CHAPTER 6

Gender and age differences in swearing

A corpus study of Twitter

Michael Gauthier & Adrien Guille
Université Lumière Lyon 2

1. Introduction

Many linguistic features (and attitudes towards them) have traditionally been gendered, that is, attributed to either women or men, and positively or negatively evaluated. The use of swearwords is traditionally associated with men and generally positively so. As Coates (2004: 97) reported, "the folklinguistic belief that men swear more than women and use more taboo words is widespread", consequently leading to the creation of pre-conceived ideas stigmatising women and men who would use a linguistic feature not generally associated with them. 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 and a way of affirming oneself (see Lakoff 2004; G. Hughes 2006; Beers Fägersten 2012; Murray 2012). Thus, the fact that one gender may be perceived as more frequent users of swearwords, or on the other hand, as swearword eschewers, may have an impact on other qualities related to power that we would inherently attribute to one gender or the other, whether these differences are real or not.

Some studies have showed that contrary to what has long been widely believed, women do not swear less frequently than men, nor do they use a drastically different register (S. Hughes 1992; Jay 1992; Coates 2004; Baruch & Jenkins 2007; Thelwall 2008; Hammons 2012; Baker 2014). Indeed, a number of these surveys have shown that what generally differs between women's and men's use of swearwords 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 envisioned that the use of "strong" swearwords (see below for a description of what "strong" swearing is) by

DOI 10.1075/pbns.282.07gau

women would increase in certain contexts (Murray 2012), especially on social media (Thelwall 2008); this seemed especially true for younger generations of users (users aged 16–19 in the case of Thelwall). Thelwall even predicted that "gender equality in swearing or a reversal in gender patterns for strong swearing, will slowly become more widespread, at least in social network sites", such that the use of "strong" swearwords among young women will eventually be more frequent than among (young) men (Thelwall 2008: 102). Thelwall's hypothesis suggests that, as adolescents are often shown to be leading linguistic changes, what he observed may apply to more than just young generations of women in the future, as even women from other generations may follow suit and adopt these linguistic preferences. According to Thelwall then, the swearing patterns displayed in MySpace in 2008 could keep evolving for a certain category of women (especially younger ones), which would correlate with a claim from Herring, who said that computer-mediated communication as a whole could be empowering for women (Herring 2003). Evidence of comparable usage of swearwords in computer-mediated communication could support this claim. Thus, the following question arises: has the prediction made by Thelwall in 2008 been fulfilled six years later, in a society where computer-mediated communication in the context of social media is firmly rooted in people's everyday lives? The aim of this chapter is thus twofold: first, it is to offer a better understanding of the patterns of swearword usage among women and men on social media, and second, it is to show the potential of these media as a source of data for synchronic (and possibly diachronic) sociolinguistic studies on a much larger scale.

The focus of the chapter is language use on the social media platform Twitter. With more than half a billion tweets emitted every day (at the time of this study) around the world, Twitter represents one of the most popular and most populated social media sites. Our study is based specifically on a corpus of just over one million tweets issued by nearly 16,000 users. The corpus used in this study is partly the same as the one we used previously (Gauthier et al. 2015), which aimed at presenting new methods and tools, which could be used by linguists to analyse Twitter data. However, for this study, the data have been refined, expanded and further processed in order to accommodate a sociolinguistic analysis. Similar application of corpus linguistic methodology can be found in this volume, see Chapters 5, 7 and 8. The corpus was populated with tweets by British users of both genders and from different age groups from throughout the United Kingdom. The geographic focus allows us to compare our results with Thelwall's (2008) UK-based findings, which showed no significant gender difference for strong swearing on Myspace, but led him to predict an eventual increase in the use of swearwords among younger women on UK-social media.

The analysis of linguistic change as documented on social media is a fairly new approach to linguistic evolution, especially in regard to the importance that social

media have now compared to the limited impact they had at the time of Thelwall's study. According to a study from Ofcom (see the 2013 Ofcom report), the time we devote to social media sites is growing every year among people from all age groups and all socioeconomic backgrounds (Smith & Brewer 2012). This chapter hopes to advance the field of swearing research with regards both to gender and the relatively new context of social media. In so doing, it also aims to further establish the use of social media in linguistic investigation and pave the way for future studies. To these ends, we first consider differences and similarities in the Twitter data. In order to interpret the data as accurately as possible, we then introduce and apply several statistical tools and methods. Based on a log-likelihood significance test, we consider the sets of swearwords that are more representative of each gender. We then present a deeper sociolinguistic investigation by analysing swearwords in context with their most relevant collocates, as identified according to the mutual information score.

2. Theoretical framework

As we have shown earlier, some studies indicate that swearing, and strong swearing in particular, among young women is increasing on social media. What Thelwall (2008) called "very strong" and "strong" swearing were the two words *cunt* and *fuck* or variations of these words. In this chapter, we focus on these words to investigate their frequency among young women and young men. However, we have also deemed that other words should be analysed as a means of comparison and to observe whether the tendency of young women swearing more than men (if true) is limited to particular words only. For the purpose of this chapter, we consider swearing in the same way as McEnery (2004: 1–2) did, and see a swearword as "any word or phrase which, when used in what one might call polite conversation, is likely to cause offence." From there, being able to determine what can be considered a swearword, and whether it is offensive or not is not an easy task, as not everyone evaluates swearing similarly. Indeed, children who swear will sometimes be severely reprimanded, whereas swearword usage may go unnoticed, or at least uncommented, among adults (Ladegaard 2004). Also, people's own perceptions of swearwords 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 considerably between generations (Harris 1990), and some people may not even consider that certain words are swearwords, whereas others will, and this is especially true when focusing on a specific region like the UK (S. Hughes 1992), as British speakers may not have the same perceptions of swearwords as other speakers of English.

