Text Mining and Twitter to Analyze British Swearing Habits

Michael Gauthier
CRTT, University of Lyon 2
michael.gauthier@univ-lyon2.fr

Adrien Guille
ERIC, University of Lyon 2
adrien.guille@univ-lyon2.fr

Fabien Rico
ERIC, University of Lyon 1

Anthony Deseille
University of Lyon 1

The way women and men speak and are expected to behave is frequently discussed. For example, women are sometimes described as speaking more than men, and men as swearing more than women. These stereotypes can alter people's expectations concerning the way we should behave. Indeed, if the idea that females generally swear less frequently than males is widespread, women who swear may be perceived as deviant from the norm, and thus be stigmatized. Clearly understanding what is true and what is not in these studies and reports is not an easy task, because there is a considerable amount of differing opinions on the topic. The way swear words are used by women and men is one of those topics which remains vague, but whose stake is great, since swearing is often considered as an act of power and a way of affirming oneself.

This article will introduce the data gathered from a corpus of tweets in order to shed a new light on new ways of analyzing specific sociolinguistic features like gendered uses of swear words on Twitter. Analyzing the linguistic behaviour of users of these media can be an interesting way of generating a most contemporary corpus representative of general trends, and computational linguistics can represent a very accurate and powerful method of analyzing the different uses people can make of certain speech patterns. In order to carry out the study, we used several tools taken from both computer science and linguistics. These tools may represent innovative methods to analyze the effect of social parameters on speech patterns displayed in Twitter corpora.

Thanks to this data, we analyze both quantitative, and qualitative instances of swear words in the corpus, to see how the linguistic gendered preferences may differ when swearing is used, but just as importantly, we see how comparable they can be. Indeed, very often when dealing with gender in corpus linguistics, small differences tend to be focused on, whereas they are actually minor compared to the similarities. As for every study, the methods used here also have certain limits that we present as well.

Without pretending to be representative of interactions other than the computer-mediated ones present in this corpus, we hope that this data can shed an up-to-date and neutral light on the way women and men use swear words on Twitter, and on the implications these results may have, as well as on new tools researchers can use in various areas of research. We believe that this study can also be useful to computational linguists/sociologists thanks to the methods used to access data not directly available and displayed by users (e.g. the age or the sex).

Introduction

The way women and men speak is a common source of discussions and debates, be it in academic research or in mainstream media. Many social attitudes and linguistic features have generally been attributed either to women or men; swearing is one of those topics traditionally associated with men. As Coates (2004) reported, "*the folklinguistic belief that men swear more than women and use more taboo words is widespread*", consequently leading to the creation of stigmas preventing women or men from using a certain type of linguistic behavior without being stigmatized. These preconceived ideas also fuel societal stereotypes and may impact people's standards concerning what is desirable from each gender. Moreover, swearing is often considered as an act of power (see Beers Fägersten (2012); Murray (2012); Lakoff (2004); G. I. Hughes (2006)) and a way of affirming oneself. Thus, the fact that one gender may be perceived as more frequent users of swear words, or on the other hand as swear words eschewers, may have an impact on other qualities related to power we would inherently attribute to one gender or the other, whether or not these differences are real. A certain number of studies have shown that contrary to what has long been widely believed, women do not swear less than men, nor do they use a drastically different register (Baruch & Jenkins, 2007; Hammons, 2012; S. E. Hughes, 1992; Jay, 1992; Baker, 2014; Thelwall, 2008; Coates, 2004). Indeed, some of these surveys have shown that what generally differs between women's and men's use of swear words is not the rate at which they are used, but the context in which they are used, as well as the kinds of words women and men use. Some studies seem to indicate that the use of "strong" swear words (see below for a description of what "strong" swearing is) by women is increasing in certain contexts, and especially on social media (Murray, 2012; Thelwall, 2008), and in the United Kingdom. Thelwall predicted that "*a reversal in gender patterns for strong swearing will slowly become more widespread, at least in social network sites*", and this seemed especially true for younger generations of users (users aged 16-19 in the case of Thelwall). Without pretending that our article would answer this question, this is a good example to illustrate how social media can be one way of highlighting new linguistic features, and try to better understand them. The importance and time we devote to social media sites is growing every year according to a study from Ofcom (see the 2013 Ofcom report), and it concerns people from all age groups, and all socioeconomic backgrounds (Smith & Brewer, 2012). It would also seem that on these media, and especially on Twitter, people tend to swear more than in face to face interactions (Wenbo et al., 2014). This paper will introduce the data gathered from a sample corpus of about one million tweets from about 18,000 users collected through Twitter in order to see how data such as age and gender can be put in relation with more (socio)linguistic-related features on Twitter, and how these features can be analyzed. Here, the goal of this pilot study is more experimental and demonstrative, as we will try to suggest innovative ways of empirically analyzing gender and language, rather than give definitive answers. In order to carry out our investigation, we collected tweets localized in the UK, then we analyzed the corpus with various tools (event detection programs, lemmatizers, named-entity recognition etc). Information concerning the users (sex, age and location) was retrieved so that we could draw statistics and study different social, contextual and linguistic uses of swear words.

