HAL
open science

Age, gender, fuck, and twitter : a sociolinguistic analysis of swearing in a corpus of British tweets

Michaël Gauthier

▶ To cite this version:

Michaël Gauthier. Age, gender, fuck, and twitter : a sociolinguistic analysis of swearing in a corpus of British tweets. Linguistics. Université de Lyon, 2017. English. NNT : 2017LYSE2079 . tel-02619761

HAL Id: tel-02619761
https://theses.hal.science/tel-02619761

Submitted on 25 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre NNT : 2016LYSE2079

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 484 Lettres, Langues, Linguistique et Arts

Discipline : Etudes anglophones

Soutenue publiquement le 30 septembre 2017, par :

Michaël GAUTHIER

Age, gender, fuck, and Twitter : A sociolinguistic analysis of swear words in a corpus of British tweets

Devant le jury composé de :

Tony MCENERY, Professeure d'université, Université de Lancaster, Président
Andrea PIZARRO PEDRAZA, Professeure d'université, Université Catholique de Louvain, Rapporteure
Vaclav BREZINA, Professeur d'université, Université de Lancaster, Examinateur
Céline POUDAT, Maître de conférences, Université de Nice, Examinatrice
Jim WALKER, Maître de conférence HDR, Université Lumière Lyon 2, Directeur de thèse
Kristy Beers FÄGERSTERN, Professeure d'université, Södertörn University, Co-Directrice de thèse

Contrat de diffusion

Ce document est diffusé sous le contrat *Creative Commons* « Paternité – pas d'utilisation commerciale – pas de modification » : vous êtes libre de le reproduire, de le distribuer et de le communiquer au public à condition d'en mentionner le nom de l'auteur et de ne pas le modifier, le transformer, l'adapter ni l'utiliser à des fins commerciales.

Thèse

Age, gender, fuck, and Twitter :
A sociolinguistic analysis of swear words in a corpus of British tweets

Michael Gauthier

2017

Université Lumière Lyon 2
U.F.R. des Langues
Département d'Etudes du Monde Anglophone
Ecole Doctorale 3LA
CRTT (Centre de Recherche en Terminologie et Traduction)

Doctorat en Etudes Anglophones

Directeur/trice de thèse :

Jim WALKER, Université Lumière Lyon 2
Kristy BEERS FÄGERTSEN, Université Södertörns högskola

Jury :

Kristy BEERS FÄGERTSEN, Université Södertörns högskola (Directrice)
Vaclav BREZINA, Université de Lancaster (Examinateur)
Tony McENERY, Université de Lancaster (Rapporteur)
Andrea PIZARRO PEDRAZA, Université de Louvain (Rapportrice)
Céline POUDAT, Université de Nice Sophia-Antipolis (Examinatrice)
Jim WALKER, Université Lumière Lyon 2 (Directeur)

ACKNOWLEDGEMENTS

People are at the basis of every human activity to begin with. They are also at the basis of academic work, as the sole purpose of research is to improve the knowledge we have of certain topics, then share our findings with others to build upon that and further refine it. It seems logical to claim that other people have been central to the completion of this PhD project then. But beyond the mere support we may thank others for in this section, I want to emphasize the crucial role that human interactions and collaborations have had in this thesis. Actually, I consider this thesis as the fruit of a collaborative endeavor and movement instead of it being the result of the efforts of one person supported by others. Without the bonds I have been lucky enough to create all along these four years of research, this experience would not have been as pleasurable, enlightening, and fruitful as it has been. Above all, and beyond the scientific knowledge I acquired through this PhD, the most valuable thing I learned is that exchanging with others is far more rewarding academically and humanly than a low p-value.

The first two people I want to thank are the ones who have been guiding this research: Jim Walker and Kristy Beers Fägersten. I thank them for their guidance, support and comments at every stage of this project.

I am also grateful for the support provided by my research laboratory (the CRTT) and my Doctoral School (the ED3LA).

Beyond the mere research sphere, I also wish to thank all the wonderful colleagues and fellow teachers I have had in Saint Etienne, Lyon and Grenoble, as they have played a major role in having me enjoy teaching as much as research.

I cannot but thank Adrien Guille for his help in giving birth to the CATS project first, which has been one of the tools used to collect the data on which this study is based, but also for his help when I (often) struggled with the Python programs I wrote to sort the data.

As I mentioned the CATS project, I have to thank the numerous people who also got involved in the various stages of development of the interface, as without them CATS would not have evolved as much.

I need to thank Tony McEnery, and everyone from the CASS research Center (and Lancaster University as a whole), for their warm welcome. My staying among them has been one of the richest academic (and human) experiences I have known.

I thank Vaclav Brezina for his advice and encouragements at various stages of the research. Our exchanges have been a true inspiration helping me get a better grasp on how I should approach corpora.

I obviously wish to thank the members of my committee for agreeing to read and assess this thesis.

Many thanks to Julie Villessèche, for being so quick and efficient to proofread this thesis, as well as for her support and ability to always find my jokes (and irony) funny.

Last but not least, I thank my family and friends for their support along these years of research.

INTRODUCTION **6**

PART 1: THEORETICAL FRAME **11**

CHAPTER 1: GENERAL CONCEPTS 13

1.1.1 What is a swear word? *14*

1.1.2 "Separate Worlds Hypothesis" *16*

1.1.3 New approach: new results? *19*

1.1.4 "Women's language": just another myth? *24*

1.1.5 Gender on the Internet: the new neutral? *28*

CHAPTER 2: SWEAR WORDS, PEOPLE AND SOCIAL MEDIA: THEIR INTERCONNECTED EVOLUTIONS 34

1.2.1 Is swearing more common than before? *36*

1.2.2 The "F-Bomb": still as "explosive" as before? *41*

1.2.3 "Overswearing" on the Internet *47*

1.2.4 Why Twitter *52*

CHAPTER 3: CORPUS LINGUISTICS AS AN ADAPTIVE SET OF TOOLS 56

1.3.1 "Is this a sociolinguistic study? Or is it corpus linguistics? Computational linguistics?" *57*

1.3.2 The limits of corpora: Quantitative Vs Qualitative? *61*

1.3.3 What I consider a swear word to be for this research *66*

PART 2: METHODOLOGY **71**

CHAPTER 4: FROM IRL TO API 72

2.4.1 What to do when no ready-made corpus fits your needs? *73*

2.4.2 About Twitter *76*

2.4.3 Some aspects of Twitter's APIs *78*

CHAPTER 5: HOW CATS HELPED ME GET A GRASP ON THE BLUE BIRD 84

2.5.1 Creating a new corpus *85*

2.5.2 Metadata available *93*

CHAPTER 6: HOW OLD IS HE OR SHE? 97

2.6.1 Inferring gender *98*

2.6.2 Inferring age *105*

PART 3: RESULTS **117**

CHAPTER 7: SOME GENERAL FIGURES 118

3.7.1 Swear word distribution according to gender and age *119*

3.7.2 About the word fuck *127*

3.7.3 How often do swear words appear among users? *129*

3.7.4 Overall swear word use by gender and age *130*

3.7.5 A gendered generational gap? *135*

CHAPTER 8: WHO PREFERS WHAT? 140

3.8.1 MWU tests *141*

3.8.2 Keyword analysis: inter-gender variations *145*

3.8.3 Keyword analysis: intra-gender variations *161*

CHAPTER 9: COLLOCATIONAL ANALYSES 171

3.9.1: About the LancsBox tool: *172*

3.9.2 About fuck: *177*

3.9.3 About fucking (and its variants): *184*

3.9.4 About wtf: *189*

3.9.5 About bitch: *194*

3.9.6 About bloody: *200*

3.9.7 About cunt: *203*

3.9.8 About the hashtag #euro2016: *208*

3.9.9 About other females: *212*

CONCLUSION **217**

BIBLIOGRAPHY **220**

APPENDICES **227**

Appendix 1: Comparison of gendered keywords for users aged 12-18 in all tweets *227*

Appendix 2: Comparison of gendered keywords for users aged 19-30 in all tweets *228*

Appendix 3: Comparison of gendered keywords for users aged 31-45 in all tweets *229*

Appendix 4: Comparison of gendered keywords for users aged 46-60 in all tweets *230*

RÉSUMÉ EN FRANÇAIS **231**

Introduction

> Gender consists in a pattern of relations that develops over time to define male and female, masculinity and femininity, simultaneously structuring and regulating people's relation to society. It is deeply embedded in every aspect of society - in our institutions, in public spaces, in art, clothing, movement.
> (Eckert and McConnell-Ginet, 2003: 33)

Speech, or how people use language to express themselves more generally, could easily be added as another defining aspect of gender. Gender norms pervade many layers of our society, and more or less strongly influence the expectations we may have of others. Among these pre-conceptions, many linguistic patterns have been said to be representative of male or female features, like tag questions, deference, turn-taking for example. As I will show in details later (see Chapters 1, 2 and 3), many of these pre-conceived ideas have been contradicted, and some are still discussed. Of all the gendered linguistic characteristics, the one which may have been the most debated is that of swear words. Swearing is indeed a subject which, even when gender is not concerned, generally provokes many tensions and debates. This is partly due to what swear words are often associated to, that is, what is called "bad language". Actually, bad language is a very general concept which can refer to swearing, but also to other aspects of language which can be considered as unacceptable such as slang, jargon, non-standard grammar, dialects, or new forms[1]. Because of a complex interplay between social expectations and power relations, swearing has traditionally been associated with men (see Chapter 1). Indeed, "the folklinguistic belief that men swear more than women and use more taboo words is widespread" (Coates, 1986: 97), consequently leading to the creation of pre-conceived ideas stigmatizing 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, 1973; G. Hughes, 2006; Beers Fägersten, 2012; Murray, 2012). Thus, the fact that one gender may be perceived as more frequent users of swear words, or on the other hand as swear word eschewers, may have an impact on other qualities related to power that we would inherently attribute to one

[1] For more details concerning all the linguistic features which are considered as part of bad language, see Trudgill and Andersson (1990) for example.

