Universität Potsdam

Philosophische Fakultät

Masters Thesis

Swearing in a public place

On the usage of swear words on reddit

Verfasser: Jonas Thomä

Deutsche Zusammenfassung

Die vorliegende Arbeit behandelt das Vorkommen von Schimpfwörtern auf der online Plattform "Reddit". Die drei zugrundeliegenden Forschungsfragen sind:

Wie oft werden Schimpfwörter benutzt?

Wie werden diese von den Lesern aufgenommen?

Beeinflusst das Thema einer Konversation die Reaktion der Leser und die allgemeine Häufigkeit der Nutzung?

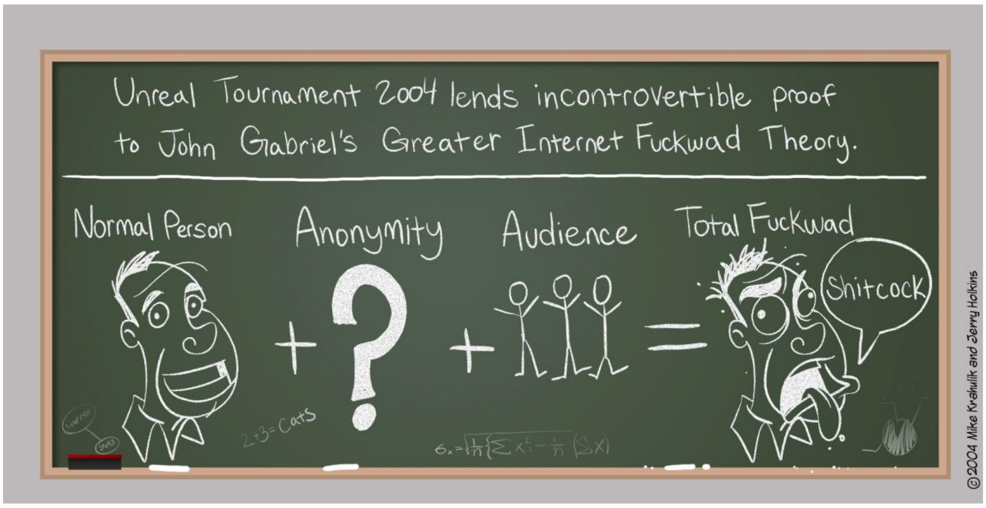
Die zugrundeliegenden Daten beinhalten fast 900 Millionen Wörter und stammen aus dem Februar 2017. Sie sind damit höchstaktuell und repräsentativ. Im Vergleich zu anderen Untersuchungen ist das Korpus damit wesentlich größer.

Zusätzlich werden im theoretischen Teil die linguistischen Grundlagen zu Schimpfwörtern erörtert. Dazu gehören u.a. Konzepte wie die Höflichkeitstheorie, das Thema Tabu und die dazugehörenden Worte und Zensur. Dies wird getan um die Faktoren, die die Benutzung und Verwendung von Schimpfwörtern beeinflussen darzulegen. Dabei wird herausgestellt, was Schimpfwörter so besonders im Vergleich zu anderen Wortgruppen macht. Zudem werden weitere Forschungsergebnisse, die aus anderen Korpora stammen dargelegt und hinterher mit den Resultaten verglichen. Dies beinhaltet Korpora die sich ebenfalls aus Onlinekommunikationen zusammensetzen, sowie Korpora die gesprochene Sprache wiedergeben. Die Ergebnisse aus allen dargestellten Korpora behandeln Ergebnisse aus der englischen Sprache.

Die Ergebnisse dieser Studie weisen daraufhin, dass die Schimpfwörter auf Reddit ungefähr gleichhäufig wie auf anderen Plattformen benutzt werden. Die Reaktionen auf diese Schimpfwörter ist überdurchschnittlich positiv, was darauf schließen lässt, dass die Benutzung von Schimpfwörtern auf Reddit nicht als unhöflich aufgefasst wird. Zudem konnte ein Einfluss des Diskussionsthemas auf die Häufigkeit und Rezeption von Schimpfwörtern festgestellt werden.

Table of Contents

1. Introduction


Unreal Tournament 2004 lends incontrovertible proof to John Gabriel's Greater Internet Fuckwad Theory.
Normal Person
Anonymity
Audience
Total Fuckwad
Shitcock
+
+
=
2+3=Cats
©2004 Mike Krahulik and Jerry Holkins

(Krahulik and Holkins 2004)

The theory above states what can be regarded as a general layman's perception about communication on the internet. People think that it is obscene and profane. With the anonymity that the internet provides to its users, repercussions for swear word usage seem minimal. With the surge of the internet new forms of communication were introduced. These new forms of communication are bound to develop their own norms of how to speak. This includes swear word usage. People from all over the world have access to these platforms. They influence how people communicate. Reddit is one of those platforms. It was founded in 2005 and has surged to be the fourth most visited website in the US. It is currently the seventh most visited website globally (cf. Alexa.com 2017). An international group of users speak in a language that is foreign to many of them. This is bound to have an effect on the way they communicate.

Swear words and taboo are a special field in the study of communication. They have many features that are unique. They were and are still censored in mass media. They can be used to express emotion and are sometimes uttered unintentionally. They are often considered to be impolite but every member of the society uses them, knows about them and has some sort of understanding when and when not to swear. They are sometimes inappropriate despite the fact that their meaning would be correct. If, for example, a cook touches a hot oven and yells "Fuck me, that's hot!" in the kitchen, this might be inappropriate if the restaurant manager stands next to him.

The research into the use of swear words is also fairly new. It only started in the 1960s (cf. Ljung 2011: 3-4). Thus, the data that is available for linguistic research is fairly limited, especially, since a lot of the research is done from a mainly psychological view. Data on the overall usage, on how much of everyday speech they make up is limited. Thus, more information on the matter can only provide further insides into a special field of linguistic research.

The corpus of this study is taken from the comment section of the website "reddit.com". It is a news aggregation website on which contents and links can be discussed in a comment section. The data is publically available. It is particularly interesting as the data is very recent, in this case from February of 2017, and very large. The corpus for this study is around 900.000.000 words. It provides a very actual and representative insight into the use of swear words. The structure of reddit allows its users and non-registered visitors to view these conversations. Thus, the discussions are public and everyone who desires to do so can see them. Users on reddit do not provide much information about themselves and remain anonymous with the exception of their username. Internet corpora also provide useful information on swear word usage opposed to staged experiments where participants are aware that they are part of a study. In the case of reddit and this corpus, the language is genuine and not affected by those factors. Other aspects that might influence the swear word usage will be discussed in detail.

The present study is exploratory. With the features and circumstances that reddit provides, three main research questions are formulated:

How often are swear words used?

How are they received?

Do conversational topics influence the amount of swear words used and their reception?

Before the questions are answered, theoretical backgrounds to swear words and connected topics will be discussed. Afterwards, recent studies on swear words in both spoken and online corpora will be presented and discussed. Furthermore, the features that are unique to reddit demonstrated to show what data can be gathered and how it can be interpreted. After discussing theoretical backgrounds that should show the different unique aspects of swear words in greater detail and providing more information on swear word usage in other corpora, the research questions will be divided into a subset of research questions which are based on the previously discussed information. Based on those questions, the methods and results will be discussed.

Before the broader backgrounds that influence swear word usage, like politeness theories and the concept of taboo are expounded, a short introduction to the general subject of swear words will be presented.

2. Linguistic background

2.1. On swearing

As swearing is influenced by a number of factors that need to be explained, it is important to first define what swearing and swear words are. The term *swearing* is ambiguous. It can mean swearing an oath or using profane or obscene language. Swear words are defined by the Webster Dictionary as "a profane or obscene oath or word" (Webster Dictionary Online). As for swearing, in the sense of offensive language use, as a whole: "The main purpose of swearing is to express emotions, especially anger and frustration." (Jay and Janschewitz 2008: 267) It has to be noted that swearing does not necessarily require the usage of unambiguous swear words, i.e. *You bloody cow!* would be considered an insult. However, *bloody* and *cow* as such would not be considered swear words without a particular context (cf. Thelwall 2008: 85). Anger and frustration are not the only reasons to use swear words. Swearing can also be used to add emphasis on what is said by using vulgar and offensive language (cf. Ljung 2011: viii). As it will be discussed later, swear words are also often uttered as a response to pain (cf. Stephens et al. 2009). Furthermore, as swear words are taboo words but taboo words are only swear words in particular contexts, the taboo words used for swearing are used in a non-literal meaning. Swearing often is subject of lexical, phrasal and syntactic constraints. (cf Ljung 2011: 4). For this present study, this has the advantage that these formulaic utterances can be researched in a corpus through electronic means, an advantage that other aspects, like euphemisms, do not have. More detail on this can be found in chapter 2.2.3.. Swear words also have to be vulgar, as words that are not vulgar, but taboo cannot be used for swearing. *Fuck* and *shag* are both English taboo words that refer to having sex as their literal meaning. However, it is only possible to swear with using *Fuck you!* but not with **Shag you!* (cf. Ljung 2011: 7). Which taboo words can be used for swearing and which cannot is not predictable (cf. ibid.: 8). Lastly, in a similar vein to what Jay and Janschewitz (cf. 2008: 267) said, swearing is emotive language and its main purpose is to express or reflect the speaker´s feelings, emotions and attitude (cf. Ljung 2011: 4).

It was mentioned above that swear words originate from taboo words or swearing uses words that are taboo in certain contexts. Thus, taboo seems to be a necessity for swearing. There are, however, insults that do not mention taboo words directly. *Your mother!* and *Your sister!* do not contain any taboo words as such, the taboo is rather implied or abbreviated (cf. ibid.: 5). Insults as such are claimed to be rare in English but common in "Romance and Slavic languages as well as in Arabic, Cantonese, Greek, Hindi, Mandarin, Turkish and others" (ibid.: 5). Through immigration, these insults have made in into languages where they were previously unknown. It would be interesting to see whether this form of insult has made its way into the English language, especially, in online forums such as reddit, on which people from all cultural backgrounds can meet. As with euphemisms, there might be difficulties in gathering reliable data[1].

As Ljung points out, although swearing has been part of language as long as language exists, linguistic research into the subject started only in the 1960s. In the following decades interest in the subject arose and studies have been done in different linguistic fields, such as psycholinguistics, sociolinguistics and historical view points (cf. ibid.: 3-4).

Swear words can originate from different semantic fields. Nowadays, swear words often refer to taboo topics such as "religion; sex acts; sexuality; genitals and

[1] More on this in 'Methodology'

sexual attributes; excretion; race, ethnic group or nationality; political affiliation; any other denigrated or oppressed group; stupidity; undesirable behavior [and] disease" (Thelwall 2008: 85). It has to be added that, although there are many subjects of taboo, the main resource for taboo words in English are sexual references and blasphemous and profane utterances (cf. Jay 2009: 154). Which category is most prevalent in the distribution of swear words is subject to change. A more in-depth discussion on taboo follows in the next chapter.

2.2. Taboo and censorship

2.2.1. Taboo - the origin of swear words

First, it should be established why the topics mentioned above are taboo and lead to offensive language. "Taboos arise out of social constraints on the individual's behavior where it can cause discomfort, harm or injury." (Allan & Burridge 2006: 1). This is rooted in the metaphysical, physical and moral risks that accompany these topics:

> "People are at metaphysical risk when dealing with sacred persons, objects and places; they are at physical risk from powerful earthly persons, dangerous creatures and disease. A person's soul or bodily effluvia may put him/her at metaphysical, moral or physical risk, and may contaminate others; a social act may breach constraints on polite behavior. Infractions of taboos can lead to illness or death, as well as to the lesser penalties of corporal punishment, incarceration, social ostracism or mere disapproval." (ibid.: 1)

These topics change over time and they are different from culture to culture. However, as mentioned above, it generally involves topics which might lead to potential physical, metaphysical or moral harm. This is grounded in actions and real life occurrences that are connected to these topics. Some topics that are taboo as actions do not have correspondent taboo words. Cannibalism would be such an example (cf. Fairman 2007: 1723). Similarly, there are taboo words that are purely linguistic. These are, however, mostly grounded in intercultural homophones, such as "Thai words *fâg* (sheath), *fág* (to hatch), and *phríg* (chili pepper)" (ibid.:1723). There are other taboos that differ from culture to culture. The topics might be the same, the actual taboo is different. Food taboos that are based in religious beliefs can differ quite a lot from one another. The special food preparation and segmentation in Judaism and the prohibition of the consumption of pork in Islam (cf. Allan and Burridge: 2006: 4) signify that case. The topics from which taboo words emerge do also have words that are not taboo. There are many medical terms to sex and bodily functions that are, especially in the context of medicine, not foul.

The level of offensiveness of a taboo word can also be determined, although it is extremely difficult. A word or utterance is not just taboo or not taboo. There are certain levels of offensiveness of taboo words. These levels can change over time and across different cultures. *Bloody* used to be an offensive word but nowadays it is fairly low in terms of offensiveness. It is argued that this was caused by certain circumstances. *Bloody* became taboo and a profanity during the Victorian era. During the first World War soldiers were exposed to such extreme situations that their slang became even more offensive. Thus, *bloody* became less offensive as stronger words replaced its use (cf. Jennings 1967: 94ff). On the other hand we have words like *fuck* which have always been very taboo. Sex, the category *fuck* can be attributed to, is still a taboo topic today, especially in the media of English speaking countries. Thus, and for other reasons[2], *fuck* remains a word which is considered taboo (cf. Fairman 2007: 1771). To determine how inappropriate and taboo a word is, Pinker says: "People treat an unpleasant word as taboo to the extent that everybody else treats it as taboo" (Pinker 2007: 357). The level of offensiveness is, thus, determined by pragmatic variables, including speaker-listener relationship, social-physical setting, choice of words and intonation and articulation - in settings where the words are spoken (cf. Jay 2009: 154). Taboo, therefore, is a social construct. People and communities agree what is ok to be said and done and what is not. This is not a distinction of black and white, it is rather a scale. An example of this is given in Table 1 - which can be found in chapter 2.2.2. on page 13.

[2] For an extensive look at and discussion of the history of *fuck* see Fairman 2007.

But what happens if we break such a taboo?

The most drastic repercussion for breaking a taboo is death. This happens in extreme cases and is only applicable if the argument that what is forbidden is therefore taboo (cf. Allan and Burridge 2006.: 5) is considered true. In that case, breaking the law and receiving capital punishment for such an offense would be an instance of death as a result of breaking a taboo. Examples for this would be the stoning of a woman who commits adultery under Sharia law. Similarly, the burning of witches in Europe because of 'You shall not permit a sorceress to live' (Exodus 22: 18) is a case of capital punishment for violating a taboo (cf. ibid.: 5).

There are, of course, breaches of taboos which cause lesser backlash than death. One of them would be participating in activities that are considered to be sins in certain religions, especially, when committing a sin is not connected to committing a crime. The depiction of god is considered to be a sin in Christianity. This does not have repercussions that are enforced by law or the state. It is rather the violation of a taboo within the realms of a belief system. It is assumed that this violation has some sort of repercussion that is or will be enforced by God. It could also lead to repercussions by fellow Christians denoting the offender as a sinner. This would cause a -temporary- loss of social status within the group. In this case, Christians have the opportunity to redeem themselves by confessing their sins to a priest and potentially doing some sort of redemption work. Language has also been subject to taboo for a long time. Blasphemy is by definition the breach of the taboo of insulting a god.

These examples show that there is generally some sort of repercussion for the violation of taboo. The general assumption there being that behavior which breaches taboo can be avoided (cf. ibid.: 6). It is also important to note that taboos are always specific for a certain community, a specific context in a certain place or time (cf. ibid.: 11). So, for example, the consuming of pork is considered a taboo in Islam but does not have any repercussions in Western culture. Whereas not wearing a hijab or headscarf as a woman is a taboo in Islam, the same piece of clothing causes debates in Western countries, in which Muslims are a minority.

The sanctions that are placed on taboos are manifold. The sanctions for breaking language taboos are of social nature. These social sanctions are placed upon behavior that is deemed "distasteful or impolite within a ceretain social context" (Allan and Burridge 2006: 237).

2.2.2. Censorship of taboo language

As previously mentioned, certain language is considered to be taboo and has been throughout history. As it was, has been or still is considered taboo, the respective words and phrases have been censored. Censorship represents the human intellect in that it is a form of conscious control over what we say and how we say it. It is associated with the human brain that is responsible for "emotional control, rational thought and tolerance" (Allan and Burridge 2006: 249). The following chapter focuses on the censorship of language.

Modern day censorship can best be shown by the infamous 'bleep' sound that is laid over the audio in US and UK media whenever certain swear words are used. Even on Youtube, where there are no censorship rules, official videos of TV-shows are still censored. When Jan Böhmermann attended the Late Night with Seth Meyers, he explained that he wanted to swear on American television just to be bleeped. The video shows him saying *fuck* quite clearly and was bleeped (2017, Youtube[3]). But measures against the use of 'foul' and taboo language are not an invention of modern media. The scope of what was censored and what was not changed through time and still differs from country to country, language to language and culture to culture.

Censorship already existed in the Roman Empire, as regulators could decide who was allowed to say what (cf. Holquist 1994: 14). The idea of taking influence on what was allowed to be said publicly has already existed for a long time. In England, laws against swearing were tried to be introduced as early as the fourteenth century with punishments including fines and being branded and placed in the stocks (cf. Montagu 1967: 108-111). It cannot be said, however, that the censorship of bad language has become less strict over time. It is rather influenced by political and cultural changes. McEnery argues that the utterance 'A turd for this argument' made by James I in the early 17th century would have had a much greater outrage in the public when Elizabeth II would have said the same

[3] Link to the to the exact moment Böhmermann swears: https://youtu.be/97XSh2t5weI?t=2m

thing during the 20th or 21st century. (cf. McEnery 2006: 52-53). Elizabeth I, with the increasing reach of printed works, established a structured censorship. The motivation back then was to influence the public through propaganda. Swear words and taboo language was not the main focus of censorship. It was targeted at topics and opinions that were deemed to be dangerous for the ruling class. However, offensive language acted as an accusation under which certain pieces of work were censored. Prosecution for publishing offensive works was allowed but rarely carried out. The Court responsible for such actions was headed by the church. A shift of focus towards censorship as a means to suppress 'bad' language happened after the Restoration stage in the second half of the 17th century. Certain parts of society did not agree with the usage of profane language in public showings. Following these emerging attitudes towards 'bad' language, the middle class started the reformation of manners. This included a desired absence of offensive language within polite conversation (cf. ibid. 69-70). The movement influenced the general attitude towards 'bad' language within the English society.

As an attempt to form their own identity, the middle class in England formed religious societies which were tasked with observing and regulating manners. Manners and moral were given a high status in order to distinguish the middle class from lower as well as higher classes. The latter thought to behave more purely than the respective lower classes. From the late 17th to the early 19th century these religious groups shifted the public opinion on 'foul' language. 'Foul' language was deemed to be a sign of lower class and, thus, something to be avoided (cf. ibid.: 71-72). The already existing religious laws that prohibited the use of offensive language in public settings were ineffective, as they were not enforced by law enforcement[4]. However, following the moral panic in the late 17th century, these laws became relevant again. To convict someone of the usage of 'bad' language only one witness was necessary. Furthermore, a financial reward for such a witness was promised (cf. ibid.: 79, 86). The background of this is that these religious groups connected the use of 'bad' language to the doings of the devil. In contrast, being morally pure and not using 'bad' language was a sign of doing good. So, the sentiment to fight 'foul' language was caused by religious interpretations. Some schoolbooks in the 18th century focused on whole sections

[4] i.e. Breaking these laws did not have any lawful repercussion.

of swearing and considered such doing as an offense to god (cf. ibid. 82)[5]It has to be noted here that the prosecution, as well as the responsible laws for such prosecutions, were created and enforced by religious groups. The first official law against swearing that was issued by the government was passed in 1745 (cf. Davis 1989: 8).