Thanks to this data, we will analyze both quantitative, and qualitative instances of swear words in the corpus, to see how linguistic gendered preferences may differ when we swear, but just as importantly, we will see how comparable they can be. As Baker (2014) pointed out, in a lot of studies dealing with gender in corpus linguistics, small differences often tend to be focused on, whereas they remain minor compared to the similarities, thus giving the erroneous idea that these represent proofs of an inherent divergence between sexes. To try to avoid that, this paper will try to remain as neutral as possible, to see whether a clear distinction can be made out of this corpus or not. As for every study, the methods used here have certain limits that we will discuss here as well.

Corpus description

When carrying out a linguistic study of any kind, the researcher needs a corpus on which they will base their analyses. The way this corpus is collected will directly impact the kind of data gathered, as well as the implications, limitations, and the reliability of this data. Thus, the corpus collection phase has to be carefully thought to have results matching the different aspects the researcher wants to investigate.

What is swearing?

As mentioned in the introduction, certain studies seem to indicate that strong swearing among women may be becoming more and more common on social media. However, being able to determine what can be considered a swear word, and whether it is offensive or not is not an easy task, as not everyone has the same latitude regarding swearing. Indeed, children who swear will sometimes be severely reprimanded, whereas it may go unnoticed among adults (Ladegaard, 2004). Also, people's own perceptions of swear words may influence how offended they are by them (Jay, 1992; Stapleton, 2010), and thus not everyone will be offended by the same words. The way swearing is perceived varies a lot between generations for example (Harris, 1990), and some people may not even consider that certain words are swear words, whereas others will (S. E. Hughes, 1992). For these reasons, it is hard to define a clear and empirical list of swear words on which everyone will agree. In order to compile a list of swear words that would be as objective and appropriate to our sample as possible, we needed to have a standard, a list of words considered swear words by most British speakers. In this regard, the study from Wenbo et al. (2014) seemed to be a good start, as they have managed to put together a list of 788 English swear words and their variations. These words were manually and independently annotated by two native speakers of English who both agreed that these words are "mostly used for cursing". We decided to use the Wenbo et al. (2014) study as a standard on which we would base certain aspects of the methodology and analyses of our pilot study, because we deemed that their study can be considered as a reference point for our own investigation partly due to the fact that their research was carried out in 2014, so it is to this day one of the most contemporary on this topic. It is also very exhaustive, as their corpus is composed of 51M English tweets from all around the world, so their results are more likely to be representative of global trends on Twitter. One of the conclusions they came to is that of the 788 words they used to define swearing tweets, "*the top seven swear words - fuck, shit, ass, bitch, nigga, hell and whore cover 90.40% of all the curse word occurrences* in their corpus. These seven words alone then represent the vast majority of the swear words repertoire of Twitter users in their sample.

Considering the afore-mentioned fact, and in order to be able to maximize the relevance of what we can consider as a swearing tweet in our sample, we chose to include the 20 most common swear words in the Wenbo et al. study. This alone should ensure a reliable representativeness of what can be used to differentiate a swearing tweet from a non-swearing one. However, it could be argued that this list may not necessarily be representative of a majority of native speakers of English, especially as it was sampled by only two native speakers. We mentioned earlier the fact that people's attitudes regarding swear words tend to vary a lot, so relying on two people only seems limited to be able to build a comprehensive list of swear words, especially when considering that for the sake of our study, we wished to focus on the UK only. Thus, what was significant in Wenbo et al.'s study may not be as significant in our own sample, as their study was based on a sample of the worldwide stream of tweets from a given period, whereas our corpus is much more localized, so there may also exist a geographical bias. In order to limit those bias as much as possible we also used the swear words mentioned in the editorial guidelines concerning the use of offensive language by the British Broadcasting Corporation (BBC) and which were not present in the list taken from Wenbo et al.. The BBC can be considered as representative of a standard in terms of what should be labelled as a swear word in the UK, especially as this concerns what is acceptable or not from audiences[1]. We deemed that this would represent a reliable addition we could use to create a list of widely accepted offensive words applicable to a British sample. In the end, the complete list of words we used to distinguish swearing tweets from the others is composed of *fuck, shit, ass, bitch, nigga, hell, whore, dick, piss, pussy, slut, tit, fag, damn, cunt, cum, cock, retard, blowjob, wanker, bastard, prick, bollocks, bloody, crap, bugger*. In other words, if a tweet contains any one, or more, of these words, it will be considered as a swearing tweet.