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 (see Chapter 1). Indeed, these investigations have showed that what generally differs between women's and men's use of swear words is not so much 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" swear words[2] by women would increase in certain contexts (Murray, 2012), specifically on social media[3] (Thelwall, 2008); this seemed especially true for younger generations of users (users aged 16-19 in the case of Thelwall). It was 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" (Thelwall, 2008: 102), such that the use of strong swear words among young women will eventually be more frequent than among (young) men. This hypothesis suggests that, as adolescents are often shown to lead linguistic changes, what Thelwall 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. Accordingly, 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 (2003), who said that computer-mediated communication as a whole could be empowering for women (see Chapter 2). Evidence of comparable usage of swear words in computer-mediated communication could support this claim. There has been, to my knowledge, no other study confirming or refuting these observations with detailed socio-demographic information to thoroughly understand their organization. Thus, the following question arises: has the prediction made by Thelwall in 2008 been fulfilled eight 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 thesis is thus twofold: first, it is to offer a better understanding of the patterns of swear word 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.

However, replicating earlier studies (e.g. Thelwall, 2008) was not an optimal solution for me as MySpace, the social medium on which some of these earlier observations were based, has suffered a considerable drop in activity and popularity since then. For this, and other

[2] I will further develop what "strong swearing" is in Part 1.

[3] Thelwall's results were based on the social network site MySpace.

methodological reasons (see Chapters 2, 4 and 6), I chose Twitter as a mode of data collection. With half a billion tweets emitted every day (at the time of this study) around the world, Twitter represents one of the most popular social media sites. This study is based specifically on a corpus originally composed[4] of just over eighteen million tweets issued by roughly 739 000 users (see Chapter 6). The corpus was populated with tweets by British users of both genders and from different age groups throughout the United Kingdom, as well as the Republic of Ireland, for practical reasons explained later. The geographic focus allows us to compare our data with earlier results of studies concentrating on the same region. Corpus linguistic methodology and tools have been used to address the sociolinguistic issues raised earlier (see Chapters 7, 8 and 9). Also, because Twitter does not provide us with a direct access to the gender or the age of the users, using computer-programming methods has been necessary to be able to study these age and gender differences (see Chapter 6).

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 now have compared to the limited impact they had when these earlier observations and predictions were made. According to a study from Ofcom (see the 2013 Ofcom report[5]), the time we devote to social media sites is growing every year among people from all age groups and all socioeconomic backgrounds (see also Smith & Brewer 2012). This thesis 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 this end, this thesis is divided into three main parts, each focusing on one of the main areas this study relies on, namely the review of the literature, the description of the methodology used, and the results. These three parts are in turn, composed of three chapters each, which are divided as such:

Part one

Chapter 1 introduces the main notions this thesis relies on, namely the notions of gender, swearing and social media, and particularly focuses on the relations there are between these.

Chapter 2 debunks some of the misconceptions about how women and men use language, or rather, how they are expected to use language according to some of the gendered stereotypes.

[4] That is, before any gender or age detection has been carried out, as I will explain later.

[5] Last accessed on June 27th, 2017. URL: https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/children-parents-oct-2013

Chapter 3 will be an opportunity to do an in-depth review of the foundations on which the branch of linguistics commonly referred to as "corpus linguistics" is based. This approach will be central in the analyses I will later present, so reviewing key concepts is necessary to clearly understand the methodological choices made.

Part two

Chapter 4 presents the advantages of using Twitter data compared to data from other social media sites. It will also be an opportunity to give more details about how the Twitter interface works, as it is of key importance in understanding how I have had access to my dataset.

Chapter 5 provides an in-depth review of the online tool used to collect the data. This interface being a central element of this study, it is necessary to understand its framework, and more importantly how I used it, in order to highlight the advantages, but also the potential drawbacks of the methodology I used.

Chapter 6 details how I managed to infer the age and the gender of the Twitter users whose tweets I collected. These two sets of information not being openly provided, I had to resort to other statistical and computer programming tools to carry this out.

Part three

Chapter 7 gives overall data on the corpus I collected. This data ranges from basic statistics regarding the number of users inside each age group, to more detailed ones regarding swear word count, and rankings of the most used swear words according to age and gender. The chapter provides an overview of how the corpus is organized, and how frequently or infrequently swearing occurs.

Chapter 8 goes more in-depth into the data, and explores which swear words are statistically more representative of each gender and age group thanks to various tests like the Mann-Whitney U test or the simple maths parameters. This chapter is an opportunity to analyze the differences there are between each gender and age groups, but also focuses on what is similar.

Chapter 9 combines both quantitative and qualitative analyses in order to focus on specific cases highlighted in earlier chapters as being representative of certain trends. This will be a way to confirm or refute the observations made earlier, and better understand their intricacies. The exploration mainly revolves around collocational analyses made possible by the LancsBox tool, and gives way to comprehensive accounts of swear word usage in tweets chosen as being characteristic of certain (sub-)groups of users.

PART 1: THEORETICAL FRAME

> Swearing constitutes a species of human behavior so little understood, even by its most devoted practitioners, that an examination of its meaning and significance is now long overdue. The temper of the times in which we live having grown somewhat more complaisant, a consideration of this once tabooed topic may not be considered out of joint.
> (Montagu: 1967)

This quotation perfectly illustrates the idea that an evolution in the way swear words are used and perceived is not new: the fact that Montagu states that swearing was a "once tabooed topic" indicates that researchers already started to perceive the necessity to analyze this further several decades ago. However, even if this need was felt at the time Montagu wrote this, the literature on whether there were different degrees of appreciation of swear words was sparse, and as Baudhuin (1973: 399) said, "[e]xcept for the studies by Baudhuin and Bostrom and Rossiter, however, no empirical investigations have been reported which dealt with the degree of "tabooness" or "objectionability" of various obscene words". Contrary to what is still regularly asserted nowadays and as we will see, a lot of material is now available on the usage and perception of swear words. Nevertheless, like any other linguistic variable, profanity is in constant evolution, and new patterns of usage keep appearing. As mentioned earlier, the patterns I am interested in are related to gendered uses of swear words online, and more specifically on Twitter. However, to fully grasp the implications that certain attitudes may have, and thus to better interpret my own results, it is necessary to review what has previously been demonstrated in this area of research. This step is also important to assess the methods used before, to replicate what has proved effective, and to improve upon features which were insufficient.

In the first chapter, I will introduce basic concepts presenting how the key notions dealt with in this thesis, namely swearing, gender and social media, are considered and how the three interact, in order to build from that for my own analysis of the topic. Women and men have often been considered as separate entities as far as language is concerned, which led to the creation of the notions of "women's language" and "men's language". I will show why these categories are no longer viable and show that swearing is not reserved to men only, be it in face to face interactions or online.

In the second chapter, I will show how, on top of actually not swearing less than men, women may be starting to swear more, and especially on social media. This will also be the opportunity

to go in-depth about the reasons why I decided to focus on Twitter for my analysis, and on the social, demographic, and linguistic reasons why this medium is more interesting than others.

In the third chapter, I will detail the kind of approach I chose in order to study language and gender, i.e. one which would traditionally be labelled as “corpus linguistic”. I will review the advantages that the literature on the topic has highlighted, and how they fit my objectives and motivations. I will also present what I am going to consider as a swear word for this study, and in particular how I selected the words in question and how they fit the population and environment I will focus on.

Chapter 1: General concepts

As mentioned in the introduction, this thesis deals with several notions (swearing, gender, social media etc.) which are analyzed in relation with each other. Contrary to what it may seem, clearly defining these notions is not easy, and as I will explain, their relations with each other make this context of study a specific one. In order to clearly understand the implications of such a specific context, it is necessary to provide details regarding how these notions have been defined before, in order to build from that and understand how they need to be approached for this particular study. To try not to isolate each of these aspects, this chapter will try to account for the intersectionality of these notions as much as possible.

In section 1.1.1 then, I explain how swearing has been defined before, and I show that it is difficult to provide a clear definition of what a swear word is, and that context is what matters most in this regard.