From 1708 to 1724 a total of 2851 prosecutions for swearing and cursing were carried out (cf. McEnery 2006: 91). McEnery goes further and assumes that prosecutions for swearing and cursing before that period were higher, as the number of prosecutions generally declined over time (cf. ibid.: 92). It can be assumed that these prosecutions, brought to court by a religious group seeing themselves as a moral observer, influenced the behavior of the English public.

It was during these times that modern swear words acquired their status as prominent swear words. *Damn* used to be the most prominent swear word, but it was slowly replaced by words like *fuck* or *piss* (cf. Porter 1991: 303-307). Additionally, synonyms for *fuck* started to decline in the sixteenth century (cf. Fairman 2007: 1718-1719). During the discussed period, 'bad' language became more defined as to what it is and what was considered to be offensive and punishable. 'Bad' language worked as a marker of distinction and, thus, a marker of social class (cf. McEnery 2006: 98). It is because of this development that the usage of swear words in public settings became very rare in the 19th century. Parallel to that, the religiously motivated limitation and censorship of 'bad' language became a more morally motivated stance against foul language. The period also shows how the attitude towards certain words changes. Words that were previously added and listed in dictionaries during the 16th, 17th and 18th century like *fuck* or *cunt* were excluded in the 19th century. This is in line with the addition of expressions of bodily functions and body parts to objectionable expressions (cf. Davis 1989: 8).

With the increasing reach and popularity of mass media in the 20th century, censorship in the UK focused even more on 'bad' language. Topics like sex and violence were subject to censorship and regulations as well (cf. McEnery 2006: 128). The public attitude towards 'bad' language, that was pro censorship and

[5] See also Allestree 1719

rooted in the preceding centuries, was carried over. With the emergence of television, the British Broadcasting Corporation (BBC) was founded. The BBC was a self-censoring entity, similar to what printers were doing around 400 years earlier (cf. ibid.: 102). With television, there was a new type of media in which censorship had to be done. This required a way to deal with offensive content. It was no longer written text that needed to be censored but sound and vision.

Meanwhile in the US, similar laws against swearing were in place. In Michigan it was prohibited to swear within earshot of women and children (cf. Fairman 2007: 1713). Sentences against such offenses included monetary fines and community services.

Following the premiere of Shaw's Pygmalion in 1914, the curse word *bloody* reentered public showings (cf. McEnery 2006: 103). It marks the reemergence of curse words in the English media. Beforehand, 'bad' language was almost completely banned from public speeches. The fact that 'foul' language was almost completely absent from public media had consequences. New shows, plays or programs that used 'bad' language gained popularity. During the Second World War, a radio show called 'Worker's Challenge' was broadcast from Germany to Britain. The aim of the show was to mobilize the British middle class against authorities. In order to do so, the presenters used British middle class slang that included bad language. As 'bad' language was associated with working class, the novelty of hearing that slang on the radio increased the number of listeners. Even the media reported on the usage of swear words in other media outlets respectively (cf. ibid.: 103-104).

Within society, the general morally absolutist opinion of things being either wrong or right changed to a more progressive and relative view. Legally, this was symbolized by a few new laws that allowed theaters to be uncensored or homosexuality to be partially decriminalized, among others (cf. ibid.: 106). In mass media this could be seen by programs including more and more 'bad' language. Especially, the BBC, which used to portray the middle class as it was thought to act, switched to a program that was aimed to a broader audience. With it, the BBC included slang and language that the lower classes were using -

including 'bad' language (cf. ibid.: 107-109)[6]. Although, this process did not happen without protest by people who still held views similar to those of the 17th to 19th century. In general, though, the view on 'bad' language changed insofar, as it was now allowed to happen in public places and in public media. However, offensive language is still censored today. The fundamental view that certain words are bad and others are not still prevails. Thus, the attitude towards offensive language that was introduced by religious groups in the 16th and 17th century is still detectable today. Furthermore, the idea to censor swear words is also prevalent in both the UK and in the US.

Nowadays, only a handful of words are still being censored. In America the Federal Communications Commission, FCC, is in charge to decide what is and what is not allowed to be broadcast. The FCC has a list of seven 'dirty' words which are *shit, piss, fuck, cunt, cocksucker, motherfucker, and tits*. Additionally, the FCC can censor any language that it considers to be 'indecent' (cf. Kaye and Sapolsky 2009: 4). It has to be noted that the FCC does not publicly announce the particular words that it would like to be censored (FCC). In shows that are broadcast live or slightly staggered, the infamous bleep sound replaces the indecent language in question. The FCC regulates language on a complaint-based basis. It weighs and balances factors which influence whether certain material is obscene, indecent or profane. Indecent and profane material cannot be shown between 6 a.m. and 10 p.m., whereas material that is considered obscene cannot be shown at all. In short, obscene material consists of mainly sexual content that excites lustful thoughts, is explicit and non-artistic, according to the US Supreme Court. Several factors can influence whether or not these factors are met for any given material. (Ljung 2011: 9-10). This means that any potential censorship is conducted on a case by case basis.

In the UK the Office of Communication, Ofcom, is responsible for regulating TV. Ofcom sets themselves the task to protect people who watch TV or listen to the radio from harmful or offensive material (cf. Ocfom.com, Ofcom 2016: 2). This includes offensive language. In order to judge which words are deemed offensive by the public Ofcom created a list of words with different levels of offensiveness.

[6] For an in-depth discussion of particular shows that influenced this development, see McEnery 2006

This could lead them to see which words could be used at which time in TV shows. Before 9 p.m., content that is unsuitable for children is restricted (cf. Ofcom 2016: 1). The list goes from milder words, medium words to strong and strongest words. What the public thinks can be broadcast and what cannot is highly dependent on contexts in which the words appear (cf. ibid.: 2). This includes the time of the broadcast as well as the contexts given within the program. So, in contrast to how the US deals with 'bad' language, in the UK the use of swear words nowadays is allowed, depending on contexts (cf. ibid.: 51). Uncut films, that contain harmful or offensive material, cannot be shown before 8 p.m. in the UK (cf. Bignell and Orlebar 2005: 11). To clarify, Ofcom censors programs before they are broadcast or sanctions broadcasters if they show offensive material on live TV or radio. The list of words looks like this:

Table 1 List of offensive words according to Ofcom

Milder words (generally of little concern)	**Medium words** (potentially unacceptable pre-watershed but acceptable post-watershed)	**Strong words** (generally unacceptable pre-watershed but mostly acceptable post-watershed)	**Strongest words** (highly unacceptable pre-watershed but generally acceptable post-watershed)
Arse	Arsehole	Bastard	Cunt
Bloody	Balls	Beaver	Fuck
Bugger	Bint	Beef curtains	Motherfucker
Cow	Bitch	Bellend	
Crap	Bollocks	Bloodclaat	
Damn	Bullshit	Clunge	
Ginger	Feck	Cock	
Git	Munter	Dick	
God	Pissed/Pissed off	Dickhead	
Goddam	Shit	Fanny	
Jesus Christ	Son of a bitch	Flaps	
Minger	Tits	Gash	
Sod-off		Knob	
		Minge	
		Prick	
		Punani	
		Pussy	
		Snatch	
		Twat	

(Ofcom 2016: 44)

Interestingly, the seven 'dirty' words that are bleeped in the US are all present in Table 1. They are not, however, considered to be the strongest words by the British public. *Tits* and *shit* are among the seven 'dirty' words but only in the Medium words category in the Ofcom research. The strongest words are all among the seven 'dirty' words in the US.

According to Ofcom, the general public maintains the belief that there should be rules and regulations on offensive language on radio and TV (cf. Ofcom 2016: 2). What this section on censorship shows is that within the UK and the US there are attitudes against offensive language on TV and radio. Offensive language is seen as something potentially harmful to children (cf. ibid.: 2) or harmful in general. This view is rooted in a claim to moral superiority that was established by religious groups in the 16th and 17th century in Britain. Back then, labeled as blasphemy, swearing and cursing was seen as an affront to god. Traces of this can still be found in modern society as the rules and regulations enforced by the FCC and Ofcom show. This is another aspect that proves the special stance that offensive language and thus swear words have in language. By contrast, in Germany the use of offensive language on TV and radio is generally allowed. That is not to say that there is no censorship, but the focus on offensive language being regulated by government entities is special to the English speaking communities of the UK and the US. It has to be said that the censorship of 'bad' language does not mean that it is not used in broadcasts. Research shows that nine out of ten programs in broadcast or cable TV in the US contain offensive language with an indecent word being spoken about every five minutes (cf. Kaye and Sapolsky 2009: 11).

One other form of censorship is self-censorship. People who know of the impact that offensive language causes can actively try to avoid such language. This is generally hard to detect, as data on omission of such language can hardly be gathered (cf. Santaemilla 2008: 244). There are however, strategies with which offensive language can be avoided. This is what the next chapter focuses on.

2.2.3. Euphemisms

To censor oneself can have many different forms, from alternating sentence structure to the omission of certain topics or just keeping quiet. Those forms are hard to detect, however. This chapter deals with the euphemisms. These are expressions that are used instead of another dispreferred expression (cf. Allan and Burridge 2006: 238) that would threaten the face of the addressee or a third party (cf. Allan and Burridge: 1991: 11). The preferred expressions are semantically similar or identical but, for whatever reason, not considered to be taboo. This is at least the case in instances of using euphemisms instead of the established taboo swear words. Another way to define euphemisms is to describe them as "the semantic or formal process by which the taboo is stripped of its most explicit or obscene overtones" (Fernández 2008: 96). The goal of using euphemisms is to avoid mentioning taboo as the use of taboo has the inherent danger of some sort of repercussion.

The use of euphemisms can be greatly motivated by the want to keep the previously mentioned 'face'. Research has shown that there are two distinctive motives for euphemism use that are connected to the notion of face. First, speakers could use euphemisms because they do not want to embarrass or discomfort the addressee, as the broad mentioning of distasteful topics threatens the positive face of the addressee. Second, euphemisms can be employed for self-presentational reasons. Avoiding taboo by using euphemisms can make the speaker appear more considerate and sympathetic, thereby saving the speaker's positive face. Research has shown that the latter motive seems to be more compelling than the former. Interestingly, if the speaker remains anonymous or does not meet the addressee, the use of euphemisms decreases (cf. McGlone and Batchelor 2003: 260). In the frame of this study, these findings would suggest that in the online discourse, in which the communicators remain relatively anonymous, the use of euphemisms should be comparatively low compared to the use of taboo terms and swear words. This assumption is supported by the study mentioned above. As the findings by McGlone and Batchelor come from an experiment in which the form of communication between the participants were e-Mails, the researchers come to the following conclusion: Politeness strategies, to which the use of euphemism belongs, are only employed if the danger of an FTA is apparent. If, like in online chatrooms or forums, the communicators are anonymous, the submitted messages might not have self-presentational currency or face value. Thus, communicators might not feel obliged or motivated to employ politeness strategies (cf. ibid: 261-262). Also, as already mentioned, taboo words and topics are not necessarily identical to swear words. The use of swear words in online discourse could, therefore, reveal different characteristics. Furthermore, reddit does have a system in which a positive self-presentation is rewarded and it does appear that "the desire to make a positive impression on the external audience [...] appeared to regulate euphemism use." (ibid.: 262). The reward system with which comments are graded and rewarded on reddit is discussed in chapter 3.3. later in this paper.

One aspect that makes detecting euphemisms in written work difficult is that they are unpredictable and can be made up a priori (cf. Domínguez 2005: 15). Even though, there are euphemisms for swear words such as *fudge* for *fuck*, they can be made up on the spot. For that to work these new euphemisms are often part of the same conceptual network. The topic of death is quite often referred to as a form of travelling, i.e. *to die* can be expressed as *to pass away*. (cf. ibid.: 12). It has to be noted that there are other forms from which euphemisms can be produced. Phonological vicinity can also play a role (cf. Bowers and Pleydell-Pearce 2011). So, they are not completely random. However, euphemisms that work in one language might not work in another (cf. Domínguez 2005: 15). For example: *(to) fudge* translates to German in a range from 'Fälschung' to 'Schmelzbonbon' or 'pfuschen'. The German translations cannot be used as euphemisms for *ficken* which is the German equivalent of *fuck.* Also, euphemisms tend to become taboo words when they are lexicalized (cf. ibid.: 11, cf. Fernández 2008: 100-102) These last findings have been contradicted, however, and are up for further research (cf. McGlone et al. 2006: 276). Furthermore, euphemisms only work as long as their interpretation is ambiguous (cf. Domínguez 2005: 10). The topic of detecting euphemisms in the chosen corpus of reddit comments that build the data set of this study will be discussed further chapter 4.3..

To sum up, euphemisms are used to avoid talking about taboo topics with taboo words. They can be used for a variety of reasons, most prominently to avoid acting out FTAs. In the realm of swear words, research has shown that euphemisms cause a much more mellow reaction than the explicit swear word (cf. Bowers and Pleydell-Pearce 2011).

2.3. Politeness and impoliteness

2.3.1. On politeness

Before focussing on swear words as such and why they have a special stance in language, further theoretical background has to be explained. Swear words are most often taboo language. Thus, speakers generally try to avoid using such words. The following segment should explain why speakers have a tendency to avoid offensive language and taboo words.

One of the most prominent additions to discourse analysis is the concept of politeness. The concept of politeness in linguistics was introduced most noticeably by Goffman (1967), Brown and Levinson (1978, 1987), Leech (1983) and has been further discussed by Lakoff (1989) and many others. "Politeness can be defined as a means of minimizing confrontation in discourse [...]" (Lakoff 1989: 102). For this study, the politeness theory proposed by Brown and Levinson (1978, 1987) is most important. The theory by Brown and Levinson involves the notion of face. They differentiate between positive and negative face (cf. Brown and Levinson 2006: 311). The positive face want is defined as "the want of every member that his wants be desirable to at least some others." And the negative face want is defined as "the want of every 'competent adult member' that his actions be unimpeded by others". (ibid.: 312) As a general rule, speakers do not want these faces to be threatened (cf. ibid.: 311). In other words, speakers are inclined to avoid so called face-threatening acts or FTAs. Brown and Levinson consider this need to avoid FTAs politeness.

There is a number of different FTAs that are defined by Brown and Levinson. Only those that are relevant for this study will be presented and discussed. The first distinction between different kinds of FTAs they make is the distinction between acts that threaten the positive face and acts that threaten the negative face (cf. ibid.: 313).

The first kind are FTAs that threaten the addressee's negative face by indicating that the speaker does not want to restrict the addressee's freedom of action (cf. ibid.: 313). This includes, but is not limited to, acts that indicate some kind of desire towards the listener or their goods. Thus, the listener feels inclined to protect themselves or their goods from the speaker. These FTAs include "expressions of strong (negative) emotions towards [the addressee] - e.g. hatred, anger, lust" (ibid.: 314). Therefore, insults are a threat to the addressee's negative face and should, according to the theory of politeness, be avoided. A factor that is already touched upon in that quote are the "strong emotions", which will seem to be a strong motivator for insults and swear words. This will be discussed in chapter 2.4..

On the other hand there are insults which also threaten the positive face by implying that the speaker does not care about the addressee's feelings and wants. Insults show that the speaker has some kind of negative evaluation of the addressee's positive face wants by attacking his characteristics, values, feelings etc. (cf. ibid.: 314). So, swear words and insults can threaten both the negative face and the positive face of the addressee.

Thus far, the FTAs mentioned above primarily threaten the addressee's face. However, the speakers can threaten their own face, too. "Acting stupid" (ibid.: 315) threatens the speaker´s positive face (cf. ibid.: 315). So, a misplaced insult or a misplaced usage of swear words could potentially hurt the speaker. Thus, the speaker should be inclined to only refer to swear words if the situation justifies their usage, i.e. if the violation of the face of the addressee seems justified. However, in online discourse the anonymity could potentially lead to a lesser desire to act politely. That is, an increase in swear words could be expected if members of a conversation feel that their nicknames do not reflect themselves and, thus, they do not have to care about face and politeness as much. But not only insults are impolite. The usage of taboo words which are not meant to be offensive can still be regarded as impolite (cf. Jay 2009: 155). For this study, I regard the politeness theory proposed by Brown and Levinson to be true to at least some extent. Also, other studies have found that verbal actions which would be considered FTAs in the English language under the principles of Brown and Levinson are not FTAs in other cultures, e.g. Persian (cf. Koutlaki 2009).

Still, the usage of offensive language and taboo words contains the danger to commit an FTA. Thus, it should be expected that such words are only used rarely. However, in online discourse and on reddit, anonymity puts a barrier between the face of the speaker and their audience. The person who submits a comment does not have to face repercussions in the real world. Could this lead to a higher usage of swearing in online discourse? Could it lead to a usage of more offensive swear words, regardless of context?

2.3.2. On Impoliteness

Although it is assumed that participants of a conversation always seek to be polite, there are instances in which impolite behavior is consciously chosen. Mock impoliteness are exchanges that are seemingly impolite but are not intended to offend. In other words, mock politeness can be seen as banter (cf. Culpeper 1996: 352). Leech goes further and connects intimacy with banter. Utterances that seem to be offensive but are not treated as such by either participant indicate social intimacy between the participants. This includes a relatively equal level of authority and social closeness (cf. Leech 1983: 144). Culpeper adds that this only works in circumstances in which the impoliteness is understood to be untrue (cf. Culpeper 1996: 352). To sum up: "insults are more likely to be interpreted as banter when directed at targets liked by the speaker." (ibid.:353)

Whereas the banter mentioned above can be observed in conversations with a small number of individuals that know and like each other, impolite utterances can also be observed in larger social groups. In these circumstances the banter is more ritualized and called sounding. Labov (1972) revealed the structure of this speech event which used to happen primarily between young black adolescents in America. The same principle as in banter can be observed: the participating group has a shared knowledge for the insults to be untrue. The purpose of the sounding is to strengthen group solidarity. In comparison to banter sounding is more ritualized and follows certain rules in which improvisation does not occur often. It is more important to know many variations of existing insults than to come up with new ones. (cf. Culpeper 1996: 353). The ritualization of insults leads to a loss of responsibility of the individual for the acts they committed (cf. ibid.: 353, Labov 1972: 352-353) - referring to the possible threat of face and act of impoliteness. Both, banter and sounding are instances of impolite words and phrases, but the acts themselves are ultimately considered to be of polite nature.

There are contexts in which impoliteness is actually intentionally used and not an instance of failed politeness. These are distinct from the bald on record strategies proposed by Brown and Levinson (1978, 1987). They propose a context in which both participants recognize that the face wants are suspended due to an emergency, the face threat is minimal and/or the power relations between speaker and hearer are very one-sided in favor of the speaker (cf. Brown and Levinson 2006: 316ff). These specific contexts allow verbal acts of lesser politeness. However, Brown and Levinson's contexts and connected strategies[7] are deployed within polite conversation. The want to maintain face is still present and only omitted for a few special contexts.

The following strategies that are deployed when the speaker does not want to maintain the hearer's face are proposed by Culpeper et al. (2002: 1554 - 1555):

> "1. Bald on record impoliteness - [...], bald on record impoliteness is typically deployed where there is much face at stake, and where there is an intention on the part of the speaker to attack the face of the hearer.
>
> 2. Positive impoliteness. - The use of strategies designed to damage the addressee’s positive face wants (‘ignore, snub the other’, ‘exclude the other from the activity’, ‘disassociate from the other’, ‘be disinterested, unconcerned, unsympathetic’, ‘use inappropriate identity markers’, ‘use obscure or secretive language’, ‘seek disagreement’, ‘make the other feel uncomfortable (e.g. do not avoid silence, joke, or use small talk)’, ‘use taboo words’, ‘call the other names’, etc.).
>
> 3. Negative impoliteness. - The use of strategies designed to damage the addressee’s negative face wants (‘frighten’, ‘condescend, scorn, or ridicule’, ‘invade the other’s space’, ‘explicitly associate the other with a negative aspect’, ‘put the other’s indebtedness on record’, ‘hinder or block the other—physically or linguistically’, etc.).