Social media and swearing

It is inadvisable to presume that the speech patterns displayed on social media are accurately representative of trends present in face to face conversations, especially on Twitter, as users are faced with a limit of 140 characters which does not apply in face to face interactions. However, it can be interesting to compare the way people swear on social media to the way they swear in oral contexts to be able to better understand how these two modes of communication can be compared, and how representative of "real life" trends swearing on Twitter can be. As we discussed earlier, the time dedicated to social media sites like Twitter increases every year. In 2013, a study from Ofcom[2] revealed that in 2012 in the UK, a vast majority of people from all age groups and socio-economic backgrounds used social media. A majority of these people also reported using social media more than once a day, which was not the case in 2011. This illustrates the growing importance that the Internet, and social media in particular, are gaining in our daily lives, consequently increasing the likelihood of daily speech patterns and evolutions of certain linguistic attitudes being present on social media and vice versa. According to Wenbo et al., "*one out of 13 tweets contains curse words*" (Wenbo et al., 2014). As corpora of tweets can be composed of a virtually unlimited number of tweets, the proportion of potentially interesting swear words to analyze thereby represents a very appealing way of generating data. As we stated earlier, the way swear words or other linguistic resources are used and perceived by a specific community can have an impact on the way this community is considered. Conversely, the way swear words are used inside a group can be an indication of evolutions in the way this community identifies itself with regards to its status, or its power for example (Beers Fägersten, 2012; Lakoff, 2004; Murray, 2012; G. I. Hughes, 2006). Since women from the United Kingdom seemed to be the most likely to use strong swear words more than men in previous studies (Thelwall, 2008), we figured that studying the use of swear words of British women and men on Twitter may reveal more profound changes in people's perception and use of swear words, at least in online communities present on Twitter.

Methodology

The main requirements we had in order to be able to carry out our study were thus the age of the informants (as younger generations of women seemed to be the most likely to experience this increase in swearing on social media sites), their gender, and they had to be localized in the UK, as this region seemed to be the most sensitive to the aforementioned phenomenon. Twitter's API (Application Programming Interface) seemed to be the perfect solution in this regard, as it can offer access to every one of these parameters. In order to collect our corpus, we used the streaming API and only requested tweets from the United Kingdom by mentioning the corresponding geolocation. We let the collection of tweets run between 7 April and 15 May 2015 and got a total number of 961,186 tweets from 18,060 users.

Inferring gender on Twitter

Users' gender is determined thanks to the name they provided. We created two repositories of female and male names

[1]For more details on the guidelines regarding what the BBC considers as offensive language, see: `http://www.bbc.co.uk/guidelines/editorialguidelines/advice/offensivelanguage/index.shtml`

[2]See the Ofcom report on Adults' media use and attitudes report, 2013.

given to British babies since the 1950s[3], and they are composed of a total of about 30,000 gendered names. For every user whose tweet we collect, the program automatically checks whether the name provided is present in one file or the other (i.e., whether the name is present in the male or the female repository), and if it is present in one, and not the other, the user is attributed the corresponding gender. As a way to avoid any bias with ambiguous names (names which can be given both to women or men, like Robin for example), if the name is present in both files, the user is considered as undefined, and is rejected.

Inferring age

The age is determined thanks to the information provided by users in the description of their profiles. We have defined a list of patterns which allows the program to automatically identify a digit sequence in a description that corresponds to the age of the user (e.g. thanks to regular expressions, the program will identify 25 from "I'm 25 yo" for example). In order to maximize the accuracy of our results, we decided to split users according to their gender and age groups. We thus categorized users into six different age groups which will be referred to as described in Table 1.

Table 1
Table of notations.

Notation	Meaning
C^m_{12-18}	Tweets published by males aged 12-18
C^m_{19-30}	Tweets published by males aged 19-30
C^m_{31-45}	Tweets published by males aged 31-45
C^m_{46-60}	Tweets published by males aged 46-60
C^f_{12-18}	Tweets published by females aged 12-18
C^f_{19-30}	Tweets published by females aged 19-30
C^f_{31-45}	Tweets published by females aged 31-45
C^f_{46-60}	Tweets published by females aged 46-60

The reason why we chose those age groups is because as many sociolinguistic studies have shown, people we spend a lot of time with can have an influence on the way we speak, especially among children (Eckert, 2008; Stapleton, 2010; Ladegaard, 2004). Since children spend most of their time at school, with peers of the same age, children of the same educational level are more likely to display similar speech patterns. Thus, until age 18, users are classified according to the academic level they are the most likely to belong to in the United Kingdom. According to the Office for National Statistics, in 2013 the average age of mothers was 30 in England and Wales[4], so age 30 will be used as a marker for two age groups. Indeed, studies suggest that parents who have children produce more standard forms than usual and avoid the use of taboo language (Stapleton, 2003; Mercury, 1995), so having babies is likely to influence the linguistic attitudes of people from these generations, hence the need to take it into account in our age classification. These age groups should allow the heterogeneousness of our sub-corpora to be limited as much as possible. Such groups also have the advantage of limiting the interference of problems caused by users who may not keep their profiles up to date for example, and who may claim to be 22 in their descriptions, whereas they would now be 23. The age reported would in this case not be the actual age of the user, but they would still belong to the most appropriate age group.

Analysis and results

To help us analyze this vast collection of tweets and gain insights into the contexts in which Twitter users swear, we leverage several data analysis tools developed in the field of machine learning. But first, some classic data concerning the demographics of our corpus will help us understand its composition.