While it may be difficult to define a list of swearwords that would satisfy evaluation from all those represented in our corpus, it is crucial for our corpus linguistic study, as it is in the studies presented in Chapters 5, 7 and 8, that a set of search words be determined. In order to compile a representative list of swearwords, we focused on words recognised as swearwords by most speakers of British English. To this end, we made use of Wang et al. (2014), which provides a list of 788 English swearwords from existing swearword lists and their variations (see also Chapter 5). In their study, the swearwords were manually and independently annotated by two native speakers of English, who both agreed that these words are "mostly used for cursing." This final list of swearwords on which both annotators agreed was what Wang et al. used to identify swearing in tweets. We decided to use the Wang et al. (2014) study as a standard on which we would base certain aspects of our methodology and analysis because their research was carried out in 2014, so it is to this day one of the most recent. It is also very extensive, as their corpus is composed of 51 million English tweets from around the world, making their results 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 swearwords –*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 swearword repertoires of Twitter users in their sample. However, we chose to examine the 20 most frequent swearwords in the Wang et al. study, in order to increase the scope of our own analysis. The resulting list is comprised of *fuck, shit, ass, bitch, nigga, hell, whore, dick, piss, pussy, slut, tit, fag, damn, cunt, cum, cock, retard, blowjob*. That this list has only nineteen words is due to the fact that we have excluded the non-English word *puta* (but see Chapter 7 for more information on *puta*). This wordlist should be reliably recognised as English swearwords, but because swearword status tends to vary among people (see, for example, Chapters 7, 10 of this volume), and the list was determined by only two native English speakers, we believe there is reason to consider even more possible candidates. Furthermore, we wished to target the UK only. While Wang et al.'s study was based on a sample of the worldwide stream of tweets from a given period, our goal is to analyse a much more localised corpus, and thus swearword usage may reflect a geographical bias. In order to account for this, we also used all the swearwords 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 Wang et al. The BBC can be considered representative of a standard in terms of what should be labelled as a swearword in the UK, especially as this concerns what is acceptable or not from

audiences.[1] This represents a reliable addition we can use to create a comprehensive list of words widely recognised as offensive and applicable to a British sample. In the end, our list of 26 swearwords reflects a selected compilation of Wang et al.'s study and the BBC list, and includes *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*.

3. Corpus building

The main requirements we had in order to be able to carry out our study was that (1) we had to have access to the gender and the age of Twitter users, (2) the tweets had to be in English, and (3) the tweets had to be localised in the UK, as this region seemed to be the most sensitive to the aforementioned potential of women using strong swearwords as often or more often than men (Thelwall 2008). 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 CATS (Collection and Analysis of Tweets made Simple), which is an interface aimed at providing tools to easily collect and analyse corpora of tweets (see http://mediamining.univ-lyon2.fr/cats/ for more information), and we only requested tweets from the United Kingdom by selecting the corresponding geolocation. We let the collection of tweets run between 7 April and 2 July 2015. In other words, we requested all the tweets that were emitted during those dates, and that were geo-tagged as being from the UK, thanks to the 'location' indicator that Twitter users can choose to provide when tweeting. According to Sloan et al. (2013), about 1% of tweets are associated with a location indicator. Even if this seems somewhat limited, it should be remembered that about half a billion tweets are produced everyday, and that even this 1% sample thus represents about 5 million geolocalised tweets every day around the world. However, this sampling method still potentially represents a bias, as it could be argued that only a specific type of person may choose to add a location to their tweets.

Also, during the collection of tweets, we further processed them to sort them according to the information we could extract from them. Part of this process involved categorising the gender and ages of the Twitter users. In order to have

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

coherent results, we also filtered the tweets we collected according to the language of those tweets, and we discarded tweets which had not been considered by Twitter as being in English at the moment of the collection, as well as all the other tweets that did not correspond to our other requirements (being able to infer gender, age and geolocation).

However, one problem for us was that Twitter's API does not directly give access to a user's gender or age. Thus, we had to find ways to access these variables using the information that users had already provided.

3.1 Inferring gender

User gender was determined according to the name provided. It should be noted that the name category refers to the name provided by the Twitter user in their profile, and which is different from the screen name, which corresponds to the alias with which users identify themselves (preceded by an '@' symbol). Thus, the name is often more likely to correspond to the actual name of the user, which is the reason why we chose this method to determine the gender of the person. We created two lists of female and male names given to British babies since the 1950s, which are composed of a total of about 30,000 gendered names. If the name provided by users whose tweets we collected was present in only one list, we assigned the related gender to that user. In order to avoid any bias with ambiguous names (names which can be given both to women or men, for example, *Robin*), if the name was present in both files, the user was considered as undefined, and was rejected.

3.2 Inferring age

User age was determined from the information provided in the user's profile. We have defined a list of patterns which allow the program to automatically identify a digit sequence which corresponds to the age mentioned by users in the description part of their profiles (e.g. thanks to regular expressions, the program will identify 25 from "I'm 25 yo" for example). In order to maximise the accuracy of our analyses, we decided to split users according to their gender and age groups. We thus categorised users into four different age groups: 12–18, 19–30, 31–45, 46–60.

4. Results

After processing the corpus, we had a total of 1,065,800 tweets from 15,793 users. In order to test Thelwall's hypothesis that "strong swearing" may be more present among younger generations of women, the results presented here will focus on

tweets from the two youngest age groups taken into consideration in our study, i.e. 12–18 and the 19–30 years old. Thelwall was not specific about the age group that he thought could reflect a reversal of gendered patterns regarding the use of swearwords in the future. What we know is that in his case, the age group in which women used strong swearwords most frequently was 16–19 years old. Our focus age groups cover and expand upon this age range. The sub-corpus on which we will focus then represents a total number of 906,199 tweets.