[7] For an in-depth discussion of strategies for bald on record behavior see Brown and Levinson (1987: 61ff).

> 4. Sarcasm or mock politeness. - The use of politeness strategies that are obviously insincere, and thus remain surface realizations. Sarcasm (mock politeness for social disharmony) is clearly the opposite of banter (mock impoliteness for social harmony).
>
> 5. Withhold politeness. - Keep silent or fail to act where politeness work is expected."

Impoliteness strategies refer to circumstances in which the threatening of face is intentional. These do not necessarily have to include swear words, although their use is a viable option (see Positive impoliteness). They are highly dependent on context, as are politeness strategies. They are not mutually exclusive. Additionally, they cannot be rated on a level of offensiveness (cf. ibid.: 1555). For this study, positive impoliteness is most important. The use of taboo language can be used in many different circumstances. It can clearly be used intentionally to attack the face of the addressee, as it can make the hearer feel uncomfortable and it can express anger targeted at the hearer (cf. ibid.: 1557). A shouted *Fuck you!* can not only indicate a highly emotional state, it can also be used to make the addressee feel responsible for the aggravated state of the speaker (cf. ibid.: 1573). It includes swear words and is directed towards the hearer. Culpeper goes further and includes prosody as a means to indicate intentional impoliteness. In the context of this study, however, prosody is irrelevant as the research is done with written text. A combination of different strategies is also possible. The phrase *What the fuck are you doing?* asks a challenging question, which belongs to the category of negative impoliteness, as well as including a swear word, which is positive impoliteness. Additionally, it could be observed that in verbal conversations the repetition of such phrases forms a sort of parallelism which increases the level of impoliteness even further (cf. ibid.: 1561). So, swear words can be part of larger structures which contain several FTAs at once. They are part of impolite acts and can be used to increase impoliteness. However, a challengingquestion is impolite whether or not a swear word is included. Thus, swear words are not necessary to complete impolite acts or FTAs. And, given that the impoliteness is intentional, repetition boosts impoliteness. If a swear word is used in a situation to threaten someone's face, it could be expected that several swear words are used in order to enhance the effect.

What might be even more interesting than the use of impoliteness, is the reaction to it. What do people do when they are confronted with an intentional act of impolite behavior?

There are two different strategies to deal with an intentional face threat. The recipient of an intentional FTA has the option to accept the FTA or counter it. The option to counter an FTA can also be divided into two. On the one hand we have the OFFENSIVE-DEFENSIVE option and on the other the OFFENSIVE-OFFENSIVE. The first one refers to instances in which a personal insult is followed by denial from the recipient. The second one refers to instances in which the recipient of the first offense answers back with an offense by himself or herself (cf. ibid.: 1562). A third option, staying silent, is also imaginable. However, Culpeper could not find such instances in the BNC. However, as my study refers to written text conversations that are open to practically everyone, this could occur much more often. Especially, since an immediate response is not expected as it would be in spoken conversation. On the other hand, comment strings are time sensitive. The sooner one answers, the better the chances are that the comment is seen. This will be further discussed in chapter 3.3..

The OFFENSIVE-OFFENSIVE strategy can be shown in this fictional example:

(1) Example A

S1: *Go to hell!*

S2: *Fuck you!*

S1: *Go fuck yourself!*

S2: *Suck my cock!*

The idea is that as a response to the first offense by S1, S2 reacts with another offense. In this case, this forms a spiral of insults including swear words of increasing level of offensiveness. So, the utterance of an intentional FTA causes the utterance of another intentional FTA by the recipient. This could escalate and cause more insults by both parties. Example B shows how the OFFENSIVE-DEFENSIVE option could look like:

(2) Example B

S1: *You are so dumb.*

S2: *No I am not.*

S3: *Yes you are.*

S2: *It wasn't my fault.*

In (2) the recipient of the first offense does not respond with another offense. S2 rather tries to just deny the accusation made by S1. Other strategies are also observable such as denying responsibility for the actions that caused S1 to utter an insult in the first place (cf. ibid.: 1565) like S2 does in his second response.

It is true that the use of swear words can increase the offensiveness of an utterance. It can be used to be impolite. It has to be noted, though, that swear words are not synonymous with insults. For example, the use of a swear word or offensive language in a joke can also be intentional impoliteness, not to the effect of an insult but to add comic value or shock to a joke. The distinction between the two is important when analyzing swear words and taboo language.

2.4. On how and why we swear

2.4.1. How (often) do we swear?

In the beginning of this study, it was already briefly discussed that swearing includes taboo but only certain taboo words and in certain contexts. This section is devoted to explain this in greater detail. In contrast to Muslim cultures, Christian cultures' swearing includes not only higher religious powers but also those of lower celestial beings, i.e. the devil and hell (cf. Ljung 2011: 6). Swear words that do not refer to religious concepts and taboos include words that are vulgar or embarrassing and include the taboos of excrement, sexual intercourse or other sexual practices and organs (cf. ibid:7). This focus on vulgarity is an important aspect of swear words. Although other words that refer to taboo are available, swearing with utterances such as *excrement, copulate* or *penis* is impossible, whereas their vulgar counterparts *shit, fuck* and *cock/prick* are swear words. This means that only a limited subset of taboo words can be used for swearing (cf. ibid.: 7).

Despite the limitation of taboo words that could be used for swearing, the potential to use a variety of words is there. A variety of studies conducted by Jay (1992, 2000) and Jay and Janschewitz (2008) show that over 70 different swear words were publicly recorded. However, the same research shows that during the period between 1986 and 2006 a set of ten words accounted for 80% of recorded swear words. These ten phrases are *fuck, shit, hell, damn, goddamn, Jesus Christ, ass, oh my god, bitch*, and *sucks*. The two most popular of those are *shit* and *fuck*. In fact, these two account for a third to a half of all accounts for swear word usage in that period. The set of swear words that are most common is relatively stable. By comparison, swear words that are regarded as extremely offensive, such as *cunt, cocksucker* or *nigger* are rarely used in public settings (cf. Jay 2009: 156). This research includes data gathered from conversations that were done verbally. The amount of swear words used compared to the overall number of words in a conversation rises the less formal a conversation is (cf. Jay and Janschewitz 2008: 273). A difference in swearing between both genders can also be observed. The age of the speaker is also a factor. Research shows that men swear more frequently in public compared to women. The difference between the two does seem to decrease, as men accounted for 67% of public swearing in 1986 but only for 55% of swearing in 2006 (cf. Jay 2009: 156). Men are also more offensive than women by using more *fuck, shit, motherfucker* than women whereas women tend to use words like *oh my god, bitch, piss* and *retard(ed)* more often than men do. Interestingly, both genders swear more often and freely when they are in conversation with the same genders (cf. ibid.: 156, cf McEnery 2006: 28-31). Note here that *oh my god!* is considered a swear word that is not represented in the list provided by Ofcom (2016: 44) where it, thus, would not be censored or considered to be taboo or indecent. Females also tend to generally rate the level of offensiveness of any given taboo word higher than their male counterparts (cf. Fägersten 2007: 32).

Swear words are used across all ages with the teenage years being the period with the highest rate of swear words used (cf. ibid.: 156, Thelwall 2008) with the frequency of 'bad' language used dropping significantly after the age of 25. (cf. McEnery 2006: 39). The same correlation of age to frequency of swear words can also be observed when correlating age and the level of offensiveness of swear words used. After reaching its peak in a group of 16 to 25 year olds the level of offensiveness generally decreases the older people get (cf. ibid.: 40). The social class of the communicators is also an influence. The higher the class the lower the frequency of 'bad' language used (cf. ibid.: 42). The findings by McEnery are taken from data from the spoken sub-corpus of the BNC; the British National Corpus. Other corpora might cause different results.

The overall rate of swear words used per total word count varies very little across most studies. The overall rate McEnery was able to observe a rate of 0.3% to 0.5% of 'bad' language words used compared to the overall word count (cf. ibid.:.45-49). Jay (1980) found a rate of 0.7% of taboo words recorded in spoken conversation from a corpus of 11.609 words. Similar results were found by Mehl and Pennebaker (2003) who found a rate of 5 words in 1000 spoken words, i.e. 0.5%, were swear words. Their research, which involved recording spoken conversation in intervals which were not known to the participants, shows that participants displayed great variety in swear word usage. Almost half of the participants used no swear words at all. In contrast to that, another participant swore at a rate of 34 swear words per 1000 words, or 3.4%, spoken. The rate of using swear words remained consistent for each participant respectively. In fact, the rate of swear word usage showed the highest consistency across all other recorded categories (cf. Mehl and Pennebaker 2003: 862-863). So, personal preference can also greatly influence the amount of swear words used. The same study also provides the rate at which other word types are used. This should give some perspective as to how common swear words are in everyday spoken conversations. Prepositions account for 8.9%, articles for 3.9% and words that reflect positive emotions such as *good* or *happy* for 3.2%. The largest part is made up by verbs in present tense at 15.9%. The only category that has a lower rate of words used is 'non-fluencies' such as *uh* or *er* (cf. ibid.: 863)[8]. Still, swear words remain a common occurrence, especially, if the fact is considered that the other categories mentioned above include words that are usable in a wider variety of circumstances than swear words. Furthermore, the already discussed issue of impoliteness and taboo influences the usage of swear words heavily. They still are

[8] There were more word categories recorded and noted, for more detail see Mehl and Pennebaker 2003

an integral part of everyday speech. The research presented indicates that with a rate of 0.3% to 0.7% and a total amount of around 16.000 words spoken each day (Mehl et al. 2007) the average person utters between 48 and 112 swear words per day. All the research presented in this part collected data of participants using the English language in spoken form. As previously discussed, other cultures and languages handle taboo and swearing differently, thus research in those languages might differ. Research that included swear words in written conversation, namely on the internet, is presented in chapter 3.2..

2.4.2. Why do we swear?

As already discussed, swearing has the main goal of conveying emotion, especially when swear words are not used for their literal meaning (Jay 2009: 155). Research by Jay found that the main reason for people to use swear words is connected to anger and frustration, expressed at a personal or an interpersonal level (cf. Jay 1992, 2000). This does not necessarily mean that these recorded instances cause the listener to regard swearing as highly aggressive or rude. The swearing recorded by Jay (1992, 2000) never led to any form of violence. It was rather regarded as conversational (Jay and Janschewitz 2008). The anger and frustration that is expressed through swear words can, therefore, be aimed at third parties that are not present at that particular conversation. However, "Taboo words are a defining feature of sexual harassment, blasphemy, obscene phone calls, discrimination, hate speech and verbal abuse categories." (Jay 2009: 155). So, while not necessarily harmful and aggressive swearing is an essential part of several harmful speech acts.

On the other hand, there are positive outcomes from the usage of taboo words. These include uses in jokes and humor, slang that is used within a particular group, self-depreciation, irony, sarcasm, storytelling, sex talk and social commentary (cf. ibid.: 155). Following an extensive summary of previous researches and a study of their own, Fägersten (2007) concludes "that the most frequently occurring type of swearing is neither that which is typically represented in offensiveness studies nor that which is considered most offensive." (ibid.: 33). They deduce that from evaluating studies in which the level of offensiveness of a word was judged on a word list without any further context. Their study opposed the ratings of a simple word list and one in which the level of offensiveness was rated based on further contexts. The ratings were very different and generally rated swear words in contexts lower than without them (cf. ibid.: 23-31). They conclude that the paradox that previous studies show - the fact that swear words rank highly on offensiveness scales but are also often used - can be explained by their findings. The relatively high number of occurrences of swear words in everyday conversations can only be explained by the fact that they must be used in contexts which render these words less offensive (cf. ibid.: 33). Thus, the reason for swearing cannot only be to express anger and frustration. However, the use of swear words always includes the risk of being impolite or offensive, even if the speaker does not intend to do any harm by using such words, especially in contexts in which a swear word is used casually as in *The food is fucking awesome.* (cf. Jay 2009: 155). Contexts like these enhance the emotional impact of the utterance but are not intended to do harm, be impolite or be humorous, unless the context suggests otherwise, of course.

Jay and Janschewitz (2008) distinguish between two different forms of swearing. On the one hand, they define propositional swearing as swearing in which the speaker consciously swears and controls the content of their utterance. This includes utterances that are either intentionally rude and offensive as well as utterances that are intentionally not so. On the other hand, they define non-propositional swearing. This includes unintentional, unplanned and incontrollable utterances of swear words. These can be observed in instances of emotional responses, as in responses to pain or surprise among others. The other possibility is swearing as a result of brain damage. Non-propositional swearing can be regarded as offensive by the listeners but does not have to be. It is claimed to be neither polite nor impolite. Whatever offense might be caused by this form of swearing is unintentional (cf. Jay and Janschewitz 2008: 269-270).

As already mentioned, swearing can be expressed as a response to pain. In an experiment conducted by Stephens et al. (2009) participants were asked to submerge their hands in 5°C cold water for up to five minutes. Afterwards, they were asked to rate the pain they perceived. Additionally, they were allowed to utter a word during that experiment. Participants who chose to utter swear words perceived their pain to be significantly lower than those who did not utter swear words (cf. Stephens et al. 2009: 1056-1059). It seems, therefore, that the reason people swear when they experience pain is to relieve stress, i.e. decrease their pain perception. It was also observed that swearing increases the heart rate (cf. ibid.: 1060). Stephens et al. speculate that swearing can be used to intentionally cause aggression within the listener, as an attempt to motivate football players or soldiers (cf. ibid.: 1060). Although speculative, it seems that swearing can also be used to motivate others.

As previously mentioned, the usage of swear words, be it their number of occurrences or their appropriateness in certain circumstances, is dependent on culture and language. This can be seen, as multilinguals rate the level of offensiveness of taboo terms in a second language differently to what native speakers would. They also seem to use them less frequently. Their usage and judgement of swear words is dependent on their individual linguistic history. (cf. Dewaele 2004a, 2004b, Jay and Janschewitz 2008). This can lead to misunderstandings. Culpeper describes an example in which a Norwegian lives with native English speakers. The Norwegian would refer to the friends of the son as 'cunts' as that was their term to greet each other. The Norwegian was unaware of the taboo that surrounds that word, thus, creating discomfort for the parents (cf. Culpeper 2011: 116). This example shows quite nicely how the usage of swear words can sometimes not be intended to be offensive, yet the usage still causes offense.

To sum up, swear words can be used in a variety of circumstances. They can be used to be intentionally impolite, to express aggression and frustration, to relieve pain but also to tell jokes, as slang within a community, to express and enhance a strong emotion that does not have to be anger or just as a conversational means, the latter being able to build social relationships through showing solidarity (cf. Daly et al. 2004). Research has shown that it is most likely that the majority of swear words is used in conversation without intending to be rude. There is, however, always the danger that swear words are perceived as rude and offensive as they are, by definition, taboo.

3. Online discourse, public discussion and reddit

3.1. Online discourse and public discussion

This paper investigates the use of swear words in online discourse, more specifically, it investigates the use of swear words on reddit. The following section aims to outline the main points of online discourse, to show previous studies on the usage of swear words online and to point out what reddit actually is.

Online discourse can have many different forms. It is difficult to define online discourse as one specific entity, as the internet provides very different modes of conversation. Werry (1996) distinguishes between three different types of conversation while only focusing on written conversations (cf. Werry 1996: 48). Within this older definition of written discourse in the online environment, none of them describe the form of conversation found on reddit exactly. E-Mail language has been described to show attributes of both written and spoken language[9]. Nowadays, online communication can also be verbally, mimicking telephone calls, or via videos in social media. One to one chatrooms, like they are frequent on social media, do not allow others to inspect and read a discussion. This is critically different from the type of conversation that can be had on online forums. Reddit in this case even allows outsiders - people who do not have an account on the website, to read most of the content shared and expressed on reddit. The exact circumstances and attributes of online discourse are often dependent on the platform they are hosted on. Therefore, I will only closely describe the form of communication that reddit offers.

3.2. Swear words in online discourse

How do people swear online? In an environment that generally allows answers with more time to formulate them, unintentional swearing seems unlikely. Commentators have full control over their utterances, "given the written communicative mode and little significance of evoking stimuli" (Dynel 2012: 37). This means that stimuli which can cause involuntary swear word usage in everyday spoken language, such as surprise and pain, do not impact swear word usage in written discourse. Answers take longer to produce, as they have to be

[9] For a summary of previous studies on the subject see Morrow 2006: 534-535

written. On the other hand, anonymity provides the avoidance of any repercussions of breaking taboo. The latter could lead to an increase in swearing, whereas the former would negate the involuntary and unintentional use of swear word, as previously discussed in this paper.

Thelwall (2008) investigated the use of swear words on the formerly popular social media site MySpace. MySpace was oriented towards the youth. Each user had their own site, their 'MySpace' on which they could present themselves to other users. Thelwall found that swear words were used in 0.2% of British profiles and 0.3% of US profiles (cf. Thelwall 2008: 93). On MySpace, then, the rate of swearing compared to other words offered on those profiles is similar to the rate found in spoken English. Thelwall goes further to investigate gender and age differences (cf. ibid.: 94-97) and comes to similar conclusions as those discussed by McEnery (2006) and Jay (2009). A difference to the findings of McEnery is that personal insults and idiomatic use seem to be more common than in spoken English, although the overall usage of swear words is slightly lower. The usage of swear words in idiomatic contexts and as playful insults is relatively high as opposed to swearing for emphasis (cf. Thelwall 2008: 99). Thelwall further concludes that MySpace language shows signs that are otherwise strongly connected to spoken language, although communication on MySpace takes place via script alone. (cf. ibid.: 97). The overall rate has to be put into perspective, as non-standard spellings, which are frequent in online discourse[10], were excluded (cf. ibid.: 98-99). Lastly, he concludes that the use of swear words on MySpace rather reflects normal behavior among the younger generation than deviant intentions (cf. ibid: 100).

In a study on the swearing habits of British Twitter users, Gauthier et al. (2015) came to find similar results. They, as well as Thelwall and also McEnery (2006), focused on the difference in swearing between both genders. First, the findings were that the majority of tweets were published by people between 19 and 30 years with the second largest group being users between the ages of 12 and 18 years. The most common swear word used was *fuck* followed by *shit, hell* and *cunt* for men. The same words and order was found for females except for the

[10] As an example: abbreviations caused by lack of space (cf. Werry 1996: 53-61)

word *cunt* which is replaced by the term *bitch*. The overall distribution of swear words among tweets was 5.8% of tweets by men contained swear words and 4.8% of women's tweets contained swear words (cf. Gauthier et al. 2015: 5). This is in line with previous findings that men seem to swear more than women. As an additional influence on swearing Gauthier et al. found that the time of day also plays a big role (cf. ibid.: 6-7) further supporting the claim that swearing is very context dependent. Lastly, they note that when mentioning entities such as people, organizations or locations, twitter users swear mostly when referring to people (cf. ibid.: 7). Thus far, research has shown that contexts and pragmatic variables such as age, gender, time of day, social class and linguistic proficiency all influence swearing habits.

In a, by their own accord, "informal" (Dynel 2012: 27) study of swearing on Youtube, Dynel shows the heavy use of swear words on an online platform that seems to be motivated by anonymity (cf. ibid.: 35). In these messages, impoliteness seems to be intended while using swear words that are very offensive. Dery also sees anonymity as a reason to drop any potential restrain one might have to use swear words. He writes:

> "[...] the wraithlike nature of electronic communication - the flesh become word, the sender reincarnated as letters floating on a terminal screen - accelerates the escalation of hostilities when tempers flare; disembodied, sometimes pseudonymous combatants tend to feel that they can hurl insults with impunity (or at least without fear of bodily harm)"
>
> (Dery 1994: 1).

Although it seems contrary to the findings in the previous studies, MySpace and Twitter are social media platforms on which people tend to register with their own names more often than on message boards or forums. Thus, the level of anonymity that Youtube or other such platforms provide is much higher. Anonymity seems to make using swear words easier because of the lack of repercussions. However, both Dery and Dynel do not provide quantitative data to support that hypothesis. Lastly, Dynel links the use of swear words on the internet with humor (cf. Dynel 2012: 40-41) whilst excluding the use of swear words to build solidarity within a group from huge relevance on the internet (cf. ibid.: 40).