In 1.1.2, I present the notion of "separate worlds" which has been used to describe how differently women and men talk. This notion is key in that it allows to understand that the way women and men talk has long been considered as two distinct entities, which further increased some of the (linguistic) inequalities existing between women and men.

In 1.1.3, I give details regarding new ways to analyze gendered speech patterns which emerged a few decades ago, and which enabled to nuance some of the early distinctions made concerning how women and men use language.

In 1.1.4, I explain how the development of these new methodologies, on top of nuancing some stereotypes, may actually lead to the cancellation of the very idea of "women's language" and "men's language".

In 1.1.5, I go in depth on the topic of the expression of gendered identities on the Internet, and how this mode of communication may be more suitable to the denial of some of the pre-conceived ideas about language and gender.

1.1.1 What is a swear word?

According to the Oxford Dictionary of English[6], a swear word is "an offensive word, used especially as an expression of anger". Although, as I will show later, swearing can be used in many more contexts other than in "anger", this is in line with McEnery's definition of bad language (2004: 1, 2), which is considered as "any word or phrase which, when used in what one might call polite conversation, is likely to cause offence". According to these definitions then, a lot of offensive words come to mind, on which many people would probably agree, like *fuck*, *shit*, *cunt*, *bitch* etc… Let's now consider this sentence: "You'd be a great fast-food clerk". If uttered in a casual conversation, to someone who is looking for a job, it may be taken as advice, and even as a compliment, but when said by a supervisor to a PhD student explaining their progress over the writing of their thesis, this may be offensive, while not containing any of the type of words mentioned above. Would "fast-food clerk" become a swear word then? It would, if we strictly stick to the definition of swear words given above. Now, let's imagine a barefoot person stepping on a Lego and crying "oh shit!" in front of a friend; is this friend going to be offended? If not, can we consider that *shit* is not a swear word in this context? Beers Fägersten (2007: 32) confirms the importance of raising this question by mentioning that in her study "context of utterance significantly affects the perceived offensiveness of swear words".

Context then, is what plays a major role in what will or will not be considered as a swear word, and therefore institutions such as the BBC come up with lists of words which, according to them, should be monitored or censored in programs according to various criteria like the time of broadcast, and thus the type of audience which is targeted for example[7]. This then constitutes a standard defining what a swear word will be in the context of audiovisual broadcast in the United Kingdom. However, even when having a defined list of words considered as swear words and rated according to their degree of offensiveness as in the case of the British Board of Film Classification (BBFC), clearly defining what should be labelled as a swear word is not easy. Indeed, despite these lists of swear words, the BBFC clearly state that context will greatly define what will influence the rating of a movie, in that "[s]trong language may be permitted, depending on the manner in which it is used, who is using the language, its frequency within

[6] See *Oxford Dictionary of English*, Oxford University Press, 2012.

[7] See the editorial guidelines of the BBC for more details on this. Last seen on December 5th 2016. URL: http://www.bbc.co.uk/editorialguidelines/guidance/strong-language/guidance-full

the work as a whole and any special contextual justification[8]". This example highlighting the importance of context applies to the case of the British broadcasting system, but also applies to other areas, as I will show later.

Sometimes, differences are also made between what swearing and cursing are, as in the case of Jay (1992) who distinguishes between cursing, profanity, blasphemy, taboo or obscenity, vulgarisms and expletives, whereas for other researchers, including me, these expressions may be used interchangeably (Mercury, 1995; Vingerhoets et al., 2013; Wang et al., 2014). I am choosing to do this because categorizing swear words according to their meaning or degree of offensiveness is not the main focus of this thesis, and I will focus on how these words are used rather than on how they could be labelled. Thus, as long as a word can appropriately be considered as a swear word in the context of this study[9], I will not make distinctions such as the ones referred to before.

Thus, different situations require different approaches and what applies in one given context may not apply in another one. This is partly what makes it difficult to establish a clear list of swear words which will apply to everyone and every situation, and this explains why the swear word classification used by the BBFC is different from the one used by Andersson and Trudgill[10] (1990: 15), which in turn is different from the one used by McEnery[11] (2004: 25) for example. The classifications made between swear words are therefore not necessarily universal and can vary depending on the resources used, the researcher's objectives, or simply because of the evolution of society. Indeed, new swear words are created, and some disappear or lose of their offensive character like *harlot* or *strumpet* for example[12], which can trigger the (dis)appearance of some categories, or the modification of others, to better account for these changes.

In my case, I am not going to try to classify swear words into categories labelling them according to what they refer to, as I will argue that what is more important is the situation in

[8] In the case of 12A/12 ratings. See the BBFC website for more details. Last seen on December 5th 2016. URL: http://bbfc.co.uk/what-classification/12a-and-12

[9] See Chapter 3 for more details regarding this.

[10] They used the classification created by Leach (1990) and which consisted of 1) words related to sex and excretion 2) words related to the Christian religion 3) words which are used in "animal abuse".

[11] He distinguishes between swear words, animal terms of abuse, sexist terms of abuse, intellect-based terms of abuse, racist terms of abuse and homophobic terms of abuse.

[12] See Villessèche (2016: 138).

which they are used, and that the pragmatic purposes and effects produced depend mainly on contextual and inter-individual factors than on which abstract labels can be associated to the swear word[13]. I do not imply that such characterizations are useless however, but in the case of a study aiming at better understanding how swear words are used by women and men on Twitter, these kinds of labels are not enough to account for the wide range of contexts in which a word can be used. For the purpose of this thesis then, it may be more appropriate to refer to functions of swearing, such as the ones described by Montagu (1967). Montagu mainly made a difference between annoyance swearing (for personal purposes, like catharsis[14]) and social swearing (for inter-individual purposes), which more accurately represent the reality of swear word usage, and this is what I will discuss in Part 3 of this thesis. Swearing, then, can be "social", in that it can be a linguistic projection of social parameters in a given social context. Gender is one of these parameters, and it is argued that the gender of a person, as any other social aspect, will influence swear word usage.

1.1.2 "Separate Worlds Hypothesis"

> According to the separate worlds hypothesis (SWH), biology is not destiny, but it is social grouping by gender that produces results that look like genetic bias, as if males and females create separate subgroup cultures.
> (Ervin-Tripp, 2001: 135)

This concept refers to the idea that men and women speak different "languages". This is what Lakoff (1973) develops when she tries to define what is "women's language". According to the separate worlds hypothesis, men and women develop differing perceptions of the world and ways of speaking partly because of the contrasts which are made between girls and boys during childhood, and because these are rooted in children's practices as same-sex interactions are favored.

Swearing is one of these features which has long been considered a male-only characteristic (Bailey, 1985; Wentworth, 1975). Descriptions of "correct" ways of speaking can even be traced several centuries ago, and it is possible to find comments about the language which was acceptable for young women in the Tudor period. Vives (1523) *De Institutione Christianae*

[13] I will detail what I am going to consider as a swear word, and which parameters I will take into account in this regard in Chapter 3.
[14] See also Goffman (1978) for more details on catharsis and swearing.

Feminae ("On The Instruction of a Christian Woman") compiled observations on what was considered appropriate language for women at the time[15]. Less than a century ago, Jespersen (1922) described women's speech and mentioned their "instinctive shrinking from coarse and gross expressions and a preference for refined and (in certain spheres) veiled and direct expressions". More recently, Wentworth (1975: xii) explained that according to him, "most American slang is created and used by males".

Swearing is, by its provocative nature, considered as an act of power, and this may be why women have been denied it a long time. During the Victorian era, being a woman, and being a lady especially was closely linked with speaking "properly"[16]. Social status played a key factor in the linguistic expectations of both genders, and the public/private dichotomy determined the spheres where women and men would have power:

> Since the private sphere is dependent on its place in the public sphere, the domestic woman's ultimate position in the social order is dependent on the place of her male relatives' position in the marketplace. And her ability to exert power and influence in the private sphere depends on how these men allocate the goods that they gain in the marketplace.
> (Eckert and McConnell-Ginet, 2003: 38)

According to this quotation then, we can hypothesize that swearing was reserved to men only because women were not "allowed" to express any kind of (linguistic) control outside of the sphere of home. Males exert power on the public domain, on everything influential, and female power is exerted on the private sphere. This is why, when she dealt with swearing, Lakoff claimed that "[t]he decisive factor is less purely gender than power in the real world" (1973: 57). She meant that the whole point of swearing, and other linguistic features which index positions of power are associated with men, who have, traditionally, had more access to most forms of power. Kira Hall's explanation of Lakoff's view can also help to understand the point that the real point of *Language and Woman's Place* (LWP) is more power than gender:

> [T]he language patterns of hippie, academic or homosexual so often appear to resemble those of the American middle-class housewife. That these disenfranchised groups are likely to use some of the same specialized lexical items as American middle-class women, she argues, points to a more general conclusion: "These words aren't, basically, "feminine"; rather, they signal "uninvolved," or "out of power"". […] While certain patterns of speech may be considered feminine because

[15] For a detailed analysis of these observations, see Juan Luis Vives, *The Education of a Christian Woman. A Sixteenth-Century Manual*, 2007.