Distribution of the number of tweets per gender and age

Table 2 and Figure 1 present basic data about the demographics of our corpus. Unsurprisingly, as shown in Figure 2, there is a huge imbalance in the representation of the different age groups taken into account, with a vast majority of our users reported as being between 12 and 30 years old. This was to be expected and this repartition also corresponds to the most represented age groups on Twitter as a whole. Also, what a manual verification revealed is that for both the youngest and the oldest age groups (i.e. the 5-11 and the 61-99), most of the users' descriptions do not correspond to actual human users, or are representative of anomalous profiles, like pages dedicated to companies or pets (for the youngest age group), or clearly untrustworthy profiles (for the oldest age group). This is mainly due to the method we used to gather information concerning the age of our users, based on regular expressions, and which does not make a difference between a profile dedicated to a 4 year old dog, and a a 4 year old boy, although a 5 year old boy is unlikely to have a Twitter profile... Our regular expressions are meant to look for profiles mentioning a number followed by "years old' in users" descriptions (or variations of "years old", like "yo", as it is a very common way to mention one's age on Twitter), among others, but does not take into account any mention of gender, or of being a human, as this is only processed thanks to the name provided. So, to prevent the potential interference of this dubious data, these age groups are never taken

[3]Sources: General Register Office, National Records of Scotland and Office for National Statistics.

[4]See the 2014 report from the Office for National Statistics.

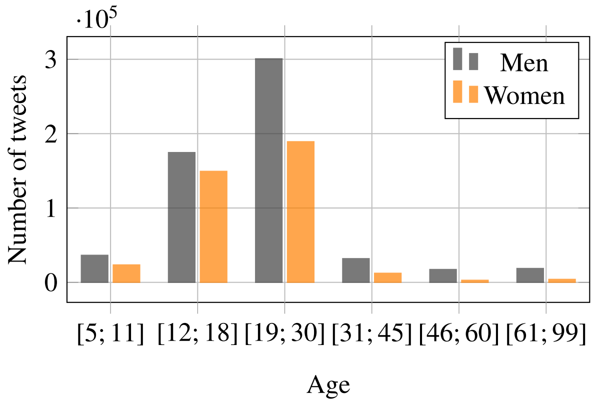

Figure 1. Distribution of the number of tweets per gender and age.

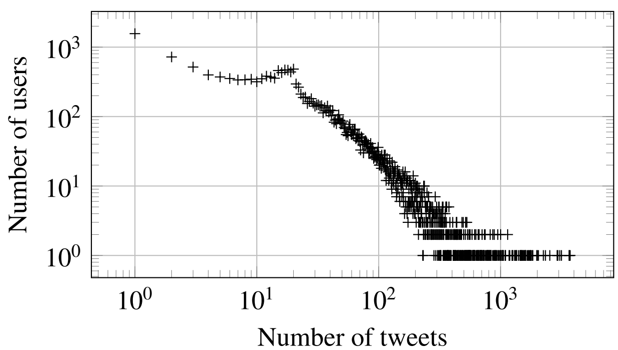

Figure 2. Distribution of the number of tweets per user on a log-log scale.

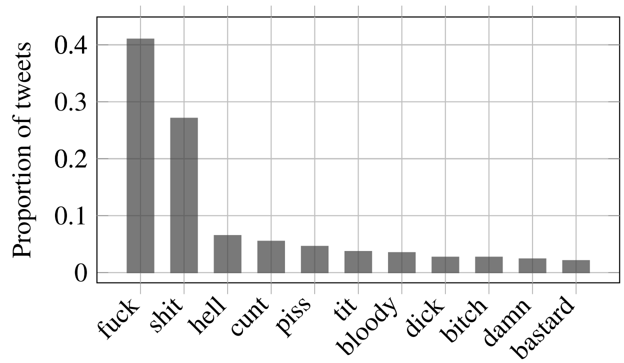

Figure 3. Most common swear words found in swearing tweets published by male users.

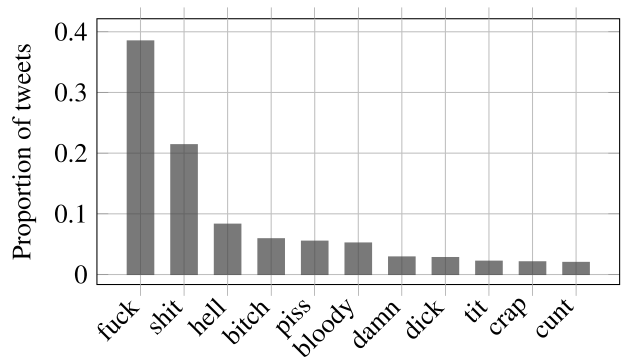

Figure 4. Most common swear words found in swearing tweets published by female users.

into account in the analyses we make, and are just presented here out of a concern for transparency.

Table 2
Basic corpus properties.