The latter is deducted from the facts that ritual abuse is difficult to distinguish from real abuse in written commentaries.

In short, swearing happens on the internet and, on social media sites, at roughly the same rate compared to spoken English. Whilst quantitative data on swearing behind a nickname is lacking at the moment, it is supposed that anonymity can increase the use of swear words. That usage also seems to be motivated by being humorous. The following section now focuses on the corpus of the present study. Its source, reddit, will be explained in detail to establish context and restraints that communication on that platform might be influenced by.

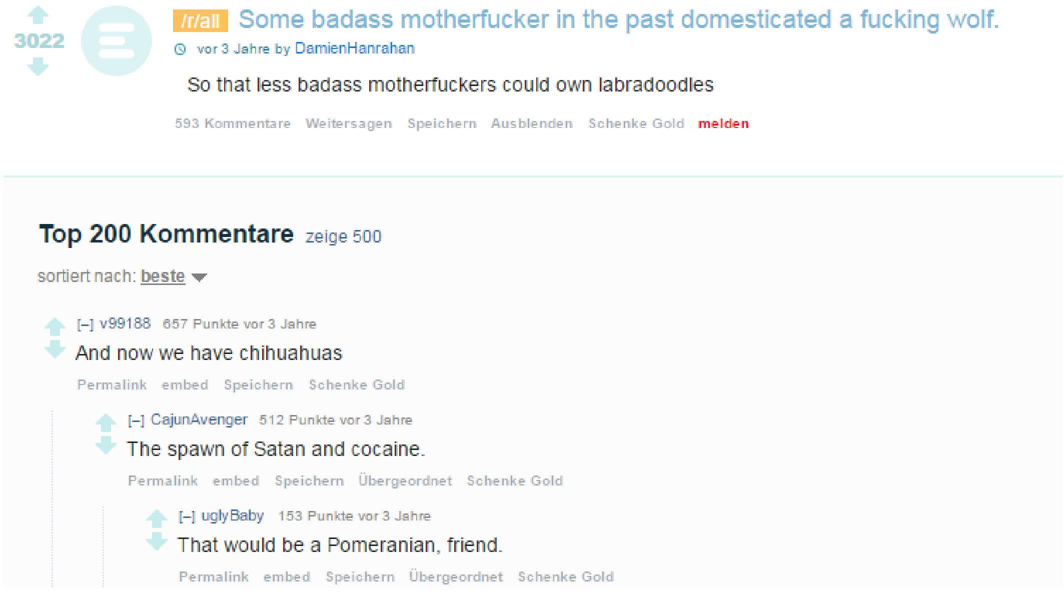
3.3. What is reddit?

Reddit is the self-proclaimed 'front page of the internet'. It was founded in 2005. It is currently the 7th most visited website on the internet worldwide. 55.4% of reddit's visitors come from the US, followed by users from the UK at 7.6%, Canada at 6.2% and Australia 3.2%. Reddit's user base is, therefore, largely based on countries where English is the native language. The largest portion of visitors that are not native English speakers come from Germany, which provides the fifth largest percentage of visitors at 2.5%. Furthermore, reddit users are predominantly male with an above average percentage of male users compared to the general internet population (cf. Alexa.com 2017). During May of 2017, reddit registered 1.3 billion visitors on their website (cf. Statista.com 2017). In data gathered from 2013, 15% of the male population and 5% of the female population of the US who are between 18 and 29 years old visit reddit.com (cf. Pew Research Center 2013). The percentage goes down as the age increases. The age group between 30 and 49 years only shows percentages of 8% males and 5% females who still use reddit. This means that the age group that is most likely to swear is also the age group that is most heavily represented on reddit. Furthermore, the gender that tends to swear more is also overrepresented. This would indicate a higher use of swear words on the site - if people online actually do swear similarly as they do in speech. According to Alexa.com, visitors on reddit are slightly better educated than the general internet population. The date taken from Alexa.com shows info from 2017. The data from the Pew Research Center could be slightly outdated.

Reddit bridges the gap between online forums and social media or, more precisely, it is an "online social system that has the attributes of a forum" (Choudhury and De 2014: 71). Web forums are one particular form of online communication. They are defined as "an online public discussion area where users exchange ideas and information" (Mann and Stewart 2000: 219). Its main attributes are that the discussions are public, they are moderated, the participants can remain anonymous, the discussions are organized by themes and topics and participants do not have to be online at the same time (cf. Witschge 2008: 80). Similar to other forums, interaction on reddit is not only possible, but the main reason for the site to exist. The format of conversation on these forums is therefore, as already mentioned, that of a public discussion for many users to participate in as opposed to conversations in a one to one scenario.

For people to participate in the forums on reddit, all one has to do is sign up with a nickname. An E-Mail account is necessary to validate the chosen password. Other than that, there are no further hurdles. This means that anyone who wants to participate in any discussion can with very little effort. Users can submit posts and comments. Posts are usually links to outside sources or self-written texts which can then be discussed in the comment section. These posts can contain content created by the user or content by other creators. In the comment section users have the opportunity to discuss whatever content was posted. Here, comment trees are created. The first comment is always a comment on the post. For subsequent comments, users have the option to either submit a comment on the post itself or submit comments on previous comments. A comment tree can be created. Such a post with a corresponding comment tree looks like this:

(4) Post in r/showerthoughts with corresponding comments



3022

/r/all Some badass motherfucker in the past domesticated a fucking wolf.

vor 3 Jahre by DamienHanrahan

So that less badass motherfuckers could own labradoodles

593 Kommentare Weitersagen Speichern Ausblenden Schenke Gold melden

Top 200 Kommentare zeige 500

sortiert nach: beste

[–] v99188 657 Punkte vor 3 Jahre

And now we have chihuahuas

Permalink embed Speichern Schenke Gold

[–] CajunAvenger 512 Punkte vor 3 Jahre

The spawn of Satan and cocaine.

Permalink embed Speichern Übergeordnet Schenke Gold

[–] uglyBaby 153 Punkte vor 3 Jahre

That would be a Pomeranian, friend.

Permalink embed Speichern Übergeordnet Schenke Gold

(Reddit.com 2014)

(4) shows a post in the subreddit r/showerthoughts. The post depicts a thought a user was having and wanted to share this with the community. The user v99188 then commented on said post. The comment underneath by CajunAvenger, which is slightly indented, is a comment on the first comment. The comment underneath, again indented, is a comment on the comment above again. Theoretically, this pattern could be repeated forever. Comments that are addressed at the initial post are on the same line as the first comment depicted in (4). Public conversations on reddit are always structured in that order. There is the possibility for users to write each other private messages. These are, as the name suggests, not public and also not part of the data set of this study. Research has shown that comments trees on reddit are hierarchical and follow a topical hierarchy (cf. Weninger et al. 2013: 583).

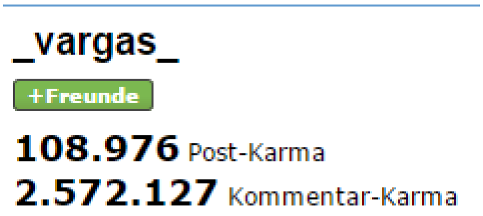
(4) also shows what a particular subreddit can be about. Reddit as a whole is divided into several subreddits. Subreddits are sub-forums which are dedicated to one particular topic each. In this case, the subreddit is devoted to so-called showerthoughts - "miniature epiphanies you have that highlight the oddities within the familiar." (reddit.com/r/showerthoughts). Subreddits are often referred to as r/'name of the subreddit' due to the reddit URL, which directs users to said subreddit, e.g. www.reddit.com/r/showerthoughts. As of the 12.06.2017, there are

1,093,464 subreddits in total. Any user can create a subreddit at any time. Users have the option to subscribe to these subreddits. Once subscribed, the most popular posts of their subscribed subreddits will be shown on their individual front page. The largest subreddit r/askreddit has 17,428,173 (cf. redditmetrics.com). Subreddits can have any topic. There are subreddits for news, politcs, sport, any particular kind of sport, but also humorous ones like r/catsonpizza on which users share pictures of cats sitting on pizza. Depending on the topic, different kinds of posts are allowed. On r/news links and political texts can be shared, whereas on r/pics only pictures are allowed. These rules are set and enforced by creators and moderators of each subreddit. The moderators are also users and can reactively delete posts and comments. This means that certain types of comment or post can be disallowed. What is allowed and what not is dependent on each community, i.e. subreddit. If a user does not find a subreddit that fits their personal interest, the user can create a new subreddit with a specific topic and specific rules.

One defining feature on reddit is the voting system. As (4) shows the post itself has a rating of 3022 and the first comment has 657 points. These points show the difference between up- and downvotes. This means that if a comment has 100 upvotes and 20 downvotes the score would be 80. The exact number of up- and downvotes is not depicted publically. The score dictates which post is shown first on any subreddit or on the front page. The same system works for comments. Users have the option to change the order of posts or comments shown, ordering them by date or score. The default setting is to show those comments and posts on top that have the highest score and most votes.

The score is also referred to as "Karma". Each user has a site on which other users can see how much Karma the other user has accumulated. The Karma of all the user's comments is summed up and the Karma from all submissions or posts is as well. Each score is displayed separately (see (5)).

(5) Collective Karma of a reddit user _vargas_



vargas

108.976 Post-Karma
2.572.127 Kommentar-Karma

(Reddit.com 2017a)

Karma can be seen as a sort of reputation users can gain on reddit. The more upvotes a user can generate for their comments, the more likely it is that the comment is shown at the top of the page. The more prominent a comment or post is displayed, the more likely other users are to participate on those submissions. It can be assumed that users have the tendency to want to submit comments or posts that get a lot of upvotes. The Karma is a sort of reward system for contributions that are deemed worthy via upvotes by the reddit community. The Karma score can, therefore, be seen as a measurement of how well a comment is received. This means that for this present study, the Karma score indicates whether or not the use of swear words was deemed appropriate or inappropriate by the community. It is also unlikely that comments that are deemed to be impolite or offensive by many users receive a high score. There are, obviously, other factors that contribute to the score a comment may receive, most notably the comment's content and the time of submission (cf. Weninger et al. 2013: 581-583).

Lastly, unless actively deleted by the poster[11], every comment or post a user has submitted can be seen. Visitors and other users only have to click on the name of the specific user they want information about. The user's behavior on the site is captured and public for others to see.

[11] as in 'submitter'

4. The Study

4.1. Research Questions

To recall the research questions formulated in the introduction:

How often are swear words used?

How are they received?

Do conversational topics influence the amount of swear words used and their reception?

Based on these questions and the previously discussed findings, this subset of questions was formulated:

- *How often are swear words used and how often are they used compared to other online platforms and spoken English?*
- *On the basis that a higher level of anonymity is one defining feature for reddit users, does anonymity seem to influence swear word usage?*
- *Is the swear word usage rated differently, depending from which semantic field the swear word originates from?*
- *Based on previous offensiveness ratings, are more offensive words rated differently?*
- *How does swear word usage and reception differ across different subreddits?*

When I refer to ratings in the questions above, I mean the score that comments, in which swear words are included, receive. The thought behind using Karma is that appreciation among the community for a comment seems to contradict the effect the breaking of a taboo is deemed to have. If a comment or a word is regarded as highly inappropriate and offensive, the comment should receive a negative score, thus, indicating that the community does not enjoy the comment made and the words used.

In order to answer those questions, two main methods were used. In the first part, the overall count and score for the selected swear words was recorded. From that, conclusions on general swear word usage on the website can be made. In a second step, the usage of swear words is divided among several different subreddits. This should provide insight in as to how much the topic of a conversation influences the amount of swear word used and the response in form of rating that comments containing swear words receive. For each procedure, the methods will be presented and the results discussed. Afterwards, a general discussion concerning the answers to the research questions will be had. But first, the dataset and acquisition is discussed and presented.

4.2. Where does the data come from?

4.2.1. The dataset

The data used in this study comes from the reddit comments and submissions from February of 2017. During February 2017, 70.609.487 comments were submitted to reddit. These comments amount to a total of 2.203.340.001 words. The data is, therefore, quite large, substantial and recent. It enables a view into online language usage at present times that is representative for online discourse on the platform reddit.

For a number of reasons, certain limitations to the dataset had to be made. The most prominent reasons being that, although the main language on reddit is English and most participants come from countries where English is the native tongue, subreddits where another language is exclusively spoken exist. These would not return a comparable number of swear words, yet they would contribute to the overall count of words. This would obviously influence the results and make them unreliable. Furthermore, as the present study focuses on public discussions, only those subreddits are part of the analyzed data to which at least 100.000 comments were submitted. Among these, only one subreddit consisted of more than 100.000 comments where the native language was not English. The subreddit in question was r/de, the subreddit in which every post and comment is German. Comments and words from this community were not included in the analyzed data. The limitation also has practical advantages. It allows to analyze the relation of a subreddit to the usage of swear words. With a total of over a million subreddits, a feasible comparison between every subreddit is not possible. Grouping them, as it is done with the 99 subreddits later on, would still be difficult to compute. Furthermore, the fact that around 40% of all comments and words are still within the corpus of this paper shows that the conversation on reddit has areas in which users agglomerate. Thus, the corpus of this study should be representative.

The result of this limitation of subreddits is that for the study 99 subreddits form the main dataset. The full list of subreddits can be found in appendix (1). To those 99 subreddits, a total of 30.322.546 comments were submitted. So, while there is only a fraction of the total number of subreddits that provide data for this analysis, 43% of all comments submitted to reddit are still included in the data set. The 30 million comments amount to a total number of 869.514.814 words, which are around 40% of the total number of words. These limitations also decrease the number of authors which contributed to the discussion. If the number of distinct authors per subreddit is added, a total of 4.679.850 different authors took part in the discussion. This number counts authors who participate in different subreddits as different authors, though. That means that a user who submitted comments to r/AskReddit and one or more comments to r/funny is counted as two different authors. The number of distinct authors, meaning that authors who contributed to several subreddits but are only counted once is 1.835.305. Thus, authors tend to comment in more than one subreddit.

The last aspect that was taken from the source material was the score or the Karma. For all the comments of the 99 subreddits, the average Karma was 10,46. This is higher than the overall average score for all comments from all subreddits which is 7,46. This means that, in general, comments submitted to subreddits where there are more active discussions, gain more recognition and appreciation on average. This seems logical as there are more people who can possibly upvote a particular comment. Yet, there are also more people who can downvote a particular comment.

So, although limitations to the complete dataset had to be made, the remaining data is still large, substantial and representative. Table 2 summarizes the attributes of the dataset below.

Table 2 Attributes of dataset

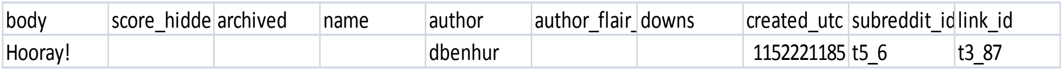
Category	Amount
Subreddits	99
Comments	30.322.546
Words	869.514.814
Distinct Authors	1.835.305
Average Score (Karma)	10,46

4.2.2. Google BigQuery

To obtain the data, I used the openly available data search program Google BigQuery. BigQuery enables users to upload large quantities of data which can then be queried using the Structured Query Language SQL. Google enables users to use two different SQL dialects, the standard SQL and Legacy SQL. The latter is a format which was introduced specifically for the usage with BigQuery (cf. BigQuery 2017a).

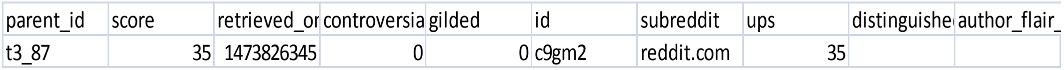
The data is stored in tables. The table from which the dataset for this thesis is taken is called "fh-bigquery:reddit_comments.2017_02" (BigQuery 2017b). The table includes, as previously mentioned, all publically available reddit comments from February of 2017. Along with the raw text of the comment, more information is saved and can be accessed. Within the table, each row represents one single comment. Below, one row taken from "fh-bigquery:reddit_comments.2006"[12](BigQuery 2016) can be seen.

(6.1) Example of row in "fh-bigquery:reddit_comments.2006" from BigQuery



body	score_hidde	archived	name	author	author_flair_	downs	created_utc	subreddit_id	link_id
Hooray!				dbenhur			1152221185	t5_6	t3_87

(6.2) Continued row of example 4.1



parent_id	score	retrieved_or	controversia	gilded	id	subreddit	ups	distinguishe	author_flair_
t3_87	35	1473826345	0	0	c9gm2	reddit.com	35		

[12] In order to obtain the data, a query had to be processed. As that query contributed to the monthly quota of 1TB for free BigQuery use, I chose a smaller table to limit my data usage. The format of the table for 2006 is identical to the format of the table for February 2017. For readability, one row of the table was divided into two rows in (6.1) and (6.2).

For clarification, I will explain what the important columns contain and what they represent. Columns that are not mentioned below are not important for this thesis, do not contain information that is visible to the public or are always empty.

The first row in (6.1) and (6.2) shows the header of each column, beneath one can see the actual first row with the fields filled in. In the columns where there is no value added, no values are available for this particular example. The first column is titled 'body' and includes the complete comment. In example (6), the comment is only one word long but each field in the 'body' column contains exactly one comment, regardless of the actual length of the comment. As it will be explained below, BigQuery and SQL allow searches for fields and texts. However, the system or the language will regard one field within a table as one single value. Thus, if words are to be searched, the text in the 'body' column has to be divided into words.

The 'score_hidden' column only has two sets of possible values. Scores, or Karma, can be hidden for different reasons. If the score was hidden, the value in the field would be 'true'. In this case, the score is not hidden, thus, no value is shown. Threads, the comment trees underneath a submission or post, can be archived. This means that users cannot comment on the subject anymore. As the field 'archived' is empty in (6.1), the comment tree from which the comment was taken is still open for further discussion.