However interesting quantitative figures may seem, as Brezina and Meyerhoff (2014) showed, when doing sociolinguistic analyses, researchers often tend to base their statistical tests and conclusions on aggregate data, which may not accurately reflect all the intricacies and inequalities of a corpus. These practices still frequently produce statistically significant results (especially with huge corpora), giving the erroneous illusion that the claims made are justified, whereas more detailed analyses may lead to more nuanced results. We thus use various means of calculating dispersion inside our different sub-corpora, in order to locate cases where a minority of users displaying extreme patterns may bias the overall results.

Table 1 presents the distribution of all the swearwords that have been statistically identified as significantly overused by one particular gender for both age groups in our sub-corpus. Log-likelihood tests were applied to compare the relative distributions of swearwords between female and male tweets, and only those that were at least significant at the level of $p < 0.05$ (LL = 6.63) are displayed in the table.

Table 1. Swearword distribution for both genders and age groups

Users 12–18			**Users 19–30**		
Swear word	***LL***	**Tendency**	**Swear word**	***LL***	**Tendency**
fuck	206,41	boy	fuck	379,93	men
shit	107,38	boy	shit	38,40	men
bitch	46,59	girl	bitch	32,29	women
pussy	8,90	boy	cunt	251,85	men
damn	13,02	boy	cock	27,89	men
cunt	198,85	boy	retard	13,04	men
cock	10,51	boy	wanker	25,34	men
retard	14,68	boy	bastard	91,98	men
bastard	18,54	boy	prick	31,06	men
prick	19,92	boy	bollock	43,97	men
bollock	7,13	boy	bloody	36,02	women
bloody	12,92	girl	crap	10,01	women
bugger	10,72	boy	fuck	379,93	men

As we can see, for the 12–18 age-group, two words are significantly more used by girls, *bitch* and *bloody*, the rest being more used by boys. For the 19–30 age-group, three words emerge as typical of women, *bitch*, *bloody* and *crap*.

At first, we may be tempted to conclude that swearword usage, according to Table 1, is more dominantly male, as eleven words are typical of men, and only two are typical of women. Indeed, this difference must be noted. However, as Baker (2014) pointed out, in many 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 genders. In this case, it should be remembered that in Table 1, only the swearwords that are used significantly more by one gender are displayed, and, also, that we basically chose 26 words as representative of swearwords for this study. Thus, there is a set of swearwords that are representative of neither gender, at least according to the log-likelihood tests. The variable use of *ass*, *nigga*, *hell*, *whore*, *dick*, *piss*, *slut*, *tit*, *fag*, *cum*, *blowjob*, *wanker*, and *crap* among 12–18-year-olds does not emerge as significantly different, and nor does the variable use of *ass, nigga, hell, whore, dick, piss, pussy, slut, tit, fag, damn, cum, blowjob*, and *bugger* among 19–30-year-olds. Bearing in mind our comparison with Thelwall's study (2008), it should be noted that in this case, "strong swearing" (i.e. *fuck* and *cunt*) is representative of men in both the age groups taken into account.

However, as mentioned earlier (see Brezina and Meyerhoff 2014), such statistics based on aggregate data are often misleading because dispersion is not taken into account, meaning that a handful of users may overuse certain patterns and bias the results towards a tendency that would not be representative of the whole group. In order to prevent that potential bias, for each corpus we classified the tweets according to the users (all the tweets from the same user were grouped together), and we partitioned the two sub-corpora based on age-groupings in ten sections of equal size. The relative frequency of each swearword was calculated inside those different sections. The results are represented as coloured matrices, where each column corresponds to a swearword, and each line corresponds to a section. Thus, each column-line pair represents the relative frequency (in percentages) of a swearword inside one of the sub-sections of the sub-corpora. As a consequence, the more homogeneous a column is, the closer the relative frequencies of a given swearword between all the different sub-sections are. In other words, if the ten sub-divisions for one single word tend to be of an approximate shade, the relative frequency of use of that swearword is uniformly distributed between these sub-sections. This procedure aims at giving a first idea of the variation in the frequency of use of each swearword among each group, so that potential anomalies, if any are present, can be spotted. Figure 1 and Figure 2 present these results for every one of the swearwords we took into account.

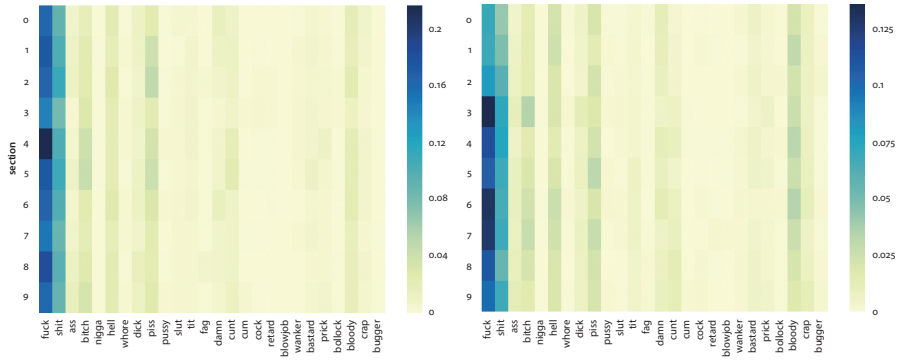

Figure 1. Frequency of use of swearwords by 12–18 girls (left) and 19–30 women (right)