	Male	Female	Total
# of users	10313	7747	18060
# of tweets	579864	381322	961186

Distribution of the number of tweets per user

Figure 2 plots the distribution of the number of tweets per user on a log-log scale. We note that it follows a power-law, with one user contributing over 3700 tweets in the corpus, while a lot of users contribute to fewer than 100 tweets, thus showing that the interference of potential spam accounts producing a great number of tweets in a short amount of time is very limited.

Proportion of swearing tweets among women and men

In our corpus, 5.8% of the male tweets contained at least one swear word, compared to 4.8% for women. Figures 3 and 4 present the proportion of tweets containing the eleven most common swear words for women and men. However, as percentages of this kind do not provide much information about the specific use of each word, we normalized the frequency of each swear word on one million words for both women and men. The results are presented below in Table 3.

Proportion of swearing tweets by gender per million words

Table 3 presents the proportions of use of all the swear words we took into account for both genders. As we mentioned before, there is an imbalance in the number of male and female users, as well as in the number of tweets for each gender. Thus, raw percentages would have been useless in that they are not comparable in such situations. To be able to efficiently compare the use of swear words by women and men, we calculated the number of instances of each swear word there would be in one million words. This then gives us an objective value on which to base our anal-

Table 3
Frequency of swear words by gender per million words.

Word	Women	Men	Overused by	*LL*
fuck	**1719.21**	**2105.07**	**Men**	**179.15**
shit	996.36	1198.84	Men	85.96
ass	63.48	63.48	Neither	0
bitch	**262.02**	**142.88**	**Women**	**168.1**
nigga	10.97	18.82	Men	9.51
hell	**344.57**	**309.89**	**Women**	**8.54**
whore	11.23	10.43	Neither	0.14
dick	124.34	133.64	Neither	1.54
piss	257.84	253.08	Neither	0.21
pussy	19.33	33.7	Men	17.92
slut	20.11	15.74	Neither	2.49
tit	**88.03**	**195.24**	**Men**	**187.47**
fag	18.02	30.28	Men	14.3
damn	117.81	102.84	Neither	4.73
cunt	**94.3**	**295.86**	**Men**	**487.84**
cum	8.62	17.79	Men	14.7
cock	82.02	93.08	Neither	3.22
retard	11.23	26.86	Men	29.71
blowjob	1.04	1.88	Neither	1.1
wanker	21.42	49.96	Men	52.82
bastard	68.44	110.54	Men	45.49
prick	47.02	84.36	Men	48.8
bollocks	14.62	37.3	Men	45.96
bugger	18.8	18.3	Neither	0.03
bloody	**227.01**	**173.51**	**Women**	**33.47**
crap	100.05	89.66	Neither	2.64

yses. For each swear word, we calculated the log-likelihood (*LL*) score, which is based on the null hypothesis that there is no difference between the observed relative frequencies of a given swear word in the two corpora (i.e. female tweets and male tweets). We can reject the null hypothesis at the level of $p < 0.01$ when the *LL* value is greater than 6.63 (Rayson et al., 2004). In that case, we consider this word to be characteristic of men or of women. In Table 3, the three most statistically significant words for women and men are highlighted. These words are, in descending order of significance, *bitch*, *bloody* and *hell* for women, and *cunt*, *tit* and *fuck* for men. It would seem that some of the findings of McEnery (2006) are verified here, as in his study of the use of swear words of women and men on MySpace, he found that *fucking, fuck, jesus, cunt* and *fucker* were more typical of males, and *god, bloody, pig, hell, bugger, bitch, pissed, arsed, shit* and *piss* were more typical of females. However, although some of the most significant words for women and men in our sample were also significant for the MySpace users, the ranking of those words is different. *Cunt* is now the most significant word for men, and *bitch* the most significant for women, which may suggest an evolution in the gendered preferences of swear words. However, it may also be due to different ways of swearing and topical differences triggered by the two social media in question (MySpace and Twitter).

Average ratio of swearing tweets per day by gender and age

As mentioned earlier, Twitter's API enables us to collect many information along with tweets themselves. The time at which those tweets are published is one of them. Figures 5 and 6 represent the average swearing ratio throughout the day for male and female users aged between 12-18 years old. The global patterns are the same for women and men for both age groups, the highest peak of swearing ratio being located in every case between 2am and 5am. In other words, this period is the one in which the proportion of swearing tweets compared to non-swearing tweets is the greatest. During the day, the pattern seems to be the same for both genders from both age groups as well, since the swearing ratio for both genders keeps increasing throughout the day. As Wenbo et al. (2014) noted, we notice that the global pattern of swearing tweets corresponds to the standard activity of human life, as users start swearing between 6am and 7am, when people usually wake up, and gradually increases throughout the day. Interestingly, we observe that there is a downfall in the swearing ratio around dinner time (around 7pm and 8pm), which suggests that people tweet, or swear less at that moment. However the ratio increases drastically after that period. As studies have shown, one of the main functions of swearing is to express strong emotions like anger, joy or sadness (Allan & Burridge, 2006; Jay & Janschewitz, 2008). Considering these interpretations, the fact that the ratio for men is constantly higher than women may suggest that they feel more comfortable expressing these kinds of emotions than women. Thus, apart from the differences between genders, what these figures suggest is that both women and men have the same attitude regarding swearing throughout the day. Even if the way women and men swear can differ lexically or quantitatively (as shown in Table 3), some aspects of swear words usage are the same, and apparently the way women and men use swear words according to the time of the day is something which is common to both genders in our corpus.