The 'author' column shows the username of the author of the comment. Each user has a unique username. The 'author_flair' are slogans that certain users can have. These slogans are displayed next to the user name and are dependent on the subreddit. In subreddits that have a certain type of sport as their topic it is quite common that users have flairs that show their affiliation to a certain club or player. The column 'downs' shows the number of downvotes the comment received. In this case, there are no downvotes, thus, the column is empty. The next four columns contain 'IDs'[13] which are only of importance for computer application. The 'parent_id' is the ID for the comment to which the comment in (6.1) and (6.2) is an answer. The other two important columns are the column 'score', which shows the difference between the up- and the downvotes. This is the

[13] Identification numbers or codes

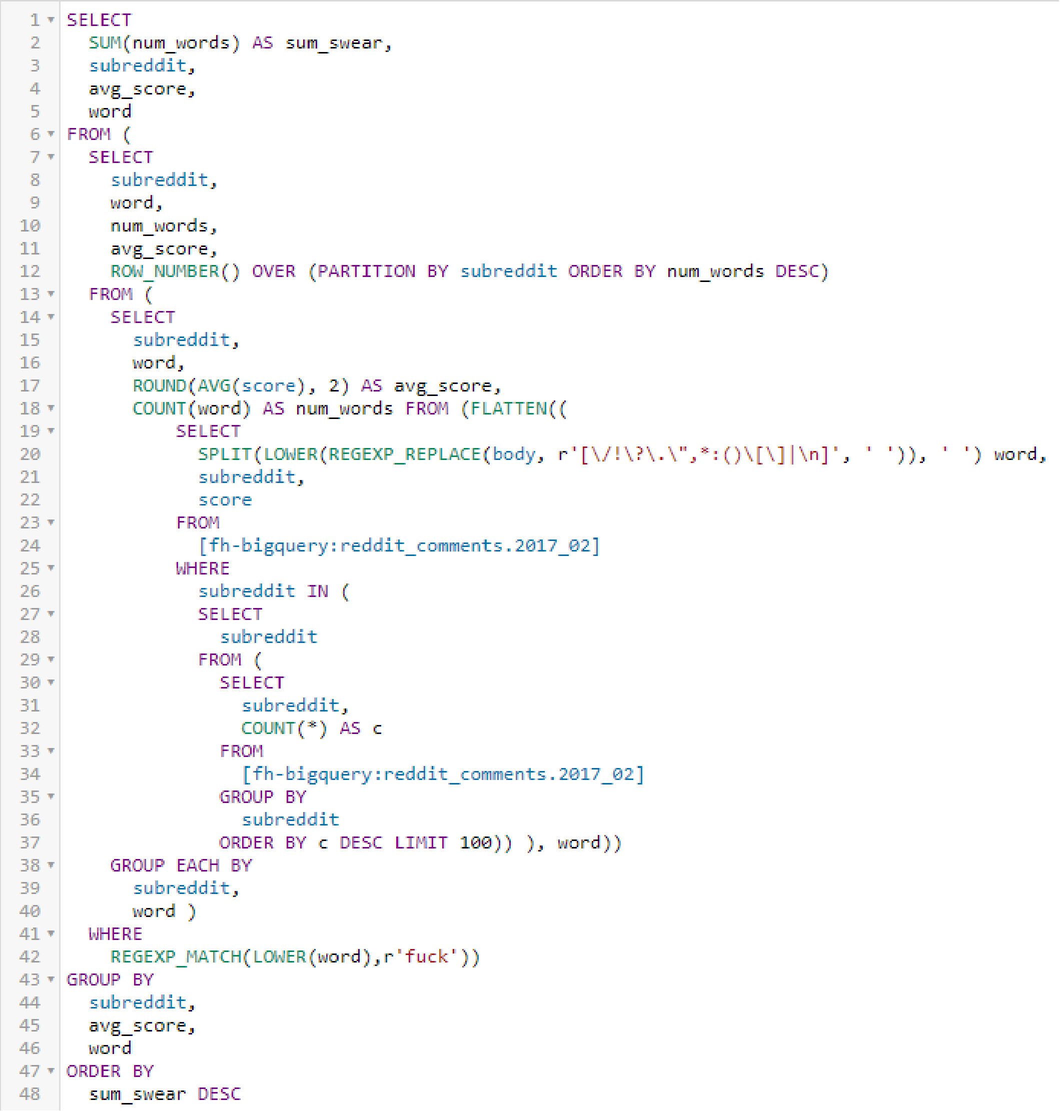
so called Karma. The other column is the 'subreddit' column, which contains the name of the subreddit. The column 'ups' shows the upvotes.

4.3. Method 1 - Overall occurrences and ratings

4.3.1. The SQL codes

Now that the structure of the source material is laid out, the code and query with which the data from the source table was obtained will be explained. In order to show how the data was taken, the query for one particular expression will be expounded in detail.

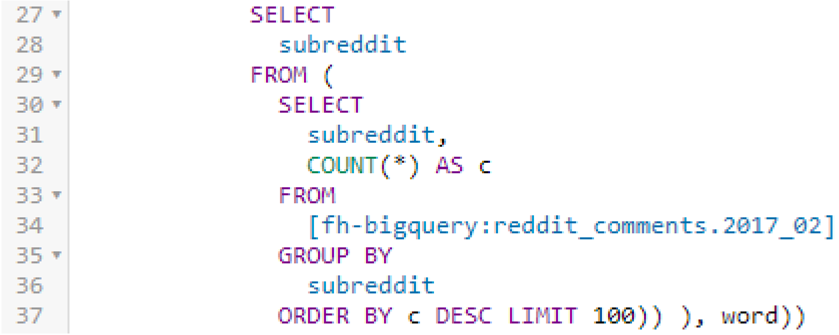
(7) Code for querying BigQuery for *fuck*



```
SELECT
  SUM(num_words) AS sum_swear,
  subreddit,
  avg_score,
  word
FROM (
  SELECT
    subreddit,
    word,
    num_words,
    avg_score,
    ROW_NUMBER() OVER (PARTITION BY subreddit ORDER BY num_words DESC)
  FROM (
    SELECT
      subreddit,
      word,
      ROUND(AVG(score), 2) AS avg_score,
      COUNT(word) AS num_words FROM (FLATTEN((
          SELECT
            SPLIT(LOWER(REGEXP_REPLACE(body, r'[\/!\?\.\",*:()\[\]|\n]', ' ')), ' ') word,
            subreddit,
            score
          FROM
            [fh-bigquery:reddit_comments.2017_02]
          WHERE
            subreddit IN (
            SELECT
              subreddit
            FROM (
              SELECT
                subreddit,
                COUNT(*) AS c
              FROM
                [fh-bigquery:reddit_comments.2017_02]
              GROUP BY
                subreddit
              ORDER BY c DESC LIMIT 100)) ), word))
    GROUP EACH BY
      subreddit,
      word )
  WHERE
    REGEXP_MATCH(LOWER(word),r'fuck'))
GROUP BY
  subreddit,
  avg_score,
  word
ORDER BY
  sum_swear DESC
```

The code above searches for the word *fuck* and all words containing *fuck* within the source table. The SELECT command specifies from which columns data is extracted. The FROM command selects the source table from which the data will be taken. This is true for all functions. As (7) shows, several functions can be processed within one query. Each function has a table as a result from which the overlaying function can take data from. The function that is written most to the right is the first function to be processed. It is, therefore, in that function that the limitation of subreddits is defined. This is described in (8).

(8) Function for top 100 subreddits based on comments submitted

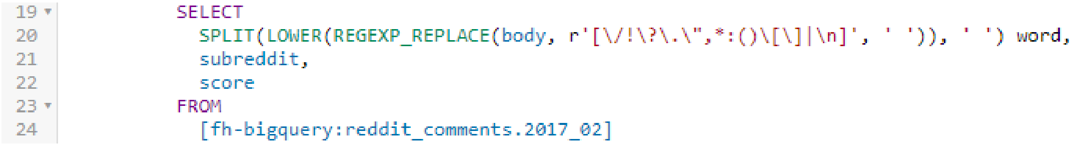


```
27              SELECT
28                subreddit
29              FROM (
30                SELECT
31                  subreddit,
32                  COUNT(*) AS c
33                FROM
34                  [fh-bigquery:reddit_comments.2017_02]
35                GROUP BY
36                  subreddit
37                ORDER BY c DESC LIMIT 100)) ), word))
```

The function searches for subreddits in the source table. Each function shows its results in a destination table. These tables can be accessed within one larger query. The COUNT command counts the number of fields in a specified column. As there is no specific column defined, the command counts all rows that are applicable. In line 32 the number counted is defined as the variable 'c'. Thus, in this function comments are counted, as each row of 'fh-bigquery:reddit_comments.2017_02' represents one comment and no further specification to the selected columns is made. The destination table, the result, shows the number of rows per subreddit. This is defined by the GROUP BY command. Then, the order of the subreddits depends on 'c', the number of comments for each subreddit in descending order. This is defined by the ORDER BY c DESC command. The limit here is set to 100. Thus, the top 100 subreddits based on comments submitted are defined by this function and displayed in the destination table. Note that, in this case, the destination table does not include the actual text of the comments. It just displays the top 100 subreddits. In order to not overload the computing power of the Google processors, the code had to be modified. Therefore, the subreddits are limited within a separate function. This means that the source table 'fh-bigquery:reddit_comments.2017_02' had to be accessed twice, once in line 34 and once in line 24 (cf. (7)). The data resulting from the query in (8) includes the German subreddit r/de. That data had to be excluded afterwards.

(9) shows the next function which is designed to divide the comment text from the table 'fh-bigquery:reddit_comments.2017_02' into single words.

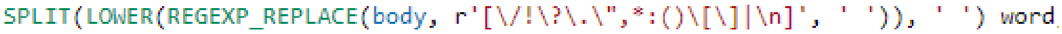
(9) Function for dissecting the 'body' column into words



```
19          SELECT
20            SPLIT(LOWER(REGEXP_REPLACE(body, r'[\/!\?\.\",*:()\[\]|\n]', ' ')), ' ') word,
21            subreddit,
22            score
23          FROM
24            [fh-bigquery:reddit_comments.2017_02]
```

The function selects three columns, the 'word', the 'subreddit' and the 'score' column. As the source table 'fh-bigquery:reddit_comments.2017_02' does not provide single words in single fields, the column in which comments are shown has to be divided into single words. In the source table, each comment is one single field in a table. The text is one entity in the 'string' format. Thus, if the 'string' would not be divided into single pieces of string, i.e. the comment would not be divided into single words, a function could only provide the number of comments which include a word or term, but not the number of words.

The division into words is done with the command.

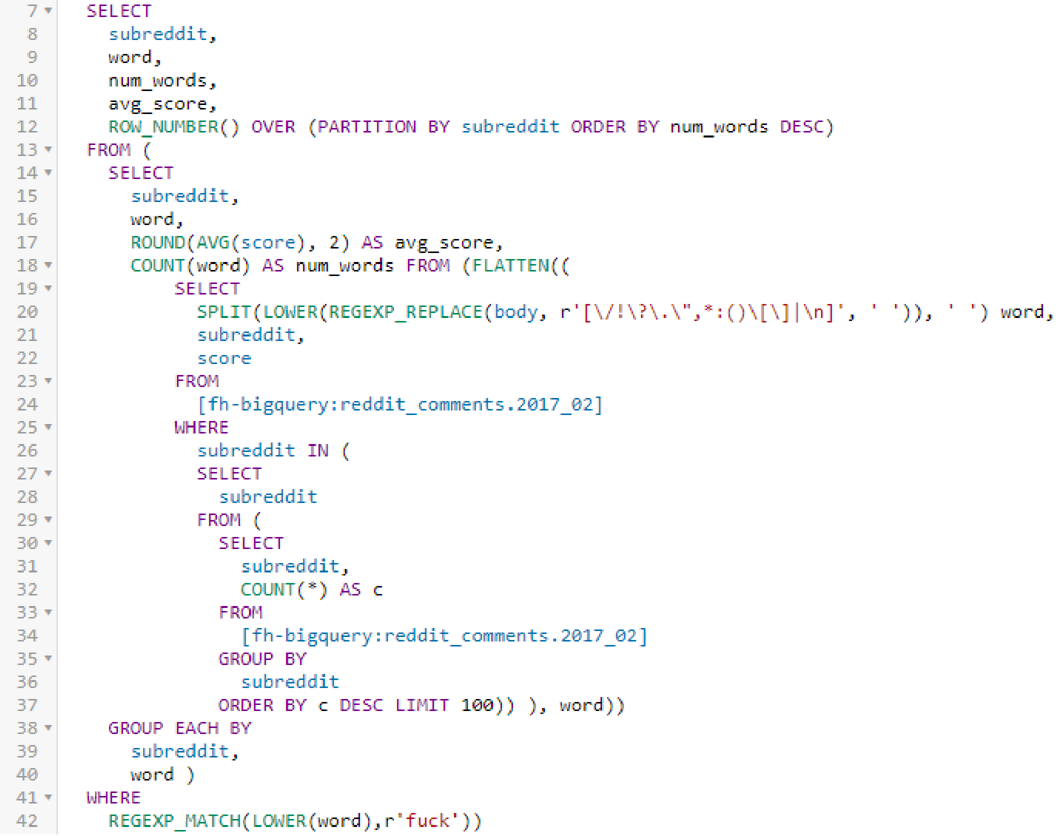


```
SPLIT(LOWER(REGEXP_REPLACE(body, r'[\/!\?\.\",*:()\[\]|\n]', ' ')), ' ') word
```

This defines how the comment is split up. The LOWER command erases any form of capital letter or sign. The REGEXP_REPLACE command erases certain defined sings. This means that, in this example, the "," "/" "!" "?" "\" """ "." "," "*" ":" "(" ")" are replaced by a whitespace, the ' ' before the last bracket in the command. That is an empty space that separates words in text documents. This means that terms like 'fuck!', 'fuck?' or 'fuck.ing'[14] would just return as 'fuck' and can be counted as the same word. For counting purposes, all forms of the word *fuck* were be added together. Lastly, the command ends with defining the dissected column 'body' as the variable or as the new column 'word'. Now, each word represents its own column within the table and can be counted.

[14] As an example.

(10) Function for counting and searching for *fuck*



```
 7   SELECT
 8     subreddit,
 9     word,
10     num_words,
11     avg_score,
12     ROW_NUMBER() OVER (PARTITION BY subreddit ORDER BY num_words DESC)
13   FROM (
14     SELECT
15       subreddit,
16       word,
17       ROUND(AVG(score), 2) AS avg_score,
18       COUNT(word) AS num_words FROM (FLATTEN((
19           SELECT
20             SPLIT(LOWER(REGEXP_REPLACE(body, r'[\/!\?\.\",*:()\[\]|\n]', ' ')), ' ') word,
21             subreddit,
22             score
23           FROM
24             [fh-bigquery:reddit_comments.2017_02]
25           WHERE
26             subreddit IN (
27             SELECT
28               subreddit
29             FROM (
30               SELECT
31                 subreddit,
32                 COUNT(*) AS c
33               FROM
34                 [fh-bigquery:reddit_comments.2017_02]
35               GROUP BY
36                 subreddit
37               ORDER BY c DESC LIMIT 100)) ), word))
38     GROUP EACH BY
39       subreddit,
40       word )
41   WHERE
42     REGEXP_MATCH(LOWER(word),r'fuck'))
```

In (10) the functions discussed in (8) and (9) are included. The lines that need further explaining are those between 12, 14 and 18, as well as the lines 41 and 42. Between lines 14 and 18 the columns which are to be shown in the result from both previous functions are selected. The `ROUND(AVG(score), 2) AS avg_score,` command selects the score and rounds that number to two digits after the comma. The command in line 18 counts all instances from the column 'words' defined in the previous function. The number is then given a name, thus, creating a new column in which the number for each word will be presented. The FLATTEN command causes values that are created by the function in (9) and that appear more than once to be treated as individual separated values. This is necessary for the program to treat the same word, e.g. *fuck,* as the same but countable value. Otherwise, the query would return one instance for 'fuck', one for 'fucking' and so forth, but not count how many of those instances there are in the corpus. FLATTEN, thus, causes the term to be countable.

Lines 41 and 42 in (10) define what word is searched for. The WHERE command selects fields from a certain column in which a certain value is found. In this case,

the definition of that value is `REGEXP_MATCH(LOWER(word),r'fuck'))`. REGEXP_MATCH returns a string value where 'fuck' is included. This means that all variations and words which contain the letters f, u, c, k in that order and in succession are selected. LOWER(word) causes the values in the column 'word' - which was created by the function in (9) - to be independent from capitalization. Lastly, the r'fuck' command defines what pattern is looked for. Thus, the whole function shown in (7) returns the sum of the words that are found in the corpus, the subreddits in which they appear, the average score of each word and the different variations of the word.

4.3.2. Advantages and disadvantages of this method

The advantage of this method is that a large corpus can be processed very quickly. It also automatically returns all variations of the word that is looked for. The dataset selected is also very recent, thus, creating contemporary results that are representative of actual language in use. The broader context, the topic in which the swear words appear, can be determined, the reception by the community has a quantifiable value which can be evaluated. There are disadvantages, however. The method limits swear word research in the sense that only words can be looked for that are determined beforehand. This method also does not allow a closer look into the context which these terms appear in. This means that swear words that are ambiguous can be looked for but the intend behind their usage cannot be determined. This is particularly important for religious swear words. Words like *hell* or *Jesus Christ* can be used in a variety of contexts in which their usage is not impolite, offensive or a swear word. The terms could be used with their actual meaning. Thus, these cannot be included in the results, as they are too ambiguous to be considered primarily swear words.

4.3.3. The swear words

As mentioned above, the terms that were looked for had to be swear words. Thus, the following terms were selected:

- bitch
- retard

- nigger
- fag
- fuck
- shit
- cunt
- ass
- damn
- piss
- cock

This search includes all variations for each word. The search for *cock* returned words like *cocktail* or *cockroach*. These were omitted from the results. The list consist of a variation of swear words taken from the Ofcom list (cf. 2016: 44), the seven 'dirty' words and the ten phrases that make up 80% of swearing in spoken English (cf. Jay 2009: 156) with the omission of words that are too ambiguous. Stronger terms like *retard* and *nigger* were also included, to see whether these terms were used comparatively often.

In addition to those words, a number of abbreviations were chosen and looked for. As there is a large amount of abbreviations in use, there is no claim to have a full list. The abbreviations were chosen based on personal experience on the reddit platform. The abbreviations that were chosen are:

Table 3 Abbreviations

Abbreviation	Meaning
af	as fuck
fu	fuck you
wtf	what the fuck?
omg	oh my god
omfg	oh my fucking god
lmao	laughing my ass off
ffs	for fuck's sake
wth	what the hell?
stfu	shut the fuck up
dafuq	the fuck?
bs	bullshit

sob	son of a bitch
pos	piece of shit
fml	fuck my life
bamf	bad ass motherfucker

4.4. Method 2: Differentiating between subreddits

Reddit does not offer as much background data of its users as the platforms MySpace and Twitter do. Thus, information concerning the age or gender of reddit users cannot be extracted from source material on BigQuery. There is, however, other information available. Especially, the subject of the subreddits provides the opportunity to connect the overlying subject of a conversation with the usage and appropriateness of swear word usage. Looking into one or two single subreddits would offer an insight into the difference between the two, but general statements concerning swear words would be hard to make. Thus, after reviewing the 99 subreddits closely, seven general topics were established. To those seven topics, five subreddits were added each. Thus, the number of subreddits is equal across all seven categories and comparisons can be made. The categories that the subreddits were distributed to were 'comedy/memes', 'conversation', 'technology', 'news/politics', 'media', 'sport' and 'games'. These categories are based on the number of occurrences of subreddits that can be distributed into these topics. Among the 99 subreddits, almost all of them could assigned to one of these five groups.

'Comedy' includes five subreddits whose topic are jokes, funny stories and internet memes. 'Conversation' includes subreddits whose purpose it is to have a discussion. Topics about these discussions can vary. 'Technology' includes subreddits which deal with computer technology. Subreddits that deal with current events of importance were assinged to 'News/politics'. In the category 'media' the topics visual media is shared and discussed. This includes videos, pictures, movies, gifs[15] and animes[16]. The category 'sport' includes subreddits which deal

[15] GIF - Graphic Interchange Format. They are basically short videos without sound.
[16] The Japanese form of cartoons.

with a specific sport each. Subreddits whose topics are anything videogame related were assigned to 'Games'. Which five subreddits were chosen per category was based on the number of words each subreddit is made up of. Thus, the five subreddits where users commented using the most words were included in the respective category.

The tables 7 and 8 containing all categories and corresponding subreddits can be seen below.

Table 4 Categories and corresponding subreddits part 1

Comedy Memes	Conversation	Technology
r/CringeAnarchy	r/AskReddit	r/Amd
r/funny	r/AskMen	r/Android
r/jokes	r/AskWomen	r/pcmasterrace
r/AdviceAnimals	r/relationships	r/buildapc
r/dankmemes	r/explainitlikeimfive	r/technology

Table 5 Categories and corresponding subreddits part 2

News Politics	Media	Sport	Games
r/politics	r/videos	r/nba	r/forhonor
r/news	r/pics	r/nfl	r/Overwatch
r/worldnews	r/gifs	r/hockey	r/NintendoSwitch
r/The_Donald	r/movies	r/soccer	r/LeagueofLegends
r/europe	r/anime	r/SquaredCircle	r/Gaming

For each of those subreddits, the number of swear words across all chosen terms and abbreviations was summed up. Afterwards, the number of swear words was divided by the total number of words for each subreddit. Thus, the percentage of swear words per subreddit was the result. For the overall percentage of swear words within each category, the sum of all swear words per category was divided by the sum of all words per category. To give further insight into the perception of swear words and the difference between these categories and subreddits, the ratio of average Karma from all comments for each of the categories and each of the subreddits and the Karma from comments containing swear words was calculated[17].