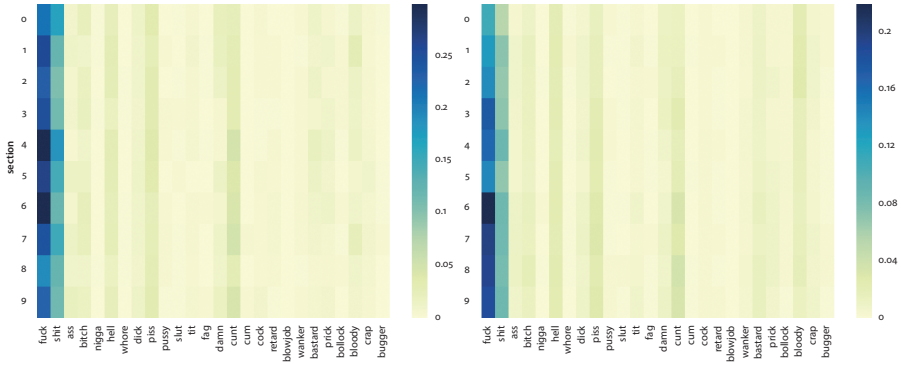

Figure 2. Frequency of use of swearwords by 12–18 boys (left) and 19–30 men (right)

As we can see, these figures indicate an equally distributed frequency of use between the sub-sections for every swearword, as the colours are mostly homogeneous. This is valid for both genders from both age groups, and reinforces the idea that there are very few outliers biasing our data by using certain words much more than other users.

This kind of visualisation provides an interesting overview of the frequency of each variable inside our different sub-corpora, but it still does not provide an accurate and quantifiable measure of dispersion inside these corpora. A widespread measurement of dispersion is the standard deviation; however, this assessment is unbounded. In other words, the similarities and differences between these figures cannot be compared using a scale that would be the same for all of these figures, so it is difficult to compare them objectively. To have an objective value enabling us to compare the distribution of swearwords between women and men from the two age groups we are focusing on, we calculated

Juilland's D (see Oakes 1998 for an overview) from the frequencies illustrated in Figure 1 and Figure 2. Juilland's D is a normalised, statistical measure of dispersion of a word's frequency across all the sub-corpora of a corpus. The dispersion is represented by a value between 0 and 1: '0' meaning that all occurrences of this word are concentrated in a single sub-corpus, and '1' meaning that this word's frequency is evenly distributed throughout the corpus. This enables us to assess whether a swearword is used by a handful of users only, or consistently used by most Twitter users in our corpus. Mathematically speaking, Juilland's D builds upon a classical statistical measure of dispersion, namely the coefficient of variation. This coefficient, noted *cv*, is defined as the ratio of the standard deviation to the mean:

$$cv = \frac{\text{standard deviation}}{\text{mean}}$$

where:

$$\text{standard deviation} = \sqrt{\frac{\text{sum of squared distances from the mean}}{\text{total number of corpus section}}}$$

The D value is based on the ratio between two coefficients of variations measured for two frequency distributions, *X* and *Y*. *X* is the frequency distribution for the word under study, taken from the corpus, while *Y* is the frequency distribution of a hypothetical word, the frequency of which is as unevenly distributed as possible. Thus, the formula for Juilland's D is:

Deriving *cv*(*Y*) shows that it only depends on the number of sub-corpora and equates to:

$$D = 1 - \frac{cv(X)}{cv(Y)}$$

The D values for all swearwords in the sub-corpora are showed in Table 2.

$$cv(Y) = \sqrt{\text{number of subcorpora} - 1}$$

As we can see from this table, the scores obtained for every one of the swearwords in every sub-group are very close to 1, meaning that swearwords are evenly distributed, thus definitely negating the possibility of a handful of users overusing swearwords and creating bias in the data.

Thanks to these detailed analyses of dispersion in our sub-groups, we can confidently assert that the log-likelihood scores we showed earlier are reliable, and not just a consequence of an unbalanced distribution of the variables between

Table 2. Swearword dispersion for both genders from both age groups

	Girls 12–18	Boys 12–18	Women 19–30	Men 19–30
fuck	0.96	0.96	0.93	0.93
shit	0.94	0.93	0.93	0.93
ass	0.80	0.79	0.84	0.81
bitch	0.91	0.91	0.90	0.94
nigga	0.72	0.77	0.70	0.75
hell	0.96	0.94	0.95	0.98
whore	0.80	0.76	0.74	0.76
dick	0.92	0.95	0.89	0.93
piss	0.92	0.95	0.93	0.93
pussy	0.76	0.80	0.77	0.80
slut	0.89	0.87	0.86	0.80
tit	0.90	0.88	0.90	0.88
fag	0.71	0.82	0.84	0.85
damn	0.87	0.86	0.92	0.91
cunt	0.87	0.93	0.86	0.89
cum	0.73	0.72	0.73	0.73
cock	0.78	0.84	0.77	0.91
retard	0.77	0.85	0.76	0.88
blowjob	0.60	0.73	0.65	0.73
wanker	0.79	0.90	0.88	0.93
bastard	0.91	0.86	0.94	0.94
prick	0.86	0.89	0.88	0.89
bollock	0.81	0.86	0.83	0.90
bloody	0.94	0.86	0.94	0.92
crap	0.89	0.88	0.90	0.91
bugger	0.73	0.80	0.81	0.85

the users present in our corpus. Thus, to come back to the comparison between Thelwall's study (2008) and ours, it can now be asserted that at least according to the log-likelihood scores, "strong swearing" cannot be said to be used by younger generations of women in our sample, as *fuck* and *cunt* are the two words which are the most strongly associated with men for both age groups, *bitch* being overall the most female-linked swearword.

As explained earlier (see Introduction), studies have shown so far that what differs between women's and men's use of swearwords may sometimes be the register they use, but most of the time it is more relevant to look at the contexts in which these words are used. An efficient way to thoroughly analyse the context in which words are used is to look at their collocates. A collocation is the frequent co-occurrence of one word with another, so it is one way to learn about the relationship that one word has with other words in specific corpora. Brezina et al. (2015: 142), talking about Phillips (1989), explain that "collocation networks […] can be used to operationalize the psychological notion of the 'aboutness' of a text." Therefore, collocations are a way of getting deeper insight into what a text is about, and thus, in the case of swearwords, collocation analysis can reveal what swearwords are used to talk about. Collocation analysis furthermore represents a useful method for looking beyond the mere quantitative aspect of our data, and studying actual differences in swearword usage patterns. For this purpose, we used GraphColl, which allows us to "create a collocation network at any level of complexity, including for instance first-, second-, third-, etc. level collocations (counting from the original node)" (Brezina et al. 2015: 149), a node being a word whose collocates we want to study.