Named-entity recognition

Named-entity recognition (NER) enables us to automatically locate specific elements in tweets, more precisely the names of people, organizations or locations. To perform this

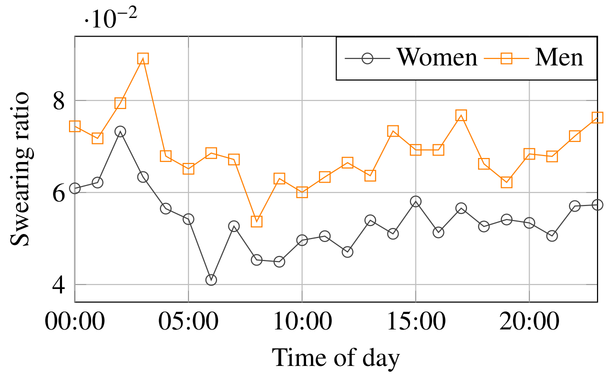

Figure 5. Swearing ratio versus time of day in C^{f}_{12-18}.

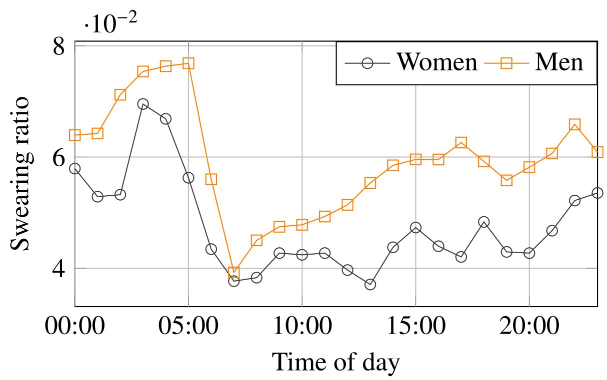

Figure 6. Swearing ratio versus time of day in C^{m}_{12-18}.

task, we use a software (Finkel et al., 2005) which implements a NER method that relies on classification rules based on features of the word sequences that constitute tweets. Table 4 reveals that on average, men use named entities more than women. Also, for both women and men, users tend to mention named entities consistently more as they get older. Figures 7 and 8 present detailed proportions of named entities per gender and age group in swearing tweets. This shows that whatever their age, both women and men majoritarily mention named entities referring to people when swearing. However, what differs is the fact that women from every age groups seem to favor locations over men, who prefer mentioning organizations. This method then highlights the fact that as far as swearing is concerned, context plays a big role. We suggest that these differences may point at gendered differences in the topics women and men focus on, at least when they swear, which may reveal the fact that the pragmatic functions of swear words for women and men of the same age groups may differ. However, more qualitative analyses would be necessary to be able to confirm this hypothesis.

Table 4
Proportion of tweets that contain named entities.

	[12; 18]	[19; 30]	[31; 45]	Average
Women	10.28%	13.41%	14.60%	12.76%
Men	14.25%	19.67%	20.59%	18.18%

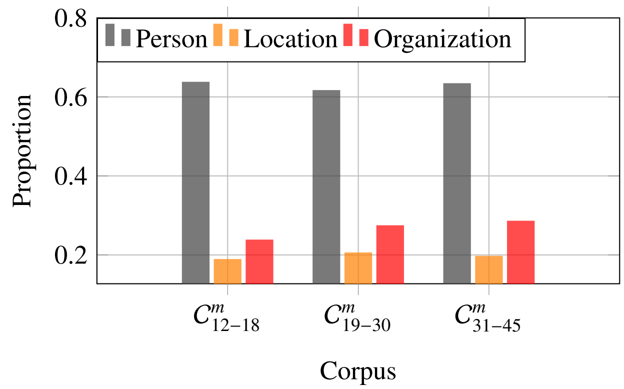

Figure 7. Distribution of the types of named entities in swearing tweets published by men.

Event detection

In order to analyze the impact that real world events may have on Twitter discussions, we studied specific reactions on Twitter triggered by the most influential of these real world events (e.g. the broadcast of a popular TV show, a political event etc...) for users. We use Mention-Anomaly-Based Event Detection (MABED), a statistical method proposed by Guille & Favre (2015, 2014) for the detection of significant events from tweets. Thanks to this method, we are able to map both macro and micro levels of gendered reactions, as it describes each event it detects with a set of words, a time interval and a score that reflects the magnitude of impact of

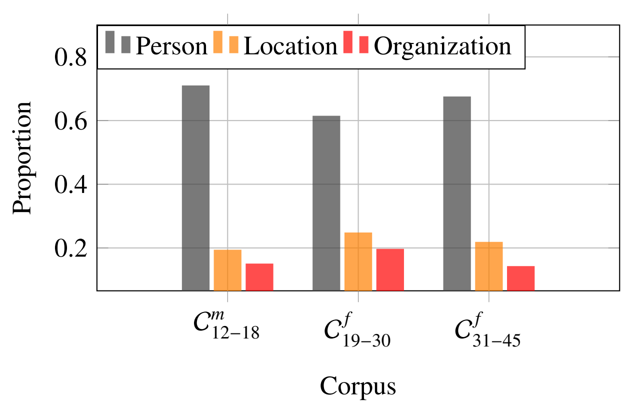

Figure 8. Distribution of the types of named entities in swearing tweets published by women.