4.5. Results part one - Overall occurrences and ratings per word

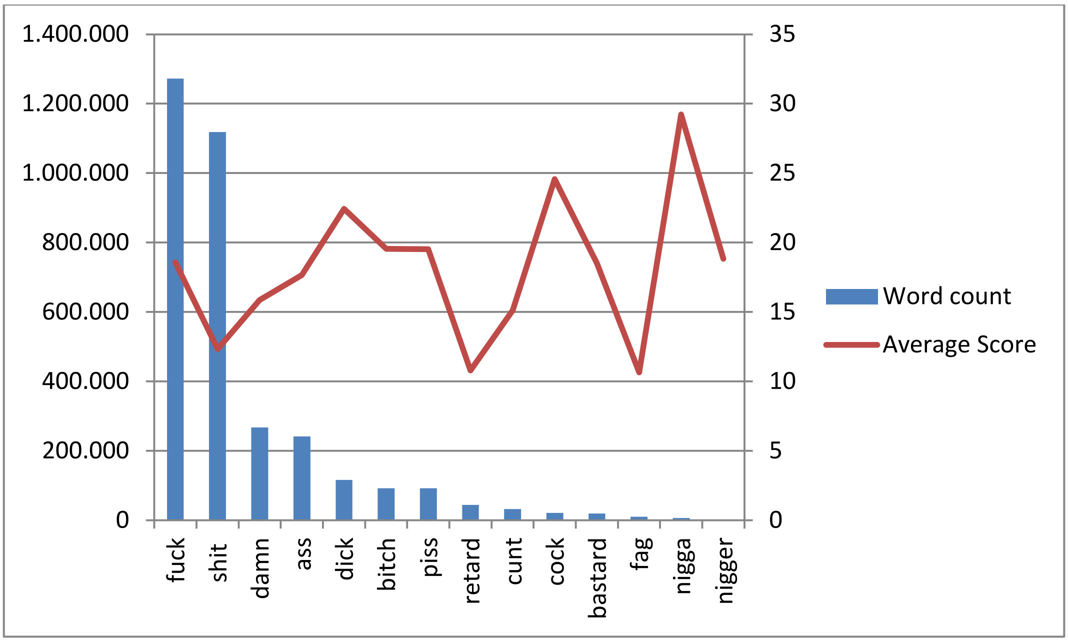
Table 6 shows the results from the first method divided into the different words, their score as well as the percentage of how much each term makes up and amount of swear words used. Table 6 deals with written-out words and Figure 1 shows the same results in a diagram to visualize distribution and ratings. The numbers and ratings are calculated from all variants of each term that were found.

Table 6 Swear words, word count and Score

Word	Word count	Average Score	Percentage of total swear words
fuck	1.272.219	18,57	38.14%
shit	1.118.610	12,33	33.53%
damn	266.959	15,85	8.00%
ass	241.153	17,65	7.23%
dick	116.553	22,43	3.49%
bitch	92.404	19,54	2.77%
piss	92.281	19,52	2.77%
retard	44.404	10,79	1.33%
cunt	32.135	15,13	0.96%
cock	21.107	24,57	0.63%
bastard	19.379	18,52	0.58%
fag	10.225	10,65	0.31%
nigga	6.205	29,23	0.19%
nigger	2.198	18,84	0.07%
Total	3.335.832	16,29	100%

[17] A full table containing the necessary data, that is word count, average rating, swear word count, swear word percentage, swear word rating, ratio for each of the subreddits that provide data for the categories can be found in appendix (2).

Figure 1 Diagram for Table 6



The last column depicts what percentage of the overall swear word usage each word is. The swear word that is used most often is *fuck,* followed closely by *shit.* The difference between them is still around 150.000 occurrences, but compared to the amount of occurrences the remaining swear words have, they are close. In total, *shit* and *fuck* account for almost 72% of all swearing within the corpus. After those two which both occur more than one million times within the corpus, there is a big gap to the next terms. Both *ass* and *damn* occur around 250.000 times. That is less than a quarter of the occurrences that each *fuck* and *shit* amount to, respectively. *Fuck* is even five times as likely to be used as *ass.* With less than half of the occurrences than *ass*, *dick* is the next most common swear word. *Dick, bitch* and *piss* all occur around 100.000 times and together make up around 9% of all swear words used. The bottom seven swear words are all considered to be highly offensive. *Bastard* and *cock* are mentioned in the Ofcom list in the category of "Strong words" (2016: 44). *Cunt* is listed in the "Strongest words" category (Ofcom 2016:44). *Retard* has become unacceptable in recent history, as well as *fag*. These terms are all within the 10.000 to 50.000 thousand occurrence range. The only term making up more than 1% of overall swear word use is *retard.* Both *nigger* and *nigga* are racial slurs and they are the only terms that appear fewer than 10.000 times. I distinguished between these two terms, as *nigger* is purely a racial slur. *Nigga* on the other hand can be used within the black community. One indication to support that claim is their score. With an overall score of almost 30 *nigga* has the highest score of all swear words. It has to be mentioned, though, that *nigger* has a comparatively high score as well. A reason for the surprisingly positive response the word *nigga* is that the subreddits that use that term the most are r/BlackPeopleTwitter and r/hiphopheads. Based on the topic, one can assume that an over-proportionally large amount of users are of dark skin color. The term *nigga* can be used within the black community without its highly offensive connotation. A higher usage and a higher rating in those communities causes the high Karma. For comparison, the term *nigga* is rated at 5,65 in the subreddit r/pics, a subreddit were photos are submitted and discussed. The score in r/BlackPeopleTwitter for *nigga* is 51,25 with 16 times the occurrences compared to r/pics.

It has to be noted that the average score in Table 6 is dependent on the occurrences. Thus, the average score for *fuck* has five times the impact on the overall average score compared to *ass*. This explains why only four words have a lower score than the overall average score and ten words are above that average.

With an average score of around 10, *fag* and *retard* were rated the worst by the community. Interestingly, *fag* and *faggot*, both terms that have a homophobe connotation, are used most often in the subreddit r/The_Donald. R/The_Donald is subreddit by and for supporters of the American president Donald Trump. The rating for that term is around 12 in r/The_Donald whereas it is 3,32 in the subreddit r/videos. With over 2.000 occurrences, r/The_Donald is responsible for more than a fifth of all occurrences within the corpus. A similar case can be made for *retard*. This term is also used most often in the Trump subreddit. However, other subreddits use the term, too, and rate them higher. The subreddit r/AskReddit, the biggest subreddit of the website is responsible for almost the same amount of occurrences, 3143 for r/The_Donald compared to 3050 for r/AskReddit, but the word is rated much higher. The ratings are 12 for Trump and 18 for r/AskReddit. It also has to mentioned here that the amount of comments and words for r/AskReddit is much higher than the numbers are for r/The_Donald. A comparison for overall swear word usage partitioned into different conversation topics will be made later on.

Fuck and *shit* have average scores that differ quite a lot from one another, yet they are in the middle compared to all the other terms. This is not surprising, as they hover around the swear word average score and they have the biggest impact on the overall score. The highest scoring terms are *dick* and *cock,* both terms that refer to male genitalia and *nigga*[18], a term mostly used between members of the black community to address each other. The overall average score for all written-out swear words is 16,29. It is around 60% higher than the average score per comment. In fact, all score averages, regardless of swear word, are higher than the average score if only slightly in the cases of *fag* and *retard*. This indicates that swear word usage is regarded as rather positive by the community. More on the implications of these results is discussed in section 5..

Figure 1 also shows how independent the amount of usage and the connected scores are. There is no linear correlation between score and number of occurrences. The amount of usage can, therefore, not be motivated purely by the desire to gather more Karma. It can be assumed that *shit* and *fuck* offer the most possibilities for usage within any given context.

Table 7 depicts the overall count and score for the abbreviations.

Table 7 Abbreviations, Count and Score

Abbreviation	**Count**	**Score**
wtf	83.447	9,57
lmao	74.207	10,07
bs	33.600	8,36
omg	31.167	11,53
af	25.791	10,55
ffs	11.397	9,23
pos	4.278	16,14
stfu	4.189	5,59
omfg	2.843	10,55
fml	2.837	20,12
fu	2.579	10,68
sob	2.266	15,41
wth	2.108	9,54

[18] with that spelling

dafuq	1.457	5,96
bamf	244	8,54
Total	282.410	10,92

As there is only a small subset of possible abbreviations, the overall fewer occurrences are natural. This subset of abbreviations is not as representative for all swearing as the list of complete words is. Still, it does provide an insight into swear word usage that is particular to the internet and computer mediated discourse. Abbreviations also provide a sort of censorship or euphemism, nevertheless it cannot be claimed that the primary motivation for using abbreviations is to censor taboo word usage. But the characteristics are similar in certain aspects. The taboo word in question is not formulated fully or, in case of *dafuq* written and spelled differently. It works on the same basis that euphemisms like *fudge* for *fuck* rely on. They are similar enough that the connection is clear in a given context. However, they differ vastly in the realm of ambiguity. These abbreviations are all unambiguous in their meaning. In that regard, they resemble spelling censorship like *f*ck* for *fuck.* There is no doubt what word is referred to, neither to author nor reader, but the effort to not spell out the word completely is still being made.

The number of occurrences and the score offer interesting insights into swear word usage and discourse on reddit. Overall, abbreviations are used less than written-out words. This is also anticipated, as the contexts in which these abbreviations can be used is much smaller than those in which a word can be used. *Fuck* in particular can be used as every constituent of a sentence, i.e. 'Fuck the fucking fuckers.' This is, obviously, not possible with abbreviations. What is interesting is that the overall score for abbreviations is much lower than the score for complete words. Part of this can also be explained by the limited contexts in which abbreviations appear. Additionally, this result indicates that a motivation behind the use of acronyms and abbreviations is not to decrease the possible harm of swear words. Using these terms does not provide a quantifiably better response by the community. Using acronyms for the sake of saving time and not having to write out phrases that are repeated often seems to be a much more likely motivation.

The difference in score between the different abbreviations shows contrasts in how swear words are used and received. In this case, the phrases that represent insults are rated lower than those of surprise or laughter. Especially, *stfu* with an average rating of 5,96 has the lowest score of all terms that were collected. This might be because *stfu* is typically directed to someone within the conversation. It, therefore, has the potential to upset people who participate in the conversation more than phrases that can be directed to outsiders of the conversation more easily. The abbreviation also addresses the very act of participating in a conversation. To *shut the fuck up* would mean to not engage in conversation any further. As conversation is the very reason for reddit to exist, this is not rated highly by other members of the community.

The other direct insult is *fu* which stand for 'fuck you'. It is rated much higher than *stfu* with an average score of 10,68, which is slightly lower than the overall score. It is used less than *stfu,* though. With 2579 occurrences it makes up only around 1% of all abbreviations. *Stfu* makes up around 1,5%. This shows that abbreviations which represent direct insults and that are considered rude, according to the score, are used very rarely. Abbreviations that represent surprise and laughter are used much more often. They are not, however, rated higher. As the examples *wtf* and *lmao* show, their use is very common compared to the other abbreviations. They are not rated higher than their counterparts. Both are rated between 9,5 and 10. Both are slightly below the overall average score.

The term that achieved the highest rating is *fml.* It can be interpreted as a self-depreciating term. As it includes the use of swear words, it is also likely to be used in rather informal contexts. It is, thus, likely to be used in contexts in which there is a comedic element to the story. As the internet is famous for the distribution of so-called 'fail' videos and stories, comments in which such a fail is described and a comedic element is added are likely to receive a positive recognition. *Fml* is, therefore, likely to be used in comedic contexts. Thus, it receives a positive response. The high score also shows that users on reddit know these factors and, seemingly, employ them effectively. This knowledge is most likely subconscious.

The highest usage is returned for phrases that represent surprise. With *wtf, omg, omfg, dafuq* and *wth* a combined occurrence of 121.022. So, a third of the abbreviations accounts for 43% of all abbreviation occurrences. Apart from *dafuq*, their ratings are around the 10 mark. Interestingly, the usage of *wtf* and *wth* is quite different. *Wtf* is the most used acronym by far, whereas *wth* is used very rarely. In terms of offensiveness, *hell* is ranked lower than *fuck*. The idea that a lesser grade of offensiveness results in higher usage is, therefore, unsubstantiated. A higher usage would indicate a wider variety of contexts in which they can appear. In this case, a higher grade of offensiveness does not represent a higher grade of inappropriateness. Both terms are rated roughly the same but their number of usage is very different. One reason behind that difference could be that in order to convey emotion via written language, stronger terms have to be used. Vocal means of conveying emotion are not available to the author. Thus, a stronger word could evoke a stronger response by the community.

The second highest rated acronyms are *sob* and *pos*. Both are insults. They have to be directed towards someone, however. It is possible that the addressee of that insult is not a participant of the conversation. Even if the addressee is a participant of the conversation other words still have to be added to make the target clear to everyone. That is one key difference to *stfu* which, theoretically, can be addressed to someone outside of a conversation as well. Without any added context like pronouns, it is directed towards another participant of the conversation, namely the author of the comment the insult is an answer to. Thus, the difference in rating can be explained. Another explanation could be that the contexts in which the usage of *sob* and *pos* is appropriate, are better understood by the community. Although, this is speculation, *pos* and *sob* could represent the emotions a community of users have towards an outsider. If that attitude is understood correctly, the use of both terms can be regarded highly by other members of the same conversation.

Table 8 Count and Score overall

	Count	Average Score
Swear words incl. abbreviations	3.618.242	15,87
Words total	869.514.814	10,46
Comments	30.322.546	10,46
Swear words per 1000 words	4	-
Swear words per comment	0,12	-
Swear words per author	1,97	-

Overall and including abbreviations, swear words make up 0,4% of the whole corpus. On average, there are swear words in every eighth comment, or, formulated differently, there is a swear word in 12% of the comments. However, this would only be a representative result, if the assumption that swear words are only used once per comment, would be true. As this cannot be granted, the number of 12% of comments containing swear words is purely mathematical. On average, reddit users used around two swear words each throughout the whole month of February. The comments containing swear words gain more Karma than those which do not. This can be interpreted in two ways. On the one hand, it suggests that swear words have a positive effect on the score. The problem with this conclusion is that other factors, most notably the actual content of the comment, including humor, timing of the comment in the comment tree, expressed opinion etc. is not taken into account. On the other hand, the results do suggest that swear words do not affect the score negatively. This is either due to swear words being used in colloquial comments which include humor and are liked by the community or that they are used to express opinions and emotions that are shared by the community. In general, the results show that the reddit user understands the discourse on the site and is able to employ means such as swear words in acceptable ways. As previously discussed, even if not intentional, the use of swear words can be understood as impolite. Thus, there is a danger of utterances that contain swear words to be rated negatively. This is not the case. The reddit user is generally aware of how, when and in what circumstances swear word usage is acceptable. Otherwise, comments containing swear words would be rated more negatively. This leads to a further indication. Although, the swear word usage is not higher compared to other online platforms[19], the ratings suggest a mostly polite use of those terms. Even the most offensive terms like *nigger* are rated higher on average than the average comment. Thus, swear word usage with the intend to offend larger parts of the participating users is minimal or ineffective. It seems that the main reason on reddit for swearing is to evoke emotion that is shared by most of the community. That does not contradict the act of swearing aimed at someone directly. A conversation within the community can still be held and be potentially offensive to one particular member or a third party. The overall positive ratings just suggest, that usually the opinion on the subject or person is shared by the community with which that conversation is held. An emphasis on that opinion via the use of swear words explains the higher ratings. This is in line with the findings by Fägersten (2007). She found that the most often found use of swear words is not impolite and not offensive.

4.6. Results part two

Table 9 Swear word percentage and Karma ration per category

Category	Swear word in %	Ratio of average Karma to swear word Karma
Comedy/Memes	0,7	1:1,14
Conversation	0,4	1:1,68
Technology	0,2	1:1,56
News/Politics	0,5	1:1,34
Media	0,5	1:1,16
Sports	0,7	1:1,24
Games	0,4	1:1,28

[19] see the section "Discussion"

The results shown in Table 9 resemble the overall results in that the score that comments with swear words received is generally higher than the overall score. Interestingly, the biggest difference between overall score and swear word score is achieved in the categories 'conversation' and 'technology' and the lowest in 'media' and 'comedy'. Especially, the low ratio in 'comedy' seems surprising, as humor has been stated to be an appropriate circumstance in which swear word usage is not impolite (cf. Jay 2009: 155). Thus, a higher difference could be expected here. On the other hand, jokes and funny comments also have the option to miss their target - making the audience laugh - regardless of swear word usage or not. The argument Jay makes is supported, though, by the relatively high number occurrences. With 0.7%, the ratio of swear words to non-swear words is comparatively high. It is almost double the overall percentage of 0.4%. The same ratio is achieved by the 'sports category'. This might be explained by the nature of conversations about sport. In r/soccer for example, so called Match Threads are opened, each being dedicated to a single football match, usually of international importance. In these threads, people comment on the events of the match. As there are usually fans from both sides, the conversation can become emotional. As previously discussed, swear words can be used to enhance an emotional expression. Therefore, a slightly higher percentage of swear words among sport subreddits is explainable. As there are conversations that are not connected to ongoing events on those subreddits, the overall percentage of swear words to non-swear words is still relatively low. Also, the use of a swear word does not improve the score drastically, compared to the other categories. Still, the ratings of swear words are 24% higher than the average score.

In contrast to the sometimes emotional conversations in sport subreddits, the conversations on subreddits which are categorized under 'technology' are filled with very few swear words. With a percentage of 0.2% the amount of swear words among those subreddits is half compared to the overall percentage and almost only a quarter compared to the categories 'sports' and 'comedy'. The usage of swear words in these technical and, thus, less colloquial conversations is much lower. However, the difference in Karma score between the overall average and comments containing swear words is high. With a ratio of 1:1,56, 'technology' has the second highest ratio. People who participate in the conversation on those subreddits use swear words less frequently but they seem to use them more effectively, when it comes to Karma score. From that it can be inferred that participants are very aware of their surroundings and more careful in the usage of swear words in those circumstances.

Although, there are categories which provide percentages that are very different compared to the overall result, most categories provide swear word occurrences that comply with the overall result. The dispersion of swear word percentages is quite small in the context of the chosen criteria. It does show, though, that the amount of swear words used does not automatically mean a better score. The Karma for swear words is better than the overall Karma but a correlation of higher swear word usage equals higher Karma cannot be made. On the contrary, swear words seem to have a much more positive effect on the score in conversations in which they are rarely used.

5. Discussion

How often are swear words used and how often are they used compared to spoken English and other online platforms?

Swear word usage on reddit is around 0.4% or four swear words in 1000 words submitted. In comparison, the swear word usage on reddit is double to that on MySpace. On the other hand, the difference between the two platforms is only 0.2% as MySpace provides an overall rate of 0.2%-0.3% (cf. Thelwall 2008: 93). Compared to the results by McEnery 2006, the swear word usage on reddit mimics the swear word usage in spoken English, when it comes to number of swear words used. McEnery recorded usages of 0.3% to 0.5% (cf. 2006: 45-49). Research by Mehl and Pennebaker returned a similar percentage in spoken English (cf. Mehl and Pennebaker 2003: 862-863). Compared to the results in spoken English, the percentage of swearing on the internet is even slightly lower with 0.2% for MySpace and 0.4% for reddit.

The results on reddit also verify the distribution of swear words. Results by Thelwall and by Gauthier et al. show that the most commonly used terms for swearing are *shit* and *fuck* both on twitter and on MySpace as Figure 2 and Figure 3 show.