[16] See Romaine S., in Holmes and Meyerhoff (2003: 104).

> women are, in her own terms, the "uninvolved," "out of power" group par excellence" (LWP 47), Lakoff is careful to note that any group in society may presumably use patterns associated with "women's language".
> (Hall, in Lakoff, *Language and Woman's Place*. 1975. New York: Oxford University Press, 2004. Print.)

Descriptions of attitudes which are reserved to men and women are not contemporary then, and in the early research on gender differentiation, the problem is that women's patterns were considered as deviant, and men's as the norm, and this is why men have for a long time been the only ones taken into account for linguistic studies (Chambers and Trudgill, 1980). Culture then, and the ideas which have been associated with it for generations, influence the perception of what will be acceptable or not, and often, this influence favors men's status. These societal distinctions eventually have an impact on men's and women's speech, as well as on the way speech will be perceived, hence swearing is still sometimes considered a male feature. As Coates (1991) explains, "the folklinguistic belief that men swear more than women and use more taboo words is widespread". To illustrate this, on Monday 11 June 2012, a Tehran cinema was shut because women were sold tickets for public screenings of the Euro 2012 football games. The reason invoked was that "[m]en, while watching football, get excited and sometimes utter vulgar curses or tell dirty jokes. [...] It is not within the dignity of women to watch football with men[17]". In this case then, men seem to have a natural right to swear, whereas women must avoid it, thus fueling the Separate Worlds Hypothesis. Furthermore, Vivian De Klerk (1991) studied two groups of teenagers from two different schools, and according to what they reported, she found out that teenagers were, generally speaking, more tolerant vis-à-vis swear words than adults were. De Klerk (1991: 164, 165) gives an explanation for this, stating that in the adult system, "[t]he overt, positively reinforced attitude is that swearing is frowned upon", and she continues by explaining that swearing may thus be a way to reject adult authority, break taboos, and affirm oneself as a member of the teenager community. This may seem out of the topic of profanity and gender since I am not dealing with girls and boys specifically, but if we take for granted the fact that teenagers swear more than adults to affirm themselves, we acknowledge that swearing is an act of power enabling oneself to break from the norm and gain authority. With this in mind, we can easily remember that "the decisive factor is less purely gender than power in the real world" (Lakoff, 1973: 57), and relate it to swearing,

[17] From Tehran police. Last seen on 21 Nov. 2016. URL: https://mic.com/articles/91067/iran-s-government-says-women-should-be-thankful-they-can-t-watch-the-world-cup-in-public

its authoritative power, and the reason why women have so long been looked down on when using this linguistic device.

Thus, if we acknowledge that female speech patterns are not purely explained because of their biological sex, but instead because of power, which may vary according to different contexts such as social status, addressee and so on, we realize that the situation is more complex and that there is a lot more depth to take into account in the study of gendered speech patterns. In order to study these situations without being influenced by pre-conceived ideas on how women and men ought to speak, a slow process involving the development of new approaches and methods had to be put in place, which I am going to review.

1.1.3 New approach: new results?

A study which can be considered one of the first articles investigating gender differences and their relation to profanity was written by J. M. Steadman (1935). His survey was carried out to analyze his college students' differences in obscene words usage. He asked 166 men and 195 women to make a list compiling as much taboo speech as they could. The informants had to classify the data into three categories: coarse or obscene words, words of a sinister or unpleasant suggestion, and innocent words. When describing the results he obtained, he reported that women, "of course, handed in less objectionable words than the men" (1935: 94). What can be a problem with such an assertion is that the researcher seems to take for granted certain characteristics generally applied to women, without trying to find other explanations for the data he obtained. When presenting things in such a deterministic way, the fact that women reported knowing fewer coarse words seems to be limited to the stereotype of the woman being a "swear word eschewer", without adding any scientific justification to support these claims. The point is that other explanations could be found to explain these results. They may have reported fewer expletives not to be stigmatized, or perceived as tomboys for example. Indeed, the fact that Steadman asked his own students to report what they knew about profanity could provoke a certain fear of the way they could be perceived, since they probably regularly saw Steadman, as he was their teacher. Moreover, we are not sure that the survey was completely anonymous, which would dramatically increase the risks of self-edition in order for the students to pass off as different from what they really were. In this case, the problem is not so much to know whether women swear more than men or the other way around, it is about rendering a truthful image of the sociolinguistic reality at work, and letting pre-conceived ideas and

stereotypes drive our interpretations is not ideal for research purposes[18]. More generally, and as McElhinny (2003: 34) pointed out, the problem with certain remarks and analyses "is that it is not at all clear that the characterizations which the investigator makes are those which are grounded in the participants' own orientation in the interaction". This can be linked with the "Hall of Mirrors" theory[19] which can sometimes influence such conclusions. Therefore, the reason why women reported knowing fewer coarse words than men in Steadman's study may not necessarily be due to their gender.

In other words, the problem with earlier analytical studies on language and gender is that stereotypes and pre-conceived ideas could sometimes guide the researcher's view of things and thus weaken the conclusions. Indeed, when the first dialectological researches began, a lot of fieldworkers based their data on men only, because they were believed to better preserve the original forms of regional dialects. This stereotype is another reason why Sapir (1929) presented women's speech as being derivational compared to that of men. This is despite the fact that the study Gauchat carried out in 1905 already hinted at the fact that the beliefs researchers had of linguistic dissimilarities were not exact, and that men were not necessarily users of "pure forms", as it was believed. It is Labov's work (1966), which promoted the use of new techniques of investigation, and highlighted drawbacks in classical dialectology. Then, from the 1970s onwards, gender and swearing started to be investigated more accurately, with more suitable technology and methodology. For example, Burgoon and Stewart (1975) analyzed gendered interactions, Bailey and Timm (1976) looked at the effects of gender and age in different (social) contexts to see how swear words were used, Staley (1978) analyzed gendered expectations in the use of swear words, Holmes (1984) offered a functional approach regarding the relation between sex and language. These studies contributed to empirically widen the knowledge researchers had of the links between language and gender, as well as to reinforce the methodology used to analyze these features. One of the new techniques which

[18] About this, see also Beers Fägersten (2012).

[19] See McConnell-Ginet in Holmes and Meyerhoff (2003: 81): "Even when each individual researcher has made only modest claims on the basis of individual studies, the combination of the sheer volume of studies and the ambient belief that the results should be positive, have led to a general impression of robust findings. In the end, then, the stereotypes are accepted as scientific fact and become part of the background of general truth about language and gender".

appeared was Conversation Analysis. For Weatherall and Gallois, this analytic approach may be a way of escaping all the stereotypes pervading language and gender analyses:

> Many gender and language studies assume that participants have an intellectualized gender identity and that people's speech is somehow related to that identity. [...] Taking a more conversation analytic approach means not treating identities as a kind of demographic or psychological facts whose relevance to behavior can simply be assumed. Instead of asking about the strength of gender identity or the kind of contexts where that identity is salient, the focus is on whether, when, and how identities are used.
> (Weatherall and Gallois, in Holmes and Meyerhoff, 2003: 500)

This quotation illustrates the fact that analytical studies, and dialectological research more generally, had to be improved and that more modern techniques of investigation would probably be more suitable. Weatherall and Gallois also refer to the fact that gender identities are changing and contextual, and that the gendered dichotomy which has long been assumed is no longer a standard, thus a more critical approach must be adopted, and this is one of the aspects advocated by Conversation Analysis (CA). CA is a branch of sociolinguistics which probes into the patterns of a face to face conversation in order to understand the relation between question and answer, or study the use of tag questions, overlapping, and so on... CA is not the only modern development which improved sociolinguistic research, but this is a good example to emphasize the reversal in the way speech, swearing, and gender were apprehended. The point is that for a long time in gender and language research, the gender of the participants itself has been considered as a factor influencing linguistic choices, and analysts often accounted for certain variables only by taking their gender into account, without trying to understand if other factors may have acted on it. So, conversation analytic studies do not aim at endorsing the "truth" of any explanation, but rather to identify the different statements given and consider the possible contradictions. CA is one example of the attempts to improve the methodology used to analyze speech, but this has been an overall general direction taken by most branches of linguistics.

These improvements then gave birth to a new approach of the way studies in linguistics had to be carried out. For Kira Hall, the direction that linguistic research on gender must take is one that "seek[s] not to describe how women's language use differs from men's, or how homosexuals' language use differs from heterosexuals', but to document the diverse range of

women's and men's linguistic repertoires as developed within particular contexts[20]". Rather than a mere inventory of the speech patterns of certain groups or sub-groups then, a contextual study of individuals or groups of people may be more fruitful. In this case, the researcher not only focuses on what is produced, but on the reasons why it is produced, which is particularly interesting for the purpose of my sociolinguistic study, that is, trying to better understand how similar or different women and men are in their use of swear words.