So, to have a more thorough understanding of women's and men's contextual uses of swearwords, we are mainly going to focus on three words, namely: *fuck*, *cunt* and *bitch*. *Fuck* and *cunt* were the two words considered by Thelwall (2008) as representative of strong swearing, and even if a quantitative analysis did not reveal any tendency supporting his hypothesis, it is still important to observe these words in greater details to see whether specific patterns can be noticed. *Bitch*, in our corpus, is the swearword that is the most used by women as a whole, at least according to the log-likelihood tests performed. We include it in our target analyses to compare the way it is used by women and by men, seeking any gender-specific patterns.

Using the GraphColl program, we chose the Mutual Information (MI) score, which according to Brezina et al. (2015: 151), "is an association measure commonly used in corpus studies and implemented in a large number of corpus tools." This process thus represents a reliable point of origin from which we can base our investigation of relevant collocates. In the GraphColl parameters regarding the MI score, we kept the default minimum frequency at 5, meaning that all collocates that do not appear at least five times are discarded. This serves as a way to only count the collocations that are representative of an actual trend, in contrast to collocations that appear only once (hapaxes); these may be due to a spelling mistake or any other reason, making the collocation irrelevant in our case. We also chose a span of five words to the left, and five to the right of the central node, meaning that a word may be considered a collocate if it is used inside that span, but will be discarded if it is

further away than this. Because of the sizes of our different sub-corpora, we set the association measure at 5 (instead of 3 with the default configuration), so that graphs could only display the most relevant collocates and be more readable; otherwise the number of potential collocates would be too large. The association measure is the score that is associated to a collocate of a word in order to assess how relevant this word is as a collocate. Thus, only collocates having an MI score of 5 or more will be displayed, keeping in mind that the higher the value is, the stronger the collocation link between two words. This will make our list of collocates much smaller and readable, while still focused on the most relevant words.

Before looking at the collocates of each word in detail, it may be useful to have an overview of what the most frequent collocates of each swearword are for women and men in each age group. Table 3 presents the fifteen most relevant collocates of *fuck*, *cunt* and *bitch* for both genders from the two age groups we focus on, as well as the total number of collocates for each word (in brackets). Collocates are ordered according to the value of the MI score obtained for each word, higher values being at the top. Collocates in bold characters are those common to women and men from the same age group and for the same word.

The first thing to notice in Table 3 is the discrepancy between the number of collocates that are common to both genders in some cases. For some, there are a lot of common collocates between women and men (*fuck* and *bitch* for the 12–18-year-olds), and for others, there are very few collocates (*cunt* for the 19–30-year-olds). There seem to be two factors at play in what will influence this: the word itself, and the age group. Indeed, for these three words at least, it is obvious that users from the age group 19–30 have on average fewer collocates in common than the age group 12–18. It is also the case that for both age groups, *fuck* is the word for which there is the greatest number of common collocates, followed by *bitch*, and *cunt* being last, displaying very few similarities. Thus, girls and boys aged 12–18 seem to use these swearwords in much more similar contexts than the users aged 19–30. There may be several interpretations for this; however, the main reason may be the span of the age groups we chose. The 19–30 age group covers a span which is almost twice as large as the 12–18 age group, and this may potentially create a bias, which in turn could represent a methodological issue that we would need to address in future studies.

As we mentioned earlier, studying collocates can be an effective way to investigate the contextual use of certain words, in our case providing information concerning how women and men use these swearwords. This kind of analysis can point to common gendered linguistic patterns, or on the other hand, to how these patterns may differ between genders. When looking at the total number of collocates of the word *fuck* for women and men, we realise that for both age groups, there are almost twice as many collocates of the word for men

Table 3. Most relevant collocates of *fuck*, *cunt* and *bitch*

12–18					
Fuck		**Cunt**		**Bitch**	
Girls (22)	*Boys (41)*	*Girls (13)*	*Boys (22)*	*Girls (22)*	*Boys (18)*
sake	**sake**	sha	scruffy	resting	babble
22nd	plants	daft	**ugly**	moody	rebel
tae	trees	**cunt**	nae	silly	madonna
holy	mongo	**ugly**	fat	**stupid**	**Lil**
off	**holy**	ma	silly	**ass**	**Ass**
shut	**off**	**ya**	**Ya**	**fat**	Ugly
dumb	**tae**	worst	tae	bye	**Fat**
themselves	gee	fucking	massive	lazy	party
actual	@ellisaoakley	absolute	stupid	**lil**	Little
da	**shut**	he's	yer	**such**	**Such**
fuck	gives	you're	wee	bitch	**stupid**
flying	**knows**	little	such	**face**	**Face**
knows	ignorant	being	**cunt**	**called**	Ya
@charlotteluwit	**dumb**		Ye	fuckin	**called**
	thick		mad	call	Tell
19–30					
Fuck		**Cunt**		**Bitch**	
Women (32)	*Men (58)*	*Women (9)*	*Men (31)*	*Women (22)*	*Men (25)*
sake	**sake**	**cunt**	bald	resting	madonna
sociology	kiwi	wee	daft	karma	basic
holy	**holy**	called	**cunt**	basic	**Lil**
pish	yourselves	fucking	fat	psycho	Ass
cunts	**shut**	such	#got	life's	Yo
shut	skinny	stop	nae	**lil**	**bitch**
actual	tae	fuck	silly	nasty	Silly
punch	outta	some	tae	ugly	Dirty
cares	**off**	little	boring	**bitch**	**stupid**
gerrard	thick		useless	face	main
off	@andymufc20		lazy	**stupid**	#bbuk
lit	@jamiebrownwhit		fake	called	Cut
gives	@nufc		Mayweather	she's	crazy
fuck	**actual**		stupid	#bbuk	damn
wit				such	ain't