Figure 2 Word distribution for MySpace

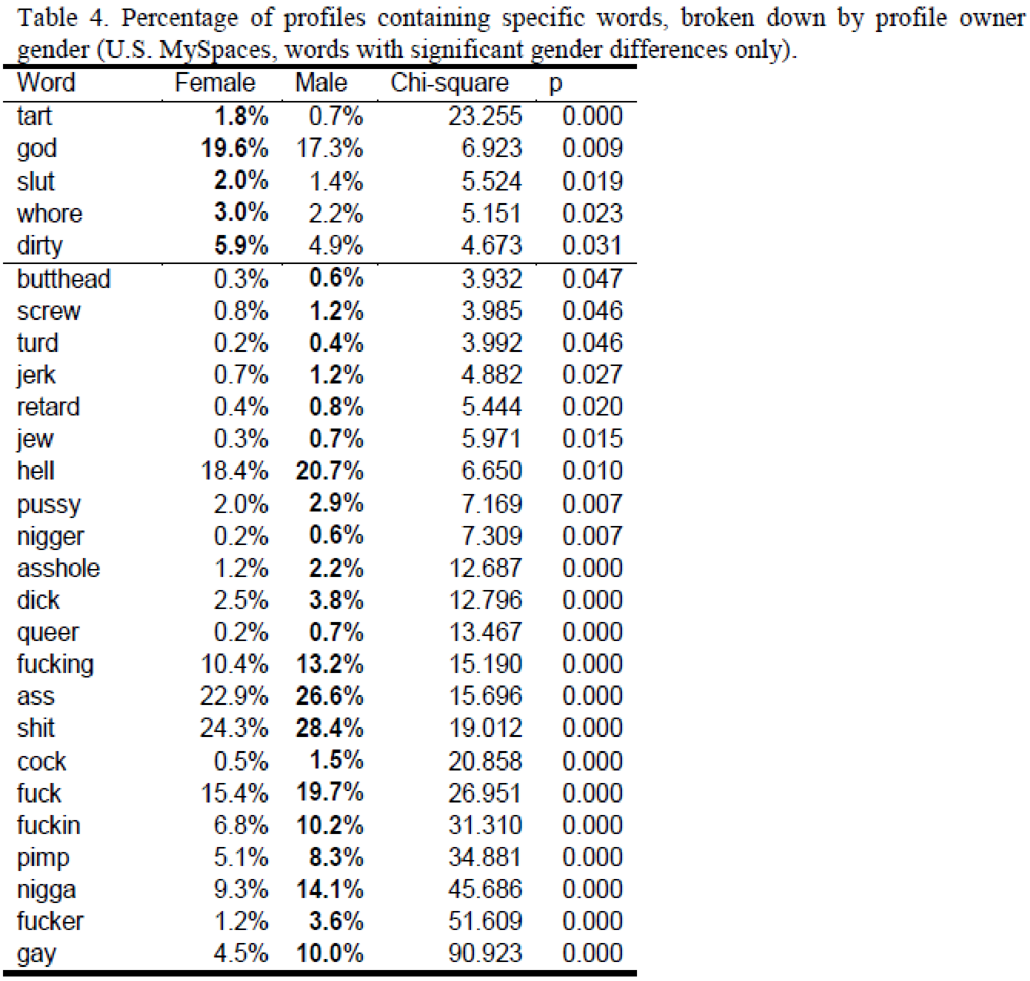


Table 4. Percentage of profiles containing specific words, broken down by profile owner gender (U.S. MySpaces, words with significant gender differences only).

Word	Female	Male	Chi-square	p
tart	**1.8%**	0.7%	23.255	0.000
god	**19.6%**	17.3%	6.923	0.009
slut	**2.0%**	1.4%	5.524	0.019
whore	**3.0%**	2.2%	5.151	0.023
dirty	**5.9%**	4.9%	4.673	0.031
butthead	0.3%	**0.6%**	3.932	0.047
screw	0.8%	**1.2%**	3.985	0.046
turd	0.2%	**0.4%**	3.992	0.046
jerk	0.7%	**1.2%**	4.882	0.027
retard	0.4%	**0.8%**	5.444	0.020
jew	0.3%	**0.7%**	5.971	0.015
hell	18.4%	**20.7%**	6.650	0.010
pussy	2.0%	**2.9%**	7.169	0.007
nigger	0.2%	**0.6%**	7.309	0.007
asshole	1.2%	**2.2%**	12.687	0.000
dick	2.5%	**3.8%**	12.796	0.000
queer	0.2%	**0.7%**	13.467	0.000
fucking	10.4%	**13.2%**	15.190	0.000
ass	22.9%	**26.6%**	15.696	0.000
shit	24.3%	**28.4%**	19.012	0.000
cock	0.5%	**1.5%**	20.858	0.000
fuck	15.4%	**19.7%**	26.951	0.000
fuckin	6.8%	**10.2%**	31.310	0.000
pimp	5.1%	**8.3%**	34.881	0.000
nigga	9.3%	**14.1%**	45.686	0.000
fucker	1.2%	**3.6%**	51.609	0.000
gay	4.5%	**10.0%**	90.923	0.000

(Thelwall 2008: 85)

If all the variations for *fuck* are combined, Thelwall's results show a percentage of 46.7% for the male population. This is higher than the results for reddit. It is comparatively even higher, as the number of terms that Thelwall recorded exceeds the number of terms that were taken from the reddit corpus. What is similar is that *fuck* and *shit* are among the most commonly used terms. As *god* could not be searched for as a swear word, comparative data to one of the most common swear words in Thelwall's study is missing from this study. Interestingly, the occurrences of *nigger* and *nigga* are much higher in the Twitter corpus than in the reddit corpus. The results for reddit show percentages of 0.07% and 0.19% respectively, whereas the results for Twitter are 0.6% and 14.1% - for male users. This is surprising as the assumption was made that the higher level of anonymity on reddit could lead to a higher usage of swear words and possibly to a higher use of very offensive terms, especially. On the other hand, conversations on reddit are lightly controlled by moderators. Offensive use that is too strong for a large part of the community can be forbidden and deleted. Also similar to Thelwall's findings is that *ass* and *asshole* are amongst the most used swear words. However, Thelwall records usages of up to 26.6% whereas the reddit corpus only shows a usage of 8%. The order in which these swear words are used is similar, the percentage of how much they make up of the overall corpus is different. What has to be noted here is that Thelwall looks at how many profiles contain a certain swear word, whereas my data comes from comments. So, the percentages for Thelwall are of all the profiles that contain swear words, which swear words were found.

Figure 3 Word distribution on Twitter

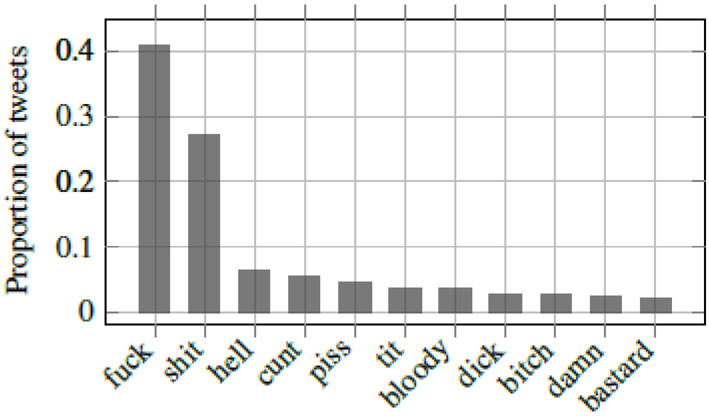


Figure 3. Most common swear words found in swearing tweets published by male users.

(Gauthier et al. 2015: 5)

Similarities can also be found by comparing the reddit corpus to the Twitter corpus by Gauthier et al. (2015). Their data might be even more comparable and representative as it is more recent. The two swear words being used most often are *fuck* and *shit* in the Twitter corpus. This mirrors the results for the reddit corpus and for the MySpace corpus. Differences can be found between *piss* and *cunt.* In the reddit corpus the former is much more common than the latter, whereas in the Twitter corpus their order is reversed. In the Twitter corpus, they are used at roughly the same quantity, whereas in the reddit corpus *piss* is used three times as often as *cunt*. The largest difference can be found by comparing *damn* between both corpora. On reddit *damn* is the third most often used swear word and on Twitter it is second to last. So, there are differences in usage between the corpora and the different online platforms. What is similar across all three of the findings is that *fuck* and *shit* are two most often used words. In the MySpace corpus, only the basic form 'shit' is mentioned, whereas all variants of 'shit' are recorded in the reddit corpus. That means that phrases like 'Are you shitting me?' and words like *shithead* contribute to the number on reddit but they do not for MySpace. Thus, the lower number for MySpace might be explained. However, as it stands, the variants of *shit* are used more often on reddit than in the other two corpora.

What might be the reason behind the fact that only two words account for over 70% of all swearing on reddit? One reason might be the different backgrounds users from these studies make up. Thelwall looked at MySpace profiles from the UK and the US and Gauthier looked at tweets that were sent from the same countries. On reddit, the user base stems from a broader variety of countries and cultures. Although, 72% of the user base come from countries were English is the native language (cf. Alexa.com 2017), the rest is international. For those users, English is the second language. Thus, their vocabulary is limited compared to native speakers. Of course, there might be MySpace and Twitter users whose second language is English but the corpora for those platforms were based on the IP addresses from which users accessed the site. Thus, most of them are residents of either the US or the UK. The statistics for reddit, which are provided by Alexa.com, are assumed to use the same method. The percentage of people for whom English is not the first language on reddit is, therfore, higher, according the statistics from Alexa.com. With a limited vocabulary, the use of the most prominent words is more likely. Especially, for users whose main connection to the English language is reddit or the internet in general. As they are mainly exposed the *shit* and *fuck,* they are likely to be most proficient in the use of those words. This is, however, a hypothesis that needs to be investigated further. The argument here is also circular as the high amount of occurrences of *shit* and *fuck* is supposed to increase their number of occurrences. A diachronic investigation into this matter would provide more insights.

To summarize, both distribution and usage of swear words in online discourse are generally similar across the monitored platforms and only differ in details. It seems, therefore, that the impact of the platform on the amount of swear words used is minimal. This can be said for the three discussed platforms at least. However, online discourse is still young compared to conversation customs of spoken language. Different settings on different platforms might influence the language used in the long run. In order to determine such claims, further platforms have to be investigated. Also, diachronic studies might provide new insides. Compared to spoken English, the usage of swear words is below 1% as well. In that regard, computer mediated discourse (CMD) seems to resemble spoken English rather than formal written English.

The fact that the overall occurrences for swearing are not higher than those in spoken language is also helped by the fact that certain types of swearing are non-existent in online discourse. Non-propositional swearing as defined by Jay and Janschewitz (2008) is unplanned and unintentional. As the form of discourse on reddit is asynchronous. Users do not have to be logged in at the same time, a response with a larger time difference between initial comment and responding comment is possible. This means that unintentional swearing is ruled out. People have time to think and formulate their answers. Even in synchronous electronic environments where people have to be logged in at the same time, i.e. chat rooms, writing an answer takes more time than saying one out loud. Still, in chat rooms a quick answer is much more important than in the CMD on reddit. Responses to pain or surprise that are unintentional are unlikely to be found on reddit and similar online platforms. Furthermore, in contrast to spoken language, users who are offended can choose to not participate in the conversation any further. The zero response is viable option. Discontent with the contents of a comment or the vocabulary that was used can be expressed by not further taking part in the discussion. This is, as reddit offers public discussions with millions of users, most likely not observable to most users. Thus, as part of the potential uses of swear words falls flat, a slightly lower number of swear words can be explained.

On the basis that a higher level of anonymity is one defining feature for reddit users, does anonymity seem to influence swear word usage?

Based on the findings in this study, it is safe to say that anonymity alone does not boost swear word usage significantly. This is true at least for reddit. Compared to Twitter and MySpace the overall swear word usage is not significantly higher or lower. On MySpace and Twitter, information can be gathered on gender and age. However, this is not publically visible on Twitter. But both platforms require profile pictures. Unless those information is published on reddit by the individual user in a comment or a post, this information is not publically available and not part of the dataset for this study. The only thing other users know about a particular user is their nickname, their comment and post history and the Karma they have gathered. People on reddit are more anonymous than they are on the other two platforms. Still, swear word usage does not increase. Other factors do have a greater influence in the amount of swear words used. Repercussions for swear word usage are also possible on reddit. Although, every user is mostly anonymous, misbehavior as judged by the community can result in a ban from the subreddit or from reddit altogether. This means that users who want to remain on the site and continue to conribute to the discussions are inclined to not step out of community guide lines. This can, but does not have to, influence swear word usage. Users are asked to "adhere to the same standards of behavior online that you follow in real life" and to "remember to be human" (Reddit 2017b). The latter rule asks users to consider posting something that they would not say to someone else in real life. Blatantly ignoring these rules could lead to an sequestration from the discussion. However, how much of an influence these rules have on the typical user behavior can only be guessed. Especially, since the user has to search for those rules after signing up initially. So, a discourse in which insults are rare is promoted by reddit itself. On the other hand, users who are excluded from reddit can sign up again very easily. If one's desire is to cause outrage, this can be done even after being banned. Creating a new account only requires an e-Mail account.

As already mentioned, the swear word usage on reddit is generally rated positively. This lead to the conclusion that comments which contain swear words express emotions and opinions that are shared by the community. The use of swear words in humorous comments seems also very likely. This leads to another conclusion, namely that the reddit community appears to be fairly homogenous in terms of opinion. Comments that contain opinions that are shared by the community can be upvoted. Swear word can be used to emphasize certain points or as an enhancer for opinions. Opinions that were uttered on reddit and that were against the general attitude within that community would be downvoted. If swear words are used to emphasize opinions and if their general ratings are above average, the conclusion is that the opinions that are uttered and emphasized using swear words are generally in line with the general opinion within that community. There will be cases which prove the opposite but the overall positive ratings for swear words support this conclusion. It has to be noted that this conclusion or assumption is only based on the data that is presented in this study. To verify or deny such assumptions, a closer inspection of contexts in which swear words and opinions are expressed would have to be made. Swear word usage on reddit appears to be used as a form to employ humor or to show solidarity with the community. These functions are connected to swear word usage by Dynel (cf. 2012: 40-41). Dynel claims that solidarity building is not relevant in online discourse, however. The result of the present study contradict that statement.

Lastly, Herring (2001) explains that online groups develop norms of practice. These norms determine what is acceptable and what is not (cf. ibid.: 622). Users on reddit seem to be very aware of what is acceptable and what is not. This explains that swear words, which regardless of intention have the ability to cause offense, are rated higher than the average comment. The discourse norms, thus, do not prohibit swear word usage and users are capable of judging situations where swear word usage is allowed correctly. Otherwise, the average score for comments with swear words would be lower than the average comment score. The large part of users who are not native English speakers seem to be aware of those circumstances as well. The difference between the overall average comment score and the score for swear words is quite substantial. Although, 70% of reddit users come from English speaking countries, mishaps by the remaining 30% of the user base should be somewhat visible. As they are not, the communicational norms of reddit seem to be generally understood by all users, regardless of nationality. The results of this study contradict the notion that swear word usage is largely boosted by the level of anonymity. Other factors appear to have a much greater influence on usage and ratings of swear words.

Is the swear word usage rated differently, depending on the semantic field the swear word originates from?

If the swear words from this study are ordered according to their average rating, the following order is established:

Table 10 Swear word ordered by rating

Word	Average Score
nigga	29,23
cock	24,57
dick	22,43
bitch	19,54
piss	19,52
nigger	18,84
fuck	18,57
bastard	18,52
ass	17,65
damn	15,85
cunt	15,13
shit	12,33
retard	10,79
fag	10,65

In general, the results show that swear words that originate from the taboo topic *genitals* are achieve higher scores than those that come from *sexual acts* and *excrements*. The terms *cock* and *dick* are rated second and third highest. The word *cunt* is on the bottom of the scale, though. It is also used less than the other two terms. *Piss* is rated fifth highest whereas *shit* is rated third lowest. The taboo topic of excrements is rated comparatively low. Both racial slurs are rated fairly highly. This is surprising as race is usually regarded as a highly offensive topic and very taboo. However, their amount of occurrences is very low, so the use in inappropriate contexts appears to be very low. The swear words whose taboo originates from sexuality and disease are rated the lowest. *Retard* and *fag* have the lowest scores among the swear words.

To answer the question, the taboo topic does appear to have an influence on the Karma they receive in certain cases. Certain topics like disease and sexuality are rated much lower than other topics. However, for those topics there is also only one word each in the selected swear words. Thus, the rating of a single word influences the rating of the whole group. This is not representative for a whole topic from which insults and swear words can come from. Thus, a general conclusion cannot be made.

Based on previous offensiveness ratings, are more offensive words rated differently?

The order shows that the potentially highly offensive term *nigga* is rated the highest. As already discussed, this supports the claim that it is used as a non-offensive term in the black community. It is also supported by the fact that it is used most often in the subreddit r/BlackPeopleTwitter.

The Ofcom classification of offensiveness reflects that *shit, arse* and *arsehole* are considered milder or medium words. This gives them a broader context in which they could be used without causing offense hypothetically. The lesser the degree of offensiveness, the harder it would be for a term to cause offense in the audience. However, all instances of *shit* are rated below the swear word average. All the variants of *ass,* which includes *asshole,* only come lowest on the list. In contrast, stronger words are rated higher. *Cock* and *dick* are the second and third highest rated terms. *Fuck* is also above *ass* with an overall higher amount of occurrences. From the Ofcom list only *cunt* is rated lower than the overall swear word average of 16,29. The other terms that are categorized in the "strongest words" column by Ofcom (cf. 2016: 44) are above the average. There is no differentiation made between *fuck* and *motherfucker* in the findings for reddit. So, a possibly different rating cannot be established. Generally, the assumption that less offensive words are rated higher than their more offensive counterparts is not true.

As previously mentioned, the offensiveness ratings are not definitive ratings. These ratings are usually established without using the swear words in a specific context. As the results show, the appreciation of the swear words on reddit is high. This would either point to very competent communicators that know very well when and when not to use an offensive term or that offensiveness is generally appreciated. It could also be that a higher rate of offensiveness enhances the rest of the comment more or increasing the humorous level of the comment more than the less offensive swear words. The ratings lead to the assumption that opinions on reddit are shared within a community. If a comment is made funnier or more compliant with the general attitude of the community by using more offensive terms, it seem logical that it receives a higher score. Furthermore, as other means to convey opinion or being humorous such as intonation, articulation, tempo and volume are all non-existent in CMD, other means have to be used. Thus, the experienced emotions have to be conveyed just by words. For that more offensive terms seem to be more fitting.

How does swear word usage and reception differ across different subreddits?

In the results it was established that the use of swear words across the different categories differs in select cases. This means that the overarching topic of the discussion has an influence on the amount of swear words used. Especially, in discussion about technological topics the use of swear words is very limited and with 0.2% only half of the overall average. In topics that feature either more humor or more emotional conversations such as comedy and sports, the use of swear words is almost double the overall average with 0.7% and almost four times as high as the discussions about technology. The category 'conversation' has an average amount of swear words used. This might be explained by the fact that conversational subreddits can feature discussions about many different subjects. In the subreddit r/AskReddit users can ask other users almost everything. Thus, a more average use of swear words can be expected.

The results for swear word usage differs even more if individual subreddits are taken into account. From the thirty five subreddits that make up the results of table 9, the one with the highest usage of swear words is r/CringeAnarchy, which is from the 'comedy' category. The subreddits with the lowest percentage of swear words is r/buildapc from the 'technology' subreddit. The percentage for r/CringeAnarchy is 1.19%. This means that per 1000 words uttered 12 of them are swear words. This is three times as high as the general average. The percentage for r/buildapc is 0.07%. The overall swear word usage is 0.4% and five times as high as the usage on r/buildapc. That means that swear words are used 17 times more often in discussions on r/CringeAnarchy than they are on r/buildapc. The differences between the subreddits are, therefore, substantial. Usage differs quite a lot depending on the subject of the discussion. Comparing this to everyday discussions this seems natural. Discussions about and advice on how to build a PC offer fewer situations in which the use of a swear word is appropriate. Especially, since those discussions are public. Humor on the other hand is proven to be a topic in which the use of swear words is not only accepted. Interestingly, the difference in Karma between the average score for all comments and those containing swear words is very small for r/CringeAnarchy with a ration of 1 to 1,1. The swear words used on that subreddit do not cause the score to increase. In contrast, the ratio of average score for all comments to comments containing swear words on r/buildapc is 1 to 2,1. It has to be mentioned that the average score on r/buildapc is 2,39 and for swear words it is 5,12. The overall ratings on that subreddit are fairly low. For comparison, for r/CringeAnarchy the average score is 9,8 and for swear words it is 10,8. It has to be noted that within the community guidelines for the subreddit r/CringeAnarchy the following rule can be found:

> "**Don't be a faggot.** If you want to make dramatic selfposts about "bullying", preach social justice topics or white knight for m'ladies, you belong in the original cringe subs."
>
> (r/CringeAnarchy 2017)

Swear word usage is allowed and a politically correct attitude is explicitly prohibited. Obviously, this influences the discussions that posts on that subreddit gather.