Using a different approach, Selnow (1985) carried out a study in which he submitted a questionnaire to 135 undergraduate students. I will take a closer look at this, and two other studies to compare their results and try to see an evolution in the way profanity was used, and especially how it was considered. Selnow's study mainly aimed at analyzing five points, which can be compared to some of the points I wish to analyze in my own study. First, he wanted to see if there was a measurable difference in the use of profanity men and women reported. Then, he wanted to see the contexts in which men and women believed it was appropriate to use profanity. He tried to analyze if women and men used profanity with differing goals in speech. He also wanted to see if the respondents' backgrounds could influence their perception and use of swear words. Eventually, his survey aimed at analyzing if women and men had differing perceptions of profanity. The overall results were that female respondents generally reported using profanity to a lesser degree than men. Female respondents also generally believed that in most of the contexts stated in the questionnaire, the use of profanity was less appropriate than males did. According to Selnow's results, all respondents disagreed with the proposition that "the use of profanity serves to demonstrate social power" (Selnow, 1985: 308), even if men disagreed less strongly than women did. About the results concerning the use of profanity by relatives of the respondents, fathers were generally reported to use more profanity at home than mothers, but, female respondents generally reported a much higher use of profanity than male respondents. Another interesting point in this study is the fact that "while women rated excretory and sexual profanities about the same as men did, it was men who rated religious profanities most severely" (Selnow 1985: 310).

What is worth noticing in Selnow's study is that the results concerning the opinions and perceptions of swear words for men and women of this study carried out in 1985 are similar to

[20] See Hall, in Holmes and Meyerhoff (2003: 375).

the results of Karyn Stapleton (2003), that is, almost twenty years later. Indeed, what stands out in Stapleton's research is that men and women all have more or less the same perception of profanity, even if their reported use of swear words seems to differ according to context. Even if it was clearer among female participants, a majority judged the vocabulary relating to female anatomy as vulgar. What differs however is the quantitative use of such terms. This may be what Selnow's results imply when it appears that women use profanity less than men, but it cannot be corroborated since Stapleton referred to the denomination for female sexual organs explicitly, and we do not know exactly what Selnow took into account. The data of a third study from De Klerk (1991) revealed great consistency in the results of sex-based groups. No matter the age, or the kind of school, the general tendency was the same. Generally speaking, boys displayed a greater tolerance vis-à-vis profanity than girls, which can be linked to the data from Stapleton's study (2003), in which she found that what differed was not so much the perception of profanity, but the degree of legitimacy men and women had of the use of profanity. Men deemed that it was more acceptable for men than for women to use terms referring to female anatomy for example. In her study, De Klerk calls that an "apparent male self confidence" (1991: 164).

What is important is the fact that even if the work from Stapleton was based on a community of practice (so the generalization of the patterns observed can be more limited), this so-called "apparent male self-confidence" is still observed some ten years after this study by De Klerk. Even if a study based on a community of practice is relatively specific, what the males from Stapleton's study suggested is that it is more acceptable to hear certain words from men than from women, so it can be asserted that this pattern does not seem to have lost of its influence over the years as the results are consistent. However, De Klerk also found that "[m]ost groups had lower tolerance towards women and children who swore" (1991: 164). This, on the other hand, shows a difference between the perceptions of profanity between the informants from the two studies. Even if we find a consistency in the "male self-confidence", the fact that De Klerk states that "most groups" are less tolerant concerning women and children swearing, contrasts with Stapleton's informants' claims. What is worth looking at in more details however is what Jenny, one of Stapleton's informants, said about a supposed legitimacy of the use of swear words: "if a word is wrong to begin with, then it doesn't matter who says it--it's still wrong. I just don't put up with people saying that sort of stuff anymore". What is interesting is the equalitarian aspect of her statement, which did not seem to be salient in previous analyses like the one from De Klerk (1991), who suggested that both male and female respondents seemed

to believe that it was more acceptable for a man to swear. It can be argued that Jenny's statement is only one isolated opinion and that it cannot be relevant to a whole generation, but Stapleton adds that it is not just Jenny's opinion, but that "the women in this study largely rejected any such notions of gender differentiation".

What can be concluded from this analysis of these three studies carried out over a period of twenty years is that the rise of new methods of investigation, and the adoption of a more analytic approach, enabled researchers to be more objective, consistent, and nuanced in their findings. The observations made in this section about the need to shift from focusing on gender itself to a contextualized approached can be summed up by one sentence from Bamman, Schnoebelen and Eisenstein (2014: 139), who analyzed the speech patterns of Twitter users, and concluded that the interpretation of their results "leads to anti-essentialist conclusions: gender and other social categories are performances, and these categories are performed differently in different situations". This notion is also found in many other studies (see for example Eckert, 2008; McConnell-Ginet and Corbett, 2013; Ochs, 1992; Schiffrin, 1996), which shows the agreement there is among researchers on this concept.

1.1.4 "Women's language": just another myth?

Thanks to these new approaches and the numerous studies which followed Lakoff's work (1973), it can now quite confidently be asserted that this women's language is another stereotype (see below) linked with how women behave. Actually, "a meta-analysis by Hyde (2005) of several hundred studies of verbal and behavioural gender differences concluded that most of the studies found that the overall difference made by gender was either very small or close to zero[21]". This does not mean that Lakoff was wrong, actually she was completely right because this women's language does exist, it is in everyone's minds. An example of this is the study that Kramer (1974) carried out on cartoons from *The New Yorker*, and which revealed that pre-conceived ideas existed in her informants' minds, who more freely associated vulgar captions taken from cartoons with male characters. De Klerk (1992: 280) also confirmed the existence of these stereotypes, as she explains that "[t]he consistency of opinion across all groupings of informants was remarkable, and rating results highlighted the profound influence of stereotypes on attitudes. Young adolescent males were seen as the most appropriate slang users by all informants, which is highly suggestive of what the "popular myth" is". A last

[21] From Baker (2014: 19).

example of this could be Edelsky's study (1976), in which children of various ages were presented with words in context and were asked to rate whether the words in question are more likely to be used by males or females. Her test children were aged 7, 9 and 12, and a growing sensitivity to gendered stereotypes can be felt as they grow older. When analyzing the results from Edelsky's study, Coates (1986: 131) reports that:

> At 7 years, only two variables get a consistent response: *adorable* is judged to be female, and *Damn it!* is judged to be male. At 9 years, this has increased to eight variables: *adorable*, *oh dear*, *my goodness*, *won't you please* are judged to be female, and *damn it!, damn* + adjective, *I'll be damned* are judged to be male (tag questions get a neutral response). At 12 years, the child judges agree on assigning every one of the twelve variables to one sex or the other: tag questions, *so*, *very*, *just* are added to the female list, and commands to the male list.

So, this cultural stereotype exists and even starts to be influential at a young age. What is sure now, on the other hand, is that the foundations on which it is built, the prototypical ideas that a lot of people have that, for example women swear less, use certain color adjectives more than men, and use more tag questions than men because of their embedded uncertainty[22], are not founded. This is echoed by the study from Bamman et al. (2014: 136), among others, who mentioned that "previous work has focused on words that distinguish women and men solely by gender. This disregards theoretical arguments and qualitative evidence that gender can be enacted through a diversity of styles and stances". These notions have also been expressed in other studies (see also Bourdieu 1977; Sewell, Jr. 1992).

To illustrate this, I will take the example of tag questions, which has been one of the most investigated topics[23] because it was one of the most popularly believed as belonging to female speech. Some studies (especially the one from Cameron D., F. McAlinden and K. O'Leary (1989) cited above) showed that, contrary to what was widely spread, tag questions may signal attitudes other than simply lack of self-confidence and uncertainty. They can be used by a speaker to indicate their involvement in the conversation, such as backchanneling[24], and to show a certain interest in what is being said, or to mark social solidarity. These studies also showed that men could be more likely to use tag questions to express uncertainty than women.

[22] See Lakoff (1973) for a list of such stereotypical representations.

[23] See Dubois B. and I. Crouch (1975), Holmes J. (1984), Cameron D., F. McAlinden and K. O'Leary (1989) for example.

[24] Backchanneling refers to the linguistic devices and strategies such as "uh-huh", "yeah", "really?" used to signal that one is listening to what a speaker is saying. See for example Eckert and McConnell-Ginet (2003: 111) for more details about backchanneling.

William O'Barr and Kim Atkins (1980) also wanted to further explore Lakoff's claim that women's language was "powerless" and "ineffective". They analyzed courtroom testimonies and they found that it was not gender, but the social position which was more likely to predict the use of "women's language". O'Barr and Atkins also played the same testimony to jurors, except that in one case they played it with people using "women's language", and in the other case, with people using a more direct style, attributed to people with a certain authority. The result was that jurors were more likely to believe the testimony in the second case.