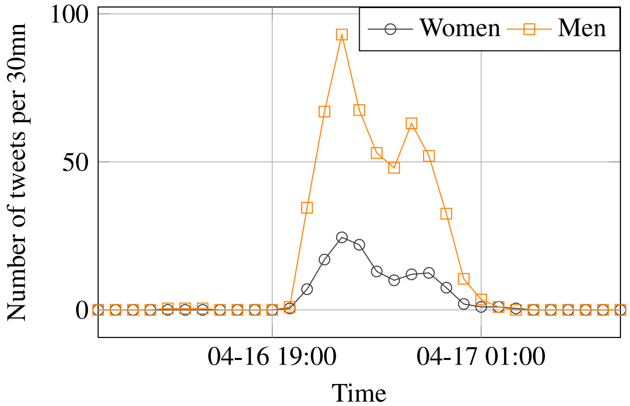

Figure 9. Evolution of the number of tweets containing #bbcdebate.

the event over users. Moreover, it is possible to analyze the tweets associated with these events, to understand their underlying composition, and the way swear words are used in our case for example. Table 5 and Table 6 present as an example the ten most significant events detected by MABED for women and men aged 19-30. These events are numbered from 1 to 10 in decreasing order of significance. The ‘event’ column presents the keywords recognized as the most representative of the event, and the last column presents the percentage of tweets containing at least one swear word inside each event. Events marked as “spam” were events which were considered as such because one single user posted the same spam tweet very often, thus virtually generating keywords considered by MABED as relevant events. These spammers, though being a minority in our corpus as shown in Figure 1, create a considerable amount of noise for our event detection method and prevent a more accurate analysis of gendered events. It is however interesting to note that in this sample, spammers are twice as more present in the female corpus than in the male one, thus suggesting that spam accounts are more likely to adopt a female name.

Apart from spams, which have been manually identified as such, what the results from MABED reveal is that generally speaking, male events could be summarized by sports (boxing and soccer) and politics, and female events by media/entertainment (birth of the Royal baby, BGT (Britain’s Got Talent)), sports (Grand National) and politics. Generally speaking, we observe that throughout those ten events, men use more swear words than women, and the event with the smallest number of swear words among men still contains more of these than the event containing the greatest number of swear words among women. The events containing the most and least amount of swear words are highlighted for each gender, and this reveals that both women and men swear more when talking about politics. What is interesting to notice is that on average, the proportion of swearing tweets in reaction to an event is higher (11.3% for men and 6.53% for women) than in classic interactions on Twitter (5.8% for men and 4.8%). It is also worth mentioning that the only common topic between women and men is the one containing the hashtag #bbcdebate. Figure 9 plots the evolution of the number of gendered tweets containing this hashtag. Though men tweet consistently more about that hashtag than women (but it must be reminded that this graph presents the raw number of tweets, so the gendered imbalance may be explained by the fact that we have more men than women in our corpus), we observe that the two patterns are very similar, and that the events are both detected roughly when the broadcast of the debate starts on television, and gradually decrease after the broadcast is over, as it triggers fewer and fewer reactions. The proportion of swearing tweets inside this common event does not differ much between women and men, which may again suggest that gendered differences in swearing are not triggered by gender alone, but by the context in which swearing occurs. In other words, women and men in the exact same context would not differ much in the linguistic attitudes they display. This would imply that swear words are not so much gendered as contextualized, which would correspond to other studies pointing to the fact that when considering gendered speech patterns, the context of use plays a greater role than gender alone (Eckert, 2008; Bamman et al., 2014; Baker, 2014; Holmes, 1995; Ladegaard, 2004). In our case, further qualitative research would however be needed to confirm or refute that hypothesis.

Limitations

This study presents certain limits. The first one concerns the way we categorized users according to their age. Though it has some advantages, it is not perfect, as some users will have children before age 30, or will leave school before age 18, so the linguistic patterns potentially influenced by those social phenomena may differ. Another potential problem is that we did not include hashtags in our swear word detection methods, and hashtags often contain swear words, thus potentially limiting our data in this regard. A manual verification of the information provided in the description of a lot of users in our sample reveals that many are students. Even if it sounds normal, as the most represented age group is the 19-30, there may thus exist a bias towards this category of users.

Conclusion

In this article, we tried to give hints about new methods which could be used to analyze specific sociolinguistic parameters on Twitter. For that purpose, we analyzed the data of a corpus of about one million tweets from users for whom we could infer both the age and the gender. Even if our data would need to be analyzed more thoroughly and qualitatively in order to draw more generalizable conclusions, our goal

Table 5
Top 10 most impactful events detected from C^{m}_{19-30}.