than for women. This far greater number of collocates may be the reason why the log-likelihood scores are so strongly in favour of men. Also, if there are more instances of use of *fuck* among men, this increases the probability that men use *fuck* in a far greater range of contexts than women. It should be noted that our corpus has not been lemmatised (i.e., not sorted by variant forms/inflections of the same word, e.g. *fuck* and *fucker* are considered different entities), and that in this case, we are presenting the results for the word *fuck* itself, and not its derivatives (*fucking*, *fucker* etc.). As such, a gendered preference of certain forms of the word cannot explain this tendency, and this reinforces the idea that the range of contexts in which the word itself is used plays an important role for men from both age groups. However, it should also be noted that despite this, the most relevant collocates of *fuck* for both genders are very similar, and this is the swearword displaying the greatest similitudes between women and men. Thus, there is a core of contexts common to both genders, while the collocates influencing the representativeness of the word for men could be considered marginal, at least compared to the top collocates.

Cunt, on the other hand, which was strongly associated with men according to the log-likelihood scores, here only has a few collocates common to women and men. What is more, the number of relevant collocates (with an association measure of at least 5) is more limited for women in both age groups, which explains the fact that only 13 of them for girls aged 12–18, and 9 of them for women aged 19–30, are displayed. Thus, in addition to being quantitatively used a lot less by women, as shown by the log-likelihood scores, it also seems that the range of contexts in which women use the word is limited, which seems to completely go against the idea that it is becoming distinctive to females. However, despite the reduced number of relevant collocates for this word, it should be noted that for girls aged 12–18, a strong collocate is *he's*. At this point it should be reminded that we set the limit of the association measure for the MI score to 5, to isolate the strongest collocates only. Thus, although *he's* appears as only the tenth most relevant collocate, it is still very pertinent. The fact that a gendered personal pronoun appears as one of the most relevant collocates of *cunt* for girls 12–18, and that no other gendered pronoun is present for boys of the same age (even when looking at the whole array of the collocates of *cunt*) may indicate that *cunt* is a female way to indirectly talk about men. This is interesting, as it directly echoes the work of McEnery (2004: 33–34) who found that in the BNC "the word *cunt* is directed exclusively at males by females. It is a pure intergender BLW [Bad Language Word] for females." Moreover, as Stapleton (2003) explains, "Risch (1987) demonstrates the way in which swearing may be used to denigrate outgroups (in this case, men), thereby strengthening both the internal bonds and the external boundaries of the ingroup." Although in our case, the word was not addressed directly at men (i.e. *he's* meaning that one is indirectly talking

about someone else) and was not necessarily 'exclusively' used to talk about men as in McEnery's study, the correlation between our study and the two quotations mentioned above may imply that *cunt* could have a specific pragmatic function for women from this age group.

Bitch is the swearword that was the most representative of women from the two youngest age groups in our corpus, so analysing the patterns related to this word in greater detail may give us clues as to the preferred linguistic usage of women when swearing. As we have seen before, looking at the use of personal pronouns can, in certain cases, help in analysing gendered preferences in the way swearing is used by women and men. In the case of *bitch*, this can enable us to spot such cases. Indeed, when looking at the graphical representation of collocates for a specific word using GraphColl, it is possible to see every collocate of the word. In the case of *bitch*, this enables us to realise that the word is used by both genders from both age groups.

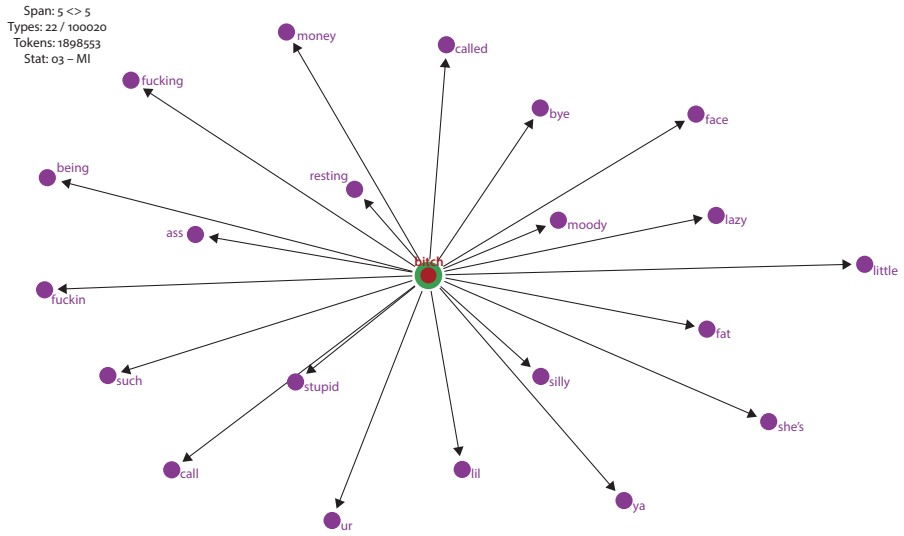

Figure 3. Visual representation of the collocates of *bitch* for girls 12–18

As we can see in Figure 3 and Figure 4, gender-specific pronouns are common to both girls and boys, with *she's* used frequently by the girls, and *she* and *her* by the boys, although they are not among the fifteen most relevant collocates (the shorter the arrow, the stronger the link between the node and the collocate). However, this is not surprising, as the term *bitch* used as a swearword (and not as designating a female dog) can be used to refer to "a spiteful or unpleasant woman" (New Oxford American Dictionary).