Christopher J. Zahn (1989) especially, also showed that the use of a so-called "powerful language" is most of the time not related to gender and has more to do with parameters such as the social situation, the occupation and so on. This would mean that Lakoff's association of "women's language" with gay men, academics and hippies could be generalized to virtually anyone and any social category in a situation of powerlessness. Thus, it means that every person with little authority could use "powerless language" (which would then be more accurate than "women's language"), as it is more likely that these people will project their social disempowerment, and not their gender, through speech.

Thus, evolutions in the interpretation of the data, data collection, and technology allowed researchers to study gendered linguistic differences in a much more reliable way. Recent technological advances have enabled linguists to collect, store and analyze vast amounts of data much more easily than ever before. This, among others, led to the creation of the British National Corpus (BNC). The BNC is a corpus of 100 million words of written and spoken British English collected in the 1990s[25]. This corpus is composed of texts from newspapers, academic journals, books, as well as transcriptions of spoken speech. The BNC is still considered nowadays as a reference providing an authoritative snapshot of what British English was like in the 1990s. These kinds of corpora are very interesting in that they allow to draw conclusions which are more generalizable, as their size and heterogeneity should provide a greater objectivity. In a study from Schmid (2003) based on the BNC, we learn that "it did appear that males and females were using language in stereotypically gendered ways - males were more likely to exploit a lexicon associated with public affairs, abstract concepts and sport while females used more words referencing clothing, colours and the home" (Baker, 2014: 21). From there, and because of the advantages of modern reference corpora I just cited, we could

[25] See http://www.natcorp.ox.ac.uk/corpus/index.xml for more details about the corpus. Last seen on November 22nd, 2016.

be tempted to conclude that this is an empirical proof of the existence of "women's language" corresponding to some of the pre-conceived ideas mentioned earlier. The *type of data* used in any study is key in having an objective representation of the (gendered) speech patterns of a panel of informants. However, the *methodology* used to collect the data may be even more important to have a reliable view of the corpus. Concerning the methodology used for the BNC data, Baker (2014: 28, 29) explains that:

> once we start to consider the context that the BNC spoken data was collected, we find an explanation for the trends towards sex difference, which raises a question about the validity of such difference. [...] Of the 320 speakers in the F1 group 261 (81%) had their conversations recorded in private settings (being tagged as 'demographic' as opposed to 'context governed' which was used for public and workplace settings). For the M2 group, of the 618 speakers, only 18 (3%) are from private settings. The larger F1-M2 difference then, is more likely to be telling us more about how people speak at work, as opposed to at home, rather than actual male-female differences.

This shows that, when analyzing the data of a corpus we must be careful about the way the data collection was carried out if we want to ensure that the conclusions drawn are reliable and representative[26]. Concerning gendered uses of swear words, the BNC, and the analyses based on it, also provide some interesting material, and Baker (2014: 34) once more highlights meaningful patterns with regards to swear word usage:

> [t]he results showed that of the 7,023 cases of these words in the corpus, they are relatively equally distributed between males and females. Males say them 888.3 times per million words while females say them 828.29 times - quite a small difference. What about dispersion? Of the 1,360 females, only 250 (18.3%) use these swear words, while 381 of the 2,448 males (15.5%) use them. Again, this is quite a small difference, although it is also interesting (and perhaps unexpected) that these words are relatively more dispersed among female speakers than males. And the 'overlooked' pattern here is that the majority of both males and females *did not swear*, at least when their speech was recorded for the corpus.

This last quotation is very important, because it may be the projection of a methodological problem in research on language and gender, and a sign that the Hall of Mirrors theory may influence researchers more often than we may think. We cannot deny the facts however, and when in a study carried out in the exact same conditions for a representative number of men and women, differences that occur and are still statistically significant cannot but to be noticed and focused on. But, it does not mean that these differences must be the sole reason to claim that men and women speak different "languages", especially if the linguistic features

[26] See Part 2 for more details regarding the importance of data collection and analysis.

highlighted remain minor compared to the whole array of linguistic resources that are similarly used among these very women and men. Thus, these examples show two important things:

- When looking at swear word usage quantitatively, and on a large scale, women and men swear as much as one another (the difference in swear word usage displayed here not being actually representative of any real difference), thus going against the idea that swear words are a characteristic of male speech. Although here again, it could be argued that the context of recording (i.e. the home/work difference) may bias these results as well.
- What should instead be focused on is not *whether* swearing is part of male *or* female speech, but rather the fact that not swearing *is* a characteristic of *both* genders in most of the cases we have analyzed so far, and that women and men are more alike in this regard.

We have so far been focusing on gendered differences in various face to face contexts to have an overview of the results provided by research on the speech patterns of women and men. We are now going to turn more specifically to the context which will be the center of our attention in this thesis, i.e. social media, and see how they can be considered as a context in themselves, and how interesting this can be as it generates a multimodal "neutrality" providing various advantages for sociolinguistic purposes.

1.1.5 Gender on the Internet: the new neutral?

Social media (and Twitter in particular) are interesting for research because they are nowadays equally used by women and men from various social backgrounds[27], thus limiting some potential sampling bias. In the case of Twitter then, the panel of potential informants offered (i.e. the users) is neutral in the sense that it provides a relatively equal demographic representation of users, in a context shared by all users, that of information and opinion diffusion[28]. This medium also seems gender-neutral in the way people (i.e. women and men) express themselves, for reasons which I will return to later, but in order to clearly understand this, it is necessary to come back to the importance of context when analyzing gendered speech patterns. The previous example from Baker (2014) and the BNC implied that the context of recording could have a major effect on the kind of speech which will be produced, and as such,

[27] See Chapter 2 for more details regarding the demographics of Twitter users.
[28] See Kwak et al. (2010) and Hughes et al. (2012).

we may wonder if context could play a bigger role than gender itself in deciding what variables will be displayed more often. In this regard, the study from Bamman et al. (2014: 148-149) can once again help us understand the importance of the context of social media, as they state that:

> All of the male-associated clusters mention named entities at a higher rate than women overall, and all of the female-associated clusters mention them at a lower rate than men overall. The highest rate of named entities is found in C13, an 89 percent male cluster whose top words are almost exclusively composed of athletes and sports-related organizations. Similarly, C20 (72.5 percent male) focuses on politics, and C15 focuses on technology and marketing-related entities. While these clusters are skewed towards male authors, they contain sizable minorities of women, and these women mention named entities at a rate comparable to the cluster as a whole — well above the average rate for men overall.

Here again, we have evidence that gender alone cannot be said to be enough to predict any quantitative use of certain lexical items, since according to the overall data, women generally use named entities at a much lower rate than men. But, in a context where named entities may be more likely to be used (technology and marketing-related entities in this case), female usage of named entities turns out to be at the same level as the usage of the cluster, and thus as men, whereas women basically were a minority to use them when aggregating the data. This confirms the pattern observed by Baker in the BNC with swear words and the fact that we need to go beyond the mere quantitative data, and look at every aspect of a study to better account for all the possible factors influencing the results observed. Additionally, and on top of showing that it can apply to other categories of words (in this case, named entities), it proves that even when being a minority to use a certain type of variables overall, one gender, when analyzed in a context favoring these very variables, use them as much as the other gender. This highlights the fact that it is not gender, but the choice of context which originally favored one gender over the other. This is important to note, because the fact that more women were recorded at home in the example from the BNC may, as Baker noted, tell us more about how people speak in certain contexts, but it did not give us details about whether women recorded at work would actually swear as much as men. The previous quotation from Bamman et al. on the other hand confirms that women use more of the variable in a context favoring it, just as men do, and thus confirms that gender may simply be one more variable, but that it may not be that alone which determines how people will speak. Thus, when analyzing how women and men speak, and in my case, how they swear, it is crucial to not only focus on mere quantitative and aggregate data. Instead, a more fine-grained approach has to be adopted, by looking at context for example, as Bamman et al., or Baker, did in these examples.

Another study about online gendered speech confirming the importance of context is that of Herring and Paolillo (2006), who analyzed the speech of women and men in web blogs, and concluded that:

> In this study of stylistic features claimed to predict author gender, we found genre effects, but no gender effect, in an analysis of entries in random weblogs. This leads us to propose that the functional requirements of the genres investigated—e.g. whether interactive or informative—lead bloggers to employ certain kinds of language, irrespective of their gender. We further propose that a more fine-grained genre analysis of apparently gendered language use in other communicative contexts might also show genre to be a conditioning factor, and that this approach should be pursued in future CMC research.

Again, this means that the type of blog, and not gender, decides which linguistic resources will be most used, based on the purpose of the blog. Thus, it is indeed not so much the linguistic resources which are meaningful here, because obviously, when one wishes to start a diary blog[29] for example, whether they are female or male, the linguistic resources used will most certainly be ones oriented towards information. So, the most meaningful thing to pay attention to in order to study gendered preferences has more to do with the type of blog (here, diary or filter), than about the linguistic resources themselves, and this is what Herring and Paolillo showed. In the case of Bamman et al. (2014) cited above, they explain the much more frequent mention of named entities by men by the fact that they generally prefer to talk about hobbies or career, and that these topics, i.e. contexts, are what accounts for the presence of named entities, and not a binary opposition of men as being more informative or explicit than women, as it has often been argued in older studies[30].