So, the topic of a subreddit influences the amount of swear words used as well as the rating. However, it is not the case that swear words are rated much higher than the average comment in subreddits where there is a lot of swearing. For the selected subreddits the opposite seems true. Rather than the format of the platform influencing swear word usage and score, the topic and conversational norms of each subreddit influence those statistics.

6. Conclusive remarks and limitations

6.1. Conclusion

This thesis was designed to show an insight into the swear word usage on the online website reddit. Conversation on the internet is still relatively young, thus, new conversational habits could develop. Therefore, an investigation into the swear word usage on a platform for which such data did not exist previously was deemed fruitful for the scientific discourse. The results show that the overall swear word usage is comparable to those found on other online platforms and to those in spoken English. As a main difference between the studied online platforms the level of anonymity was made out. This does not affect the overall usage of swear words. It is rather the topic of a particular discussion and the conversational norms of the subreddits that influence both usage and reception of swear words. A subset of abbreviations was found to show significant differences in usage between the individual terms. The same can be said about the written out words. In general, written out words are used more often than abbreviations, which is partly due to a broader set of contexts that the written out words have. Two words form more than 70% of all swear words used in the corpus.

As a second measure, the reddit-specific scoring system, or Karma, was used in order to determine how swear words are received by the community. These ratings show that swear words have a positive effect on the reception of a comment. This leads to the conclusion that users on that platform are very aware on when, where and how swear words can be used acceptably. It further suggests, that swear word usage is mainly done not to offend and attack other users. In general, swear words on reddit are not used impolitely. This leads to the assumption that swear words on reddit are more often used to express emotions that are shared by larger parts of the community and to convey humor. Both these uses explain the overall higher ratings for swear words. A general rejection of swear words cannot determined for reddit users. Measures such as censorship of single words, like it is done in other mass media, is not required or carried out. On the contrary, the effect of swear words is rather positive.

As an addition to previous findings which found significant differences between the genders and age groups when it comes to swear word usage, the findings in this study add that the topic of a conversation influences the usage and reception of swear words. The conversational norms of the sub-communities are also greatly influential. These two factors are more influential on the overall ratings and usage of swear words than the base platform and its level of anonymity.

6.2. Limitations and further research

The statements made above have to be put into the context of the limitations of this study. The positive effect of swear words on the comment's rating is not the only factor that influences the score of a comment. Other influencing factors such as timing and, most importantly, the actual content of the comments are not recorded. It is safe to assume that the content of a comment is more influential on the score than whether or not swear words were used. Also, the method for acquiring the data does not allow to search for terms that are made swear words by their context. This affects mostly swear words like *oh my god!* and *Go to hell!* whose taboo originates from the taboo of religious terms. Swearing which uses such terms could not be recorded. Furthermore, the act of selecting a set of swear words before looking into the corpus influences the overall percentage of swear words used. Although, the results suggest that most of the swearing is recorded, as the results are similar to other studies concerning swear words, not all swearing could be processed. Additionally, the method of data acquisition also did not allow for the search for particular phrases. Thus, statements could only be made about the general word. Different variations of those words were recorded, but did not provide any significant results. Thus, they were added into the score and count of the base word. Also, a search for euphemisms was not possible as the results did not provide an insight into the contexts in which the words were used. Contexts could only be established by the topic or theme of individual subreddits.

Another limiting factor that concerns the results of the Karma is the fact that discussion on reddit and other similar platforms are moderated. Moderators have the ability to block or ban users that to not comply to the guidelines for each subreddit. Furthermore, they have the power to delete comments. This, obviously, affects the usage of swear words and utterance of offensive comments.

A few assumptions were made in this study. To further deny or verify them, further research would have to be carried out. The data that can be gathered via BigQuery could provide such data sources. Smaller corpora could be compiled to look at a set of individual comments to research Karma reception an swear word usage further. As previously discussed, the use of euphemisms is motivated by "the desire to make a positive impression on the external audience" (McGlone and Batchelor 2003: 260). A study of the use of euphemisms on reddit could provide interesting new insights into language use online. The need for euphemisms to avoid swear words does not appear to be there. Swear words are responded to very positively. To explore their use in online discourse is only one example which the reddit data could provide new insights to.

References

Alexa.com (2017), " reddit.com Traffic Statistics". <http://www.alexa.com/siteinfo/reddit.com> (13.06.2017).

Allan, Keith & Kate Burridge (1991), *Euphemism and Dysphemism, Language Used as Shield and Weapon*. Oxford-New York: Oxford University Press.

Allan, Keith & Kate Burridge (2006), *Forbidden words: taboo and the censoring of language*. Cambridge: Cambridge Univ. Press.

Allestree, Richard (1719), *The Whole Duty of Man, Laid down In a Plain and Familiar Way, For the Use of All, but especially the Meanest Reader. Divided into XVII Chapters; One whereof being read every Lord's-Day, the Whole may be read over Thrice in the Year. Necessary for all Families. With Private Devotions for several Occasions.* London: Roger Norton.

Bignell, Jonathan & Jeremy Orlebar (2005), *The Television Handbook.* London: Routledge.

BigQuery (2016), " fh-bigquery:reddit_comments.2006" <https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2006?pli=1&tab=schema> (25.06.2017)

BigQuery (2017a), "Migrating to Standard SQL". <https://www.google.de/search?q=BigQuery+2017+https%3A%2F%2Fcloud.google.com%2Fbigquery%2Fdocs%2Freference%2Fstandard-sql%2Fmigrating-from-legacy-sql&gws_rd=ssl> (08.07.2017)

BigQuery (2017b), " fh-bigquery:reddit_comments.2017_02". <https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2017_02?pli=1&tab=details> (25.06.2017)

Bowers, Jeffrey. S & Christopher W. Pleydell-Pearce (2011), "Swearing, Euphemisms, and Linguistic Relativity", *PLoS ONE* 6:7*,*1-8.

Brown, Penelope % Stephen C. Levinson (1987), *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press. [First published 1978 as part of Esther N. Goody (ed.): Questions and Politeness. Cambridge University Press], 311-323.

Brown, Penelope & Stephen C. Levinson (2006), "Politeness: Some universals in language usage", in Adam Jaworski & Nikolas Coupland (eds.) *The Discourse Reader.* London: Routledge.

Choudhury, Munmun De & Sushovan De (2014), "Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity", *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media,* Palo Alto, California: The AAAI Press, 71-80.

Culpeper, Jonatha (1996), "Towards an anatomy of impoliteness", *Journal of Pragmatics* 25, 249-367.

Culpeper, Jonathan (2011), *Impoliteness: Using Language to Cause Offence*, Cambridge: Cambridge University Press.

Culpeper, Jonathan, Derek Bousfield & Anne Wichmann (2002), " Impoliteness revisited: with special reference to dynamic and prosodic aspects", *Journal of Pragmatics* 35, 1545-1579.

Daly, Nicola, Janet Holmes, Jonathan Newton and Maria Stubbe (2004), "Expletives as solidarity signals in FTAs on the factory floor", *Journal of Pragmatics* 36, 945-964.

Davis, Hayley (1989), "What makes bad language bad?", *Language & Communication 9*:1, 1-9.

Dery, Mark (1994), *Flame Wars: The Discourse of Cyberculture.* Durham, NC: Duke University Press.

Dewaele, Jean-Marc (2004a), "The Emotional Force of Swearwords and Taboo Words in the Speech of Multilinguals", *Journal of Multilingual and Multicultural Development* 25:2-3, 204-222

Dewaele, Jean-Marc (2004b), " Blistering barnacles! What language do multilinguals swear in?!", *Estudios de Sociolingüística* 5:1, 83-105.

Domínguez, Pedro J. Chamizo (2005) "Some theses on euphemisms and dysphemisms". *Studia Anglica Resoviensa* 25, 9–16

Dynel, Marta (2012), " Swearing methodologically: The (Im)politeness of expletives in anonymous commentaries on Youtube", *Journal of English Studies* 10, 25-50.

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

Fairman, Christopher M (2007), "Fuck", *Cardozo Law Review* 4, 1711-1772.

Fernández, Eliecer Crespo (2008), " Sex-Related Euphemism and Dysphemism: An Analysis in Terms of Conceptual Metaphor Theory", *Atlantis* 30:2, 95-110.

Gauthier, Michael, Adrien Guille, Anthony Deseille & Fabien Rico (2015), "Text Mining and Twitter to Analyze British Swearing Habits", in C. Levallois, M. Marchand, T. Mata and A. Panisson (eds) *Handbook of Twitter for Research*. Lyon: EMLYON, 27-43.

Goffman, Erving (1967), *Interaction Ritual: Essays on Face-to-Face Behavior*. New York: Doubleday.

Herring, Susan (2001), "Computer-mediated Discourse", in D. Schiffrin, D. Tannen and H. Hamilton (eds) *The Handbook of Discourse Analysis*. Oxford: Blackwell, 612–34.

Holquist, Michael (1994), "Introduction: Corrupt Originals: The Paradox of Censorship", *PMLA* 109:1, 14-25.

Jay, Timothy (1980), "A frequency count of college and elementary school students' colloquial English", *Catalogue of Selected Documents in Psychology* 10, 1.

Jay, Timothy (1992), *Cursing in America*. Philadelphia: John Benjamins.

Jay, Timothy (2000), *Why we curse*. Philadelphia: John Benjamins.

Jay, Timothy (2009), "The Utility and Ubiquity of Taboo Words", *Perspectives on Psychological Science 4*:2, 153-161.

Jay, Timothy & Kristin Janschewitz (2008), "The pragmatics of swearing", *Journal of Politeness Research 4*, 267-288.

Jennings, Gary (1967), *Personalities of Language.* London: Gollancz.

Kaye, Barbara K., & Barry S. Sapolsky (2009), "Taboo or Not Taboo? That is the Question: Offensive Language on Prime-Time Broadcast and Cable Programming", *Journal of Broadcasting & Electronic Media 53*:1, 1-16.

Krahulik, Mike & Jerry Holkins. "Green Blackboards (And Other Anomalies)", <https://www.penny-arcade.com/comic/2004/03/19/green-blackboards-and-other-anomalies> (13. May 2017).

Labov, William (1972), *Language in the inner city: Studies in the black English vernacular.* Oxford: Blackwell.

Lakoff, Robin (1989), "The limits of politeness", *Multilingua* 8: 101-129.

Leech, Geoffrey (1983), *Principles of pragmatics*. London: Longman.

Ljung, Magnus (2011), *Swearing: a cross-cultural linguistic study*. Basingstoke: Palgrave Macmillan.

Mann, Chris and Fiona Stewart (2000), *Internet Communication and Qualitative Research: A Handbook for Researching Online*. London: Sage

McEnery, Tony (2006), *Swearing in English: bad language, purity and power from 1586 to the present*. London: Routledge.

Mcglone, Matthew S. & Jennifer Batchelor (2003), "Looking Out for Number One: Euphemism and Face" *Journal of Communication* 53:2, 251-264.

McGlone, Matthew S., Gary Beck & Abigail Pfiester (2006), "Contamination and Camouflage in Euphemisms", *Communication Monographs* 73:3, 261-282.

Mehl, Matthias R. & James W. Pennebaker (2003), " The Sounds of Social Life: A Psychometric Analysis of Students' Daily Social Environments and Natural Conversations", *Journal of Personality and Social Psychology* 48:4, 857-870.

Mehl, Matthias R., Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher, James W. Pennebaker (2007), "Are Women Really More Talkative Than Men?" *Science* 317, 82.

Montagu, Ashley (1967), *The Anatomy of Swearing,* London: Macmillan.

Ofcom (2016), *Attitudes to potentially offensive language and gestures on TV and radio.*

Ofcom.com (2017, March 30), "What is Ofcom?" < https://www.ofcom.org.uk/about-ofcom/what-is-ofcom> (25. May 2017),

Pew Research Center (2013), " 6% of Online Adults are reddit Users". <http://www.pewinternet.org/2013/07/03/6-of-online-adults-are-reddit-users> (15.06.2017).

Pinker, Steven (2007), *The Stuff of Thought : Language as a Window Into Human Nature*. New York, NY: Viking.

Porter, Roy (1991) *English Society in the Eighteenth Century,* London: Penguin.

r/CringeAnarchy (2017) "sub Reddit for alt right trolls" <https://www.reddit.com/r/CringeAnarchy/> (10.07.2017).

Reddit.com (2014), " Some badass motherfucker in the past domesticated a fucking wolf. • r/Showerthoughts" <https://www.reddit.com/r/Showerthoughts/comments/264k8p/some_badass_motherfucker_in_the_past_domesticated> (15.06.2017).

Reddit.com (2017a), " u/_vargas_" <https://www.reddit.com/user/_vargas_> (15.06.2017).

Reddit.com (2017b), "Reddiquette". <https://www.reddit.com/wiki/reddiquette> (10.07.2017)

Santaemilia, José (2008), "The Translation of Sex-Related Language: The Danger(s) of Self-Censorship(s)", *TTR* 212, 221–252.

Statista.com (2017), " Reddit.com: unique visitors 2017". <https://www.statista.com/statistics/443332/reddit-monthly-visitors> (15.06.2017).

Stephens, Richard, John Atkins & Andrew Kingston (2009), "Swearing as a response to pain", *NeuroReport 20*, 1056-1060.

Thelwall, Mike (2008), "Fk yea i swear: Cursing and gender in a corpus of MySpace pages", *Corpora* 3(1), 83–107.

Webster Dictionary Online, "Swear words" < https://www.merriam-webster.com/dictionary/swear%20words> (13. May 2017)

Weninger, Tim, Xihao Avi Zhu & Jiawei Han (2013), "An exploration of discussion threads in social news sites: A case study of the Reddit community", *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, Niagara Falls, ON: IEEE. 579-583.

Werry, Christopher C. (1996), "Linguistic and Interactional features of Internet Relay Chat" in Susan C. Herring (ed.), *Computer-Mediated Communication: Linguistic, social, and cross-cultural perspectives.* Amsterdam: Benjamins. 47-64.

Witschge, Tamara (2008), "Examining online public discourse in context: A mixed method approach", *Javnost-The public*, 15:2, 75-91.

Youtube (2017), "Jan Böhmermann Is the Seth Meyers of Late Night German TV". <https://www.youtube.com/watch?v=97XSh2t5weI> (20. May 2017)

Appendix

(1) List of all 99 subreddits:

FIFA
FireEmblemHeroes
Fitness
canada
2007scape
AdviceAnimals
Amd
Android
anime
AskMen
AskOuija
AskReddit
AskWomen
australia
aww
baseball
BlackPeopleTwitter
buildapc
cars
CFB
ClashRoyale
CollegeBasketball
conspiracy
CringeAnarchy
dankmemes
DBZDokkanBattle
DestinyTheGame
DotA2
EnoughTrumpSpam
europe
explainlikeimfive
FFBraveExvius
ffxiv

pcmasterrace
personalfinance
Philippines
Jokes
Futurology
Games
gaming
gifs
GlobalOffensive
GlobalOffensiveTrade
gonewild
hearthstone
heroesofthestorm
hiphopheads
hockey
IAmA
india
jailbreak
leagueoflegends
LifeProTips
magicTCG
mildlyinteresting
MMA
movies
Music
nba
news
nfl
NintendoSwitch
nottheonion
OkCupid
Overwatch
pathofexile

forhonor
funny
pics
pokemongo
pokemontrades
politics
PS4
Rainbow6
relationships
RocketLeague
RocketLeagueExchange
rupaulsdragrace
Showerthoughts
Smite
Sneakers
soccer
space
SquaredCircle
technology
teenagers
television
The_Donald
TheSilphRoad
todayilearned
trees
TwoXChromosomes
ukpolitics
unitedkingdom
videos
worldnews
wow
WTF
xboxone

(2) Table with subreddit Categories, Word Count, SW Count, percentage and ratio

Category	Subreddit	Word Count	Average Score	SW count	SW Score	SW per Sub in %	ratio	SW %
memes comedy	CringeAnarchy	2307495	9,8	27553	10,8	1,19	1:1,1	0,7
	AdviceAnimals	5822563	12,45	31463	12,28	0,54	1:1	
	funny	9399116	14,36	63612	15,25	0,68	1:1,1	
	dankmemes	1653656	6,97	16771	8,59	1,01	1:1,2	
	jokes	1709532	16,15	11059	21,16	0,65	1:1,3	
conversation	AskReddit	119196074	17,27	584099	29,77	0,49	1:1,7	0,4
	AskMen	4505193	8,83	20627	12,78	0,46	1:1,4	
	AskWomen	4712424	8,67	10760	12,93	0,23	1:1,5	
	explainitlikeim5	8181162	7,83	9362	18,01	0,11	1:2,3	
	relationships	16366194	14,91	35247	22,49	0,22	1:1,5	
Technology	technology	5725382	11,46	18084	13,72	0,32	1:1,2	0,2
	buildapc	6335094	2,39	4733	5,12	0,07	1:2,1	
	AMD	3349576	4,65	5173	5,71	0,15	1:1,2	
	Android	3196471	7,54	7009	9,55	0,22	1:1,3	
	pcmasterrace	9294859	5,49	25862	10,91	0,28	1:2	
news politics	politics	72714711	10,93	282406	15,83	0,39	1:1,4	0,5
	news	26875576	10,94	130070	14,78	0,48	1:1,4	
	worldnews	30467356	11,29	119646	15,39	0,39	1:1,4	
	The_Donald	33710481	10,02	237334	13,06	0,7	1:1,3	
	europe	7285543	7,5	19462	8,88	0,27	1:1,2	
media	videos	13505147	14,68	73262	16,84	0,54	1:1,1	0,5
	pics	13992774	13,7	75410	17,97	0,54	1:1,3	
	gifs	6521913	17,66	42733	19,13	0,66	1:1,1	
	movies	10610119	13,91	48870	16,25	0,46	1:1,2	
	anime	8171402	7,81	29135	8,89	0,36	1:1,1	
sport	nba	16712204	12,37	128452	15,99	0,77	1:1,3	0,7
	nfl	11217407	12,12	69824	15,31	0,62	1:1,3	
	soccer	8685167	11,89	54917	13,73	0,63	1:1,2	
	hockey	7715944	6,63	66504	7,69	0,86	1:1,2	
	squadcircle	10555143	8,59	63068	10,55	0,6	1:1,2	
games	forhonor	13481192	3,94	64046	4,58	0,48	1:1,2	0,4

	overwatch	12808815	11,18	45534	10,85	0,36	1:1	
	nintendoswitch	12662559	5,15	23610	6,89	0,19	1:1,3	
	leagueoflegends	14214239	5,45	60641	8,85	0,43	1:1,6	
	gaming	8475484	12,53	46207	15,78	0,55	1:1,3	

(SW = swear word. Ratio compares the average score overall with the average score for the swear words per subreddit.)

Erklärung

Ich bestätige hiermit, dass ich von der Plagiatsregelung am Institut für Anglistik/Amerikanistik Kenntnis genommen habe und durch die Teilnahme an diesem Seminar diese ausdrücklich anerkenne.

Datum Unterschrift