Rank	Event	% of swearing tweets
1	*spam*	
2	fight maywheather he maypac watch pacquiao	13.00%
3	bbcdebate ed farage about natalie milliband	10.70%
4	messi ronaldo best goal ever boateng world lionel what	10.10%
5	snp seats have	9.60%
6	**tories labour have more you**	**15.50%**
7	*spam*	
8	bournemouth league premier next all well play football	11.70%
9	**exit polls ge2015 have wrong hope right lib**	**9.40%**
10	seat his lost	10.40%

Table 6
Top 10 most impactful events detected from C^{f}_{19-30}.

Rank	Event	% of swearing tweets
1	*spam*	
2	**royalbaby princess kate girl charlotte diana baby name**	**2.50%**
3	*spam*	
4	bgt dog antanddec ant me max omg	6.60%
5	*spam*	
6	baby royalbaby girl princess kate	4.40%
7	National grandnational bets	7.90%
8	*spam*	
9	**bbcdebate nigel up ed nhs would**	**9.20%**
10	Grand national bets	8.60%

here was to show that by combining techniques from both computer science and linguistics, it is possible to provide innovative ways of studying the way women and men swear on Twitter. These tools showed that beyond mere quantitative data which could lead to erroneous impressions and generalizations on the reasons why women and men swear, contextual parameters are sometimes more important in being able to determine what is influential, as we concluded with the event detection and NER analyses. This work is then the continuation of prior studies which showed that gender is often enacted in subtil ways, hence the necessity to develop more tools to explore these questions. We believe that some of the tools presented here can be used improved in future research based on Twitter data, so that the analyses presented in this paper can be refined, especially in order to be more qualitative.

References

Allan, K., & Burridge, K. (2006). *Forbidden words: Taboo and the censoring of language*. Cambridge University Press.

Baker, P. (2014). *Using corpora to analyze gender*. Bloomsbury Publishing.

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

Baruch, Y., & Jenkins, S. (2007). Swearing at work and permissive leadership culture: When anti-social becomes social and incivility is acceptable. *Leadership & Organiation Development Journal*, *28*(6), 492–507.

Beers Fägersten, K. (2012). *Who's swearing now? The social aspects of conversational swearing*. Cambridge Scholars Publishing.

Coates, J. (2004). *Women, men and language: a sociolinguistic account of gender differences in language*. Pearson.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*, 453–476.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *In proceedings of the 43nd annual meeting of the association for computational linguistics* (pp. 363–370).

Guille, A., & Favre, C. (2014). Mention-anomaly-based event detection and tracking in Twitter. In *Proceedings of the ieee/acm international conference on advances in social network analysis and mining* (p. 375-382).

Guille, A., & Favre, C. (2015). Event detection, tracking and visualization in Twitter: a mention anomaly based approach. *Social Network Analysis and Mining*, *5*(1), 1-18.

Hammons, J. W. (2012). *Wgaf: Swearing, social structure and solidarity in an online community*. Unpublished doctoral dissertation, Ball State University, Indiana.

Harris, R. (1990). Lars porsena revisited. In *The state of the language* (p. 411-421). Faber.

Holmes, J. (1995). *Women, men and politeness*.

Hughes, G. I. (2006). *An encyclopedia of swearing: The social history of oaths, profanity, foul language, and ethnic slurs in the English-speaking world*. Sharpe.

Hughes, S. E. (1992). Expletives of lower working-class women. *Language in Society*, *21*(2), 291-303.

Jay, T. (1992). *Cursing in America: a psycholinguistic study of dirty language in the courts, in the movies, in the schoolyards, and on the streets*. John Benjamins.

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, *4*(2), 267–288.

Ladegaard, H. J. (2004). Politeness in young children's speech: context, peer group influence and pragmatic competence. *Journal of Pragmatics*, *36*.

Lakoff, R. (2004). *Language and woman's place*. Harper and Row.

McEnery, T. (2006). *Swearing in English*. Routledge.

Mercury, R. E. (1995). Swearing: A "bad" part of language; a good part of language learning. *TESL Canada Journal*, *13*(1), 268–36.

Murray, T. E. (2012). Swearing as a function of gender in the language of Midwestern American college students. In *A cultural approach to interpersonal communication: Essential readings* (pp. 233–241). Blackwell.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th international conference on statistical analysis of textual data* (p. 926-936).

Smith, A., & Brewer, J. (2012). *Twitter use 2012* (Tech. Rep.). Pew Research Center.

Stapleton, K. (2003). Gender and swearing: A community practice. *Women's and Language*, *26*(2).

Stapleton, K. (2010). Swearing. *Interpersonal pragmatics*, 289–305.

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

Wenbo, W., Lu, C., Krishnaprasad, T., & Amit, P. S. (2014). Cursing in English on Twitter. In *In proceedings of the acm conference on computer supported cooperative work and social computing* (p. 415-425).