Bamman et al. (2014: 148) also applied this to swear words, as they grouped their Twitter users according to various clusters which they labelled with letters and numbers (e.g. A1, A2 etc…). When analyzing the patterns observed inside these clusters, they found the same kind of pattern, in that:

> [t]aboo terms are generally preferred by men (0.69 versus 0.47 per hundred words), but several male-associated clusters reverse this trend: C10, C13, C15, and C20 all use taboo terms at significantly lower rates than women overall. Of these clusters, C10 and C15 seems to suggest work-related messages from the technology and marketing spheres, where taboo language would be strongly inhibited.

[29] In their study, Herring and Paolillo focused on diary and filter types of blogs.
[30] See Bamman et al. (2014: 149) for more details on this.

On social media as well then, context seems to be a more important factor influencing the use of swear words (as well as any other lexical item) than gender alone, and only confirms what I showed before about the need to take other parameters into account when trying to make sense of statistical differences *and* similarities, between genders. Thus, no matter the kind of corpus one focuses on, and whether these are based on face to face interactions, literary texts, or online discourse, the need to pay close attention to context is paramount. Although it cannot be denied that women and men actually use certain words or grammatical categories more in certain cases, and that differences do exist, we need to be careful before attributing our conclusions to gender only[31]. Bamman et al. (2014: 154) sum this up perfectly by saying that "[w]hile the statistical relationships between word frequencies and gender categories are real, they are but one corner of a much larger space of possible results that might have been obtained had we started with a different set of assumptions".

Thus, it would seem that the online context could prove to be a relatively neutral place where women and men express themselves in a way that is very similar, at least as far as swearing is concerned, although a lot of the studies mentioned suggest that this could be generalized to more than just swearing. However, this neutrality has not always been there, and earlier studies on online gendered speech patterns seemed to reveal different results than more recent ones, and they "problematized claims of gender-free equality in cyberspace[32]". Earlier research on language and gender on the Internet mainly focused on the (relative) anonymity provided online. One of the goals was to see whether gendered differences would disappear when the gender of the person one is addressing was not as obvious as during face to face interactions, and the results were not convincing. Women and men were reported to diverge linguistically on several levels; men were reported to dominate interactions (Selfe and Meyer, 1991), be aggressive (sometimes sexually, to women) (Dibbell, 1993; Herring, 1999), post longer messages than women and be more vulgar (Sutton, 1994; Herring, 1992; Kramarae and Taylor, 1993). Concerning the male domination of the discussions, Herring and Stoerger (2014) explain that this feature was present "both under normal conditions and under conditions of anonymity", implying that the speech patterns would not be contextual and influenced by the gender of the interlocutor, but actually representative of inherent gendered patterns. An explanation given is that "gender is often visible in Computer-Mediated Communication

[31] See Bamman et al. (2014: 154) for a reference to this as well.

[32] Herring and Stoerger (2014). See their article for an interesting review of several early studies on language and gender online, as most of the related references in the following paragraph were taken from this article.

(CMC) on the basis of features of a participant's discourse style – features that the individual may not be consciously aware of or able to change easily" (Herring and Stoerger, 2014). This implies that women and men do have means of expressing themselves driven by their gender, and that these could be acquired and reproduced to "sound" male or female. Although the studies I mentioned so far tend to prove the opposite, one explanation to account for the gap between what early and recent research on online gendered speech patterns found could be the fact that roughly twenty years separate studies pointing to these two different positions. Thus, evolutions in the modes of communication, or simply in the accessibility of these media, influenced the modes of expression online. Herring and Stoerger (2014) also provide a review of studies of gendered online language carried out some time later, in the early 2000s, and the results already seemed to level out to some extent, even if the authors mention that "the research results are mixed". This time, the differences appeared to be only based on style, rather than on word choice and several other levels as suggested by earlier studies. Thus, although online anonymity does not seem to play any role in the linguistic patterns used by women and men, it seems that nowadays, social media are a relatively neutral place where the differences in the way women and men express themselves are much less relevant than they were.

As will be detailed in Chapter 3, the place, accessibility and influence of the Internet (and particularly social media) in our lives, has greatly evolved in the last few years. Thus, it is obvious that the status they had twenty years ago was radically different, which may influence ways of expressing oneself online. As we will see in the next chapter, it could also simply be that gendered speech patterns more generally, evolved and that now we are reaching a point where the gap between the way women and men speak (or write, in the case of CMC) has reduced to a point where it is no longer as visible as before. Indeed, Herring and Stoerger claimed that "gender is visible in CMC" (see above). So, if gender has been visible in online contexts before as in the case of Herring and Stoerger, the contrast with other results from more recent studies showing that gender is *not* visible may imply that online gendered differences may no longer be relevant.

Conclusion

I have established that even when using modes of expressions which are very different from face to face spoken speech, the speech patterns of women and men are nowadays actually much closer than what stereotypes predict. These observations also show that whatever the mode of expression, or genre used, women and men do express themselves in a manner which is more similar than different. Thus, in order to analyze gendered speech patterns, whether online or in face to face interactions, context of utterance should actually prevail over gender alone in order to make sense of the differences or similarities observed. I have shown that many gendered differences were reported in earlier studies, and that differences are still reported now, but that a more objective and empirical approach has mitigated these results and interpretations over time. This evolution may be twofold: i) evolutions in the methodology used to analyze gendered speech patterns more objectively probably encouraged more nuanced interpretations ii) as we have seen in this chapter, and as I will show in the next one, evolutions in the way women and men express themselves may also explain this change.

Last of all, I highlighted the fact that the differences observed should not be associated with inherent gendered traits, as patterns which are observed among women are also observed among men in certain situations and vice versa. It is a sign that we have to move from a purely binary gendered opposition to a mainly contextual one. So, although some differences remain, this nuances a lot of previous research on language and gender favoring a linguistic dichotomy between women and men, and I will have to build on these findings for my own methodology and the interpretation of my data.

Chapter 2: Swear words, people and social media: their interconnected evolutions

We live in an age where bad language can become worrying not because it is getting worse, but, paradoxically, because it is no longer bad enough.
(Harris, 1990: 421)

This quotation refers to an opinion which is not an isolated one; for several decades now, there has been a growing impression that swear words are becoming more present in people's speech[33]. Words which were reported to be barely whispered forty years ago can now be heard in the street or in popular TV shows. Are these words actually more present, or is it just a persistent impression? If this impression is confirmed by figures, does it mean that these words are now more accepted, or are they still as taboo as before? Answering these questions will help us understand the place that swear words have in our modern society, and analyzing the attitudes of women and men regarding their potential increase in the use of swear words is crucial if we want to depict the sociolinguistic situation we are in.

In 1.2.1. then, I review evidence that indicates context-specific increases in the frequency of swear word usage. However, and as we will see, certain words, or categories of words, are considered as more unacceptable than others, indicating that this evolution is not uniform.

In 1.2.2, I give specific examples, and focus especially on the word *fuck*, which is one of the words which has evolved the most in terms of its (un)acceptability. By focusing on previous studies which focused on this word in detail, I give hints as to what triggered this greater acceptability and frequency of use. However, I also show that, according to other findings, there may be what has been reported as a swearing paradox, and that swear words may in the real world not be as offensive as perception studies claim them to be.

In 1.2.3, I present studies showing how the Internet, and social media in particular, may be the place where swearing is the most common. These words seem to be more present on these media than in face to face conversations, or any other part of the Internet, and I try to understand why. Also, I show that social media seem to particularly encourage women to swear, and in a

[33] See "Are swears becoming so common they aren't even profanity anymore? F--- that!". Last seen on June 21st, 2017. URL: http://nationalpost.com/news/are-swears-becoming-so-common-they-arent-even-profanity-anymore-f-that/wcm/a1a81edf-ccc4-4dd4-817b-bd3d0b4045b8

See also "Expletive deleted". Last seen on June 21st, 2017. URL: https://www.theguardian.com/uk/2002/nov/21/britishidentity.features11

way which sometimes surpasses that of men, and we will see that this could be the sign of more profound social changes.

At last, in 1.2.4 I explain further why I chose Twitter as a source of data, and how interesting it can be in terms of volume, demographics and linguistics. I also explain the differences and similarities there are between the speech present on Twitter and face to face conversations, and how comparable they can be.