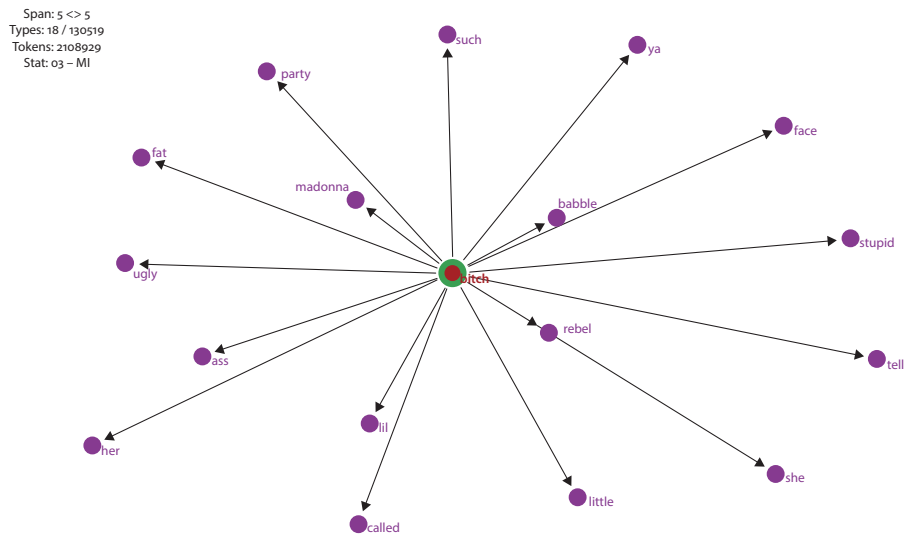

Figure 4. Visual representation of the collocates of *bitch* for boys 12–18

One interesting aspect of GraphColl is that it enables the researcher to navigate between the different nodes by expanding the context of each collocate. In other words, it is possible to click on any collocate displayed in the graph in order to see its own collocates, thus treating it as a new node. In the case of Figures 3 and 4, the central node is *bitch*, but it is possible to expand the context of any collocate of *bitch* to see its most relevant collocates. It is then possible to have a look at a network of collocates in the same figure, which gives a much more accurate idea of how words are linked to one another in different sub-corpora. Let us take the example of *moody* for example, which is the second most relevant collocate of *bitch* for girls in the age group 12–18. *Moody* is extremely relevant for girls in that age group, but not at all for boys. In fact, it is the only group in which *moody* seems to be relevant, since it does not appear as a collocate of *bitch* for women and men aged 19–30. By expanding the context of *moody* as shown in Figure 5, we may be able to better understand why it is so specific to those users.

As we can see, by extending the context of *moody* as well, we quickly realise that the only relevant collocate of that word is the structure *I'm*, which enables us to understand that for girls of this age group, the expression "*I'm a moody bitch*" is very relevant, highlighting the fact that this expression is very specific to this sub-category of users. Thanks to the visual representation of collocates in nodes, we are thus able to better understand specific linguistic patterns inside different sub-corpora, and in our case, this helps us understand how women and men use swearwords, and which patterns are most representative of each. *Moody bitch* is

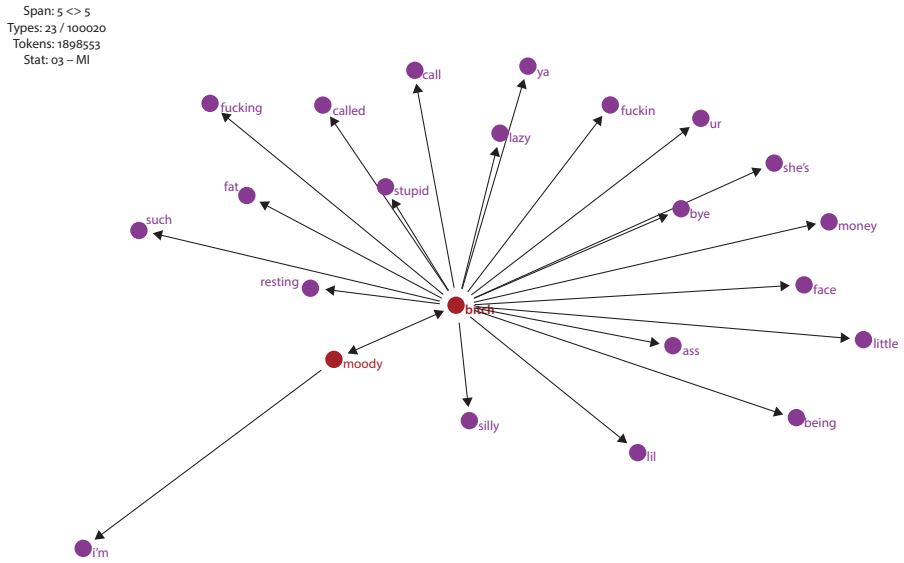

Figure 5. Visual representation of the collocates of *moody* and *bitch* for girls 12–18

used only by girls to talk about themselves, and as such, the specificity of this intra-gender expression may be one factor determining why it is *bitch*, and not *cunt* or *fuck*, that is statistically more representative of female tweets.

5. Conclusion

Thelwall (2008) hypothesised that women from younger age groups would gradually start using strong swearwords more than men on social media, and especially in the UK. Such a linguistic change could have an impact on other qualities we could attribute to both genders, and could therefore contribute to a redefinition of gendered expectations. Our study aimed at confirming or refuting that hypothesis through the analysis of a corpus of one million tweets emitted from the UK, and from users of different age groups. Statistical tests revealed that no matter the age of the users, the two swearwords that accounted for strong swearing in Thelwall's study (*cunt* and variations of *fuck*) were highly associated with men, *bitch* being the swearword that was the most statistically significant for women. A more detailed examination of the collocates of those three words revealed that not only did men use *cunt* and *fuck* more often than women, they also used it in a greater range of contexts. This detailed exploration also revealed specific gendered patterns that enabled us to understand the implications of certain swearwords, and the role they can play for both genders.

