element is the list of swear words. There are many more words that could be included in the study. The researcher could include compounds and other inflections, or perhaps only search for all forms and compounds of a single swear word. The frequencies of different forms and their relations with each other deserve more attention. Multilinguals might also use other than English swear words in their English tweets, which could also be examined. Of course, there is no reason why other languages could not be included and inspected in the study.

Like many other Twitter studies, a study of swearing on Twitter could also use gender as a variable, in addition to language and location. The resulting investigation would be like a combination of Coats (2016b or 2017) and Thelwall (2008). Males and females generally swear differently, as pointed out in Thelwall (2008). Possible research questions in a study of this type could be e.g. 'do American males swear more than American females? How about Norwegian males and females?' or 'do Finnish females use more offensive swear words than Swedish females?' The immense amount of metadata in every tweet grants a wide variety of possible variables to consider. For example, the *created_at* fields could reveal how the time of day affects

language use, the *source* fields could tell what kind of tweets are sent from which devices, and with help from the *favorite_count* fields the linguistic structure of popular tweets could be figured out.

The styles of Twitter English in the US and in other Nordic countries should also be investigated. This study only supposed that English on Twitter in the Nordic countries is casual and interactional. If this supposition was proved, and a similar examination (as in Coats, 2016a) was performed on the US, reliable inspections could be made about how the nature of context affects swearing. If the assumption that Nordic Twitter English is more conversational than American Twitter English turns out to be true, an interesting question emerges. Why does English on Twitter in the US retain more swearing, if the context is more formal and narrative?

Considering the amount of research that linguists, social scientist and others have done with and about Twitter, it is very curious how swearing on Twitter is so little examined. It is obviously nice to be in the front lines, but the subject deserves without a doubt more attention. It is very probable that Twitter continues to function as a common, contemporary source for research in various disciplines. It is also probable that someone eventually takes on the topic of swearing on Twitter.

The frequency and relative frequency of swear words in this paper are presented in Table 6.

Table 6. Swear word frequency in the present study.

Swear word	**Frequency**	**Relative frequency (%)**
Any swear word	175	0,02425

References

Beers Fägersten, K. (2007). A sociolinguistic analysis of swear word offensiveness. *Saarland Working Papers in Linguistics*, *1*, 14-37.

Chandler, T. (2011, March 31). The Death of MySpace. Retrieved from https://www.youngacademic.co.uk/features/the-death-of-myspace-young-academic-columns-953

Coats, S. (2016a). Grammatical feature frequencies of English on Twitter in Finland. In Lauren Squires (Ed.), *English in computer-mediated communication: Variation, representation, and change* (p. 179–210). Berlin: de Gruyter Mouton.

Coats, S. (2016b). Grammatical frequencies and gender in Nordic Twitter Englishes. In D. Fišer and M. Beißwenger (Eds.), *Proceedings of the 4th conference on CMC and social media corpora for the humanities* (p. 12–16). Ljubljana: U. of Ljubljana Academic Publishing.

Coats, S. (2016c). *Nordic Englishes on Twitter* [PDF document]. Retrieved from http://cc.oulu.fi/~scoats/DHiNCoats.pdf

Coats, S. (2017). Gender and grammatical frequencies in social media English from the Nordic countries. In D. Fišer and M. Beißwenger (Eds*.), Investigating social media corpora* (p. 102–121). Ljubljana: U. of Ljubljana Academic Publishing.

Davis, M. & Phillips, A. (2009). BCP 47 - Tags for Identifying Languages. Retrieved from https://tools.ietf.org/html/bcp47

Dewaele, J. (2004). Blistering Barnacles! What language do multilinguals swear in? *Estudios de Sociolingüística*, *5*(1), 83-106.

Hutchinson, A. (2017, March 21). Top Social Network Demographics 2017. Retrieved from https://www.socialmediatoday.com/socialnetworks/

Jay, T.B. (1992). *Cursing in America*. Philadelphia: John Benjamins Publishing Company.

JSON - Introduction. (n.d.). Retrieved from https://www.w3schools.com/js/js_json_intro.asp

McCormick, T.H.; Lee, H.; Cesare, N.; Shojaie, A.; & Spiro, E.S. (2017). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods & Research, 46*(3), 390-421.

Mocanu, D.; Baronchelli, A.; Perra, N.; Gonçalves, B.; Zhang, Q.; & Vespignani, A. (2013). The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE*, *8*(4), p. 1-9. https://doi.org/10.1371/journal.pone.0061981

Pavalanathan, U. & Eisenstein, J. (2015). Confounds and consequences in geotagged Twitter data.

Retrieved from http://arxiv.org/abs/1506.02275

Roesslein, J. (2015). Tweepy. Python package [computer software]. Retrieved from www.tweepy.org

Stenström, A.-B. (2006). Taboo words in teenage talk: London and Madrid girls' conversations compared. *Spanish in Context*, *3*(1), p. 115-138.

Thelwall, M. (2008). Fk yea I swear: cursing and gender in Myspace. *Corpora*, *3*(1), p. 83-107.