1.2.1 Is swearing more common than before?

The evolution of profanity follows the same path as the evolution of ways of living, as De Klerk (1992: 288) implies when she says that "[i]t is obviously not so much socioeconomic changes but shifts in social attitudes and lessening inhibitions that influence expletive usage". As the quotation in the introduction of this chapter suggests, in 1990, Harris perceived an increase in swear word usage and acceptability. Jonathan Margolis, a bit more than ten years later, confirmed that impression by saying that "today, any 12-year-old from the dodgiest comp to Eton would say fuck if they so much as grazed a knee, I doubt my dad would have said it even if a flying saucer landed on the patio and a Martian laser-gunned the shed[34]". Although he referred to the word *fuck* specifically, he implies that this increase in the use of swear words would be generational, and younger generations would be less likely to be offended by swear words as older ones. The two quotations presented so far both refer to an increase in the use of swear words, and to a weakening of their power to offend, and it seems that surveys confirm that these are not just impressions, but refer to an actual evolution of the offensiveness of swear words. Indeed, reports from the Broadcasting Standards Authority[35] (BSA) indicate that the responses from the survey they carried out "indicate a continuing softening of attitudes" (2013: 3). The first thing which can be noticed when comparing the data from 1999 and 2013 is the increasing tolerance regarding nearly all kinds of swear words. Detailed analysis of the data shows that even the words which are consistently rated as the most offensive are continuously considered as less offensive over time, with the most offensive one consistently being *cunt*, and the least offensive being *bugger*. Although this overall increase in acceptance is reported to be shared by both genders, it somehow seems that "[m]ales tend to be more accepting of the words or phrases than females" (2013: 12, 13), and the report gives detailed results by gender for the five words rated as being the most unacceptable (i.e. *cunt*, *nigger*, *Jesus fucking Christ*, *mother fucker*, *cocksucker*), which indeed indicate a greater tolerance from men[36]. However, despite the gap in acceptance (or rather, *unacceptance* in this case) between women and men, it should be noted that apart from *cocksucker*, both women and men are a majority to report the four most offensive words as being unacceptable. So, although a gap is indeed present, the most noticeable

[34] Jonathan Margolis, "Expletive deleted". The Guardian, on November 21, 2002. Last seen on December 12th, 2016. URL: https://www.theguardian.com/uk/2002/nov/21/britishidentity.features11

[35] This New Zealand agency has been carrying out surveys in 1999, 2005, 2009 and 2013 in order to monitor how acceptable or non-acceptable public audience finds the use of swear word in broadcasting.

[36] For the word *cunt*, 79% of the women found it unacceptable compared to 60% of the men. For *nigger*, 75% of the women found it unacceptable, compared to 53% of the men. 69% of the women found the word *motherfucker* unacceptable, compared to 53% of the men. 66% of the women found *Jesus fucking Christ* unacceptable compared to 55% of the men. 67% of the women found *cocksucker* unacceptable compared to 44% of the men.

pattern here is that women and men are a majority to agree that the words are unacceptable. What applies to New Zealand media also applies elsewhere, and this is the sign of a global phenomenon leading to a greater acceptance of swear words in all sorts of media, and this can be seen by the presence of swear words in successful TV shows (e.g. the American animated comedy South Park[37]), on Youtube (Beers Fägersten, forthcoming), or on online forums (Jaffe, forthcoming). However, digital media are not the only contexts where an increase in the use of swear words can be witnessed, as data from books reveal when Google Ngram Viewer is applied:

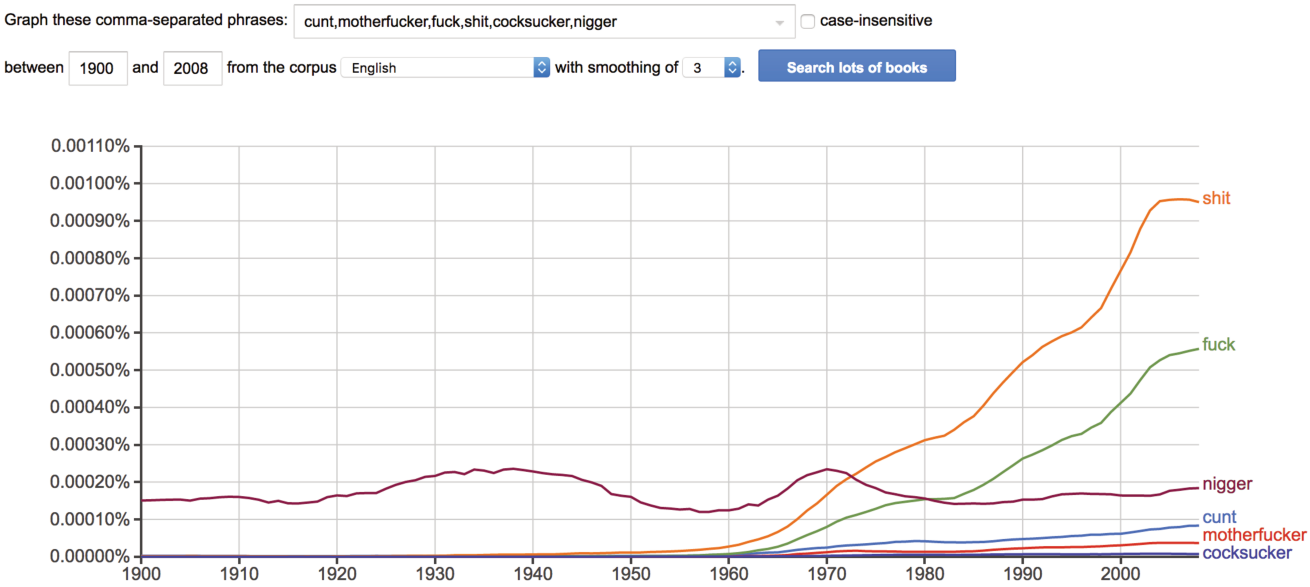


Figure 2.1: Frequency of occurrence of some swear words in the Google books database

This chart represents the evolution in the frequency of the selected swear words (i.e. *shit*, *fuck*, *nigger*, *cunt*, *motherfucker*, *cocksucker*) in the Google Books database between 1900 and 2008. I chose to focus on these words in particular because they were designated as being some of the most offensive by either the BBC[38], or in the BSA report mentioned above. As can be seen, for most of these words, an increase in their usage is clear, even if moderate in certain cases (e.g. *motherfucker* and *cocksucker*). It should be noted that this list is not exhaustive however for readability concerns, and that for some words not included in the chart, no particular increase can be felt. Nevertheless, for the most offensive ones at least, a trend emerges, which seems to indicate that from the 1960s onwards swear words are used increasingly more often. I do not

[37] See Grimm (2003) for more details on this.

[38] BBC editorial guidelines. Last seen on December 14th, 2016. URL: http://www.bbc.co.uk/editorialguidelines/guidance/strong-language/guidance-full

wish to expand on the causes of this increase in the 1960s as this is not the focus of this thesis, but this example shows that the greater tolerance and use of swear words is a global one which is not limited to online or digital content, and Jay (1992: 155) noted this overall increase in the US as well.

Although sometimes mixed, other results pointed to the fact that swear words have gradually been considered as less taboo, while being used more often. It can sometimes be hard to have global evidence of the evolution of the frequency of use of certain words, as reference corpora like the BNC are rarely diachronic and as McEnery (2004: 40) mentioned:

> [t]his is a difficult question to address given the corpus resources available. If in the future an equivalent to the spoken BNC is produced, it may be possible to explore changing patterns of BLW [Bad Language Words] use over time. As it stands, however, this is not possible at the moment.

Although other reference corpora have been created since the BNC (COCA, COHA, GlowBe etc…), the sample, and their orientations being different, studying swear word usage over time with steady standards is not as easy as it may seem. Indeed, huge corpora like the GlowBe (Corpus of Web-Based Global English), or the COHA (COrpus of Historical American English) seem to indicate an increasing frequency of swear words over time, but as these are for a great part based on online content, itself growing exponentially as technology becomes more and more accessible, we can legitimately wonder whether this increase in swear words usage is due to an actual increase, or to a greater likelihood of swear words appearing because of a much greater quantity of data appearing on the Web. The frequencies observed in the two corpora mentioned previously being the raw frequencies, and not, for example, the normalized frequencies per million words, this question remains unanswered as it stands.

It is possible however to analyze that more objectively by studying the attitudes and patterns of people from various age groups. However, using this method can be risky, because as Baker (2010: 58) pointed out:

> McEnery et al. (2000a, 2000b) found that in the spoken section of the BNC, after age 16 there was an inverse relationship between swearing and age, with younger speakers generally swearing more than older speakers. Could this finding be used as evidence to argue that swearing overall is on the increase in British society? Unfortunately we would need further information. Swearing may instead be an aspect of 'age grading', where people only use a particular linguistic feature only at certain points in their lives. The BNC does not provide any evidence about whether the older speakers actually swore more when they were younger, or whether the younger speakers' swearing behaviour will decline as they age. Differences in age at a given point in time may be suggestive of diachronic changes, but are certainly not proof.

Thus, although this method is not ideal, reviewing several other studies in order to understand how swear word usage evolved over time seems to be one of the only methods available at this point.

The report from the BSA (2013: 4) states that "[y]ounger respondents tend to be more accepting than older respondents" in their evaluation of the acceptability of swear words in broadcasting. Williamsson (2009: 2) also noted that "[y]oung speakers and adolescents had a higher frequency use than other age-groups regarding these words". McEnery (2004: 40), talking about the data he observed in the BNC, mentioned that he:

> will assume, until evidence to the contrary presents itself, that what is observed here is what researchers have expected to see for