In this case then, Thelwall's theory does not seem to hold, at least according to our sample. However, it should be remembered that the corpus on which he based his study was taken from MySpace, which, due to its content and orientation, may attract a different category of people from that of Twitter, potentially explaining the discrepancy observed between Thelwall's results and ours. Indeed, MySpace was originally aimed at sharing and discussing music-oriented content, so this aspect alone may be the source of discrepancies between the Twitter and MySpace population. Another factor that may be at play here is the constraint of Twitter in terms of the length of the tweets. The 140-character limit makes it a very specific mode of communication, which alone may incite linguistic patterns that are different from patterns observed in other modes of written or spoken communication, either increasing the use of swearwords for some people or decreasing it for others.

Despite those findings, and perhaps more importantly, this study highlighted the fact that beyond the representativeness of certain swearwords for each gender, the majority of the swearwords considered did not show significant gendered differences in representation. Thus, certain words can be said to be overused by women or men, but as Baker (2014) pointed out, it should not overshadow the fact that both genders actually use swearwords in a way that is more similar than different.

References

Baker, Paul. 2014. *Using Corpora to Analyze Gender*. London: Bloomsbury Publishing.

Baruch, Yehuda, and Stuart Jenkins. 2007. "Swearing at Work and Permissive Leadership Culture: When Anti-Social Becomes Social and Incivility is Acceptable." *Leadership & Organisation Development Journal* 28 (6): 492–507. doi: 10.1108/01437730710780958

Beers Fägersten, Kristy. 2012. *Who's Swearing Now? The Social Aspects of Conversational Swearing*. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Brezina, Vaclac, and Miriam Meyerhoff. 2014. "Significant or Random? A Critical Review of Sociolinguistic Generalisations Based on Large Corpora." *International Journal of Corpus Linguistics* 19 (1): 1–28. doi: 10.1075/ijcl.19.1.01bre

Brezina, Vaclav, Tony Mcenery, and Stephen Wattam. 2015. "Collocations in Context. A New Perspective on Collocation Networks." *International Journal of Corpus Linguistics* 20 (2): 139–173. doi: 10.1075/ijcl.20.2.01bre

Coates, Jennifer. 2004. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. Edinburgh: Pearson.

Gauthier, Michael, Adrien Guille, A. Deseille, and Fabien Rico. 2015. "Text Mining and Twitter to Analyze British Swearing Habits." *Handbook of Twitter for Research*. Lyon: Emlyon Press.

Hammons, James. 2012. *WGAF: Swearing, Social Structure and Solidarity in an Online Community*. Unpublished Doctoral Dissertation, Ball State University, Indiana.

Harris, Roy. 1990. "Lars Porsena Revisited." *The State of the Language*: 411–421.

Herring, Susan. 2003. "Gender and Power in Online Communication". In *The Handbook of Language and Gender*, ed. by Janet Holmes and Miriam Meyerhoff, 202–228. Oxford: Blackwell. doi: 10.1002/9780470756942.ch9

Hughes, Geoffrey. 2006. *An Encyclopedia of Swearing: The Social History of Oaths, Profanity, Foul Language, and Ethnic Slurs in the English-Speaking World*. London: ME Sharpe.

Hughes, Susan. 1992. "Expletives of Lower Working-Class Women." *Language in Society* 21 (2): 291–303. doi: 10.1017/S004740450001530X

Jay, Timothy. 1992. *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets*. Amsterdam: John Benjamins. doi: 10.1075/z.57

Ladegaard, Hans. 2004. "Politeness in Young Children's Speech: Context, Peer Group Influence and Pragmatic Competence." *Journal of Pragmatics* 36: 2003–2022. doi: 10.1016/j.pragma.2003.11.008

Lakoff, Robin. 2004. *Language and Woman's Place*. New York: Oxford University Press.

McEnery, Tony. 2004. *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.

Murray, Thomas. 2012. "Swearing as a Function of Gender in the Language of Midwestern American College Students." In *A Cultural Approach to Interpersonal Communication: Essential Readings*, ed. by Leila Monaghan, Jane E. Goodman, and Jennifer Meta Robinson, 233–241. Hoboken: Blackwell.

Oakes, Michael. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Phillips, Martin. 1989. *Lexical Structure of Text* [Discourse Analysis Monograph 12]. Birmingham, UK: University Of Birmingham.

Risch, Barbara. 1987. "Women's Derogatory Terms for Men: That's Right, 'Dirty' Words." *Language in Society* 16 (3): 353–358. doi: 10.1017/S0047404500012434

Sloan, Luke, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. 2013. "Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter." *Sociological Research Online* 18 (3): 7. doi: 10.5153/sro.3001

Smith, Aaron, and Joanna Brewer. 2012. *Twitter Use 2012*. Washington, DC: Pew Research Center.

Stapleton, Karyn. 2003. "Gender and Swearing: A Community Practice." *Women and Language* 26 (2): 22–33.

Stapleton, Karyn. 2010. "Swearing." *Interpersonal Pragmatics*: 289–305.

Thelwall, Mike. 2008. "Fk Yea I Swear: Cursing and Gender in a Corpus of Myspace Pages." *Corpora* 3 (1): 83–107. doi: 10.3366/E1749503208000087

Wang, Wenbo, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. "Cursing in English on Twitter." Proceedings of the *ACM Conference on Computer Supported Cooperative Work and Social Computing*: 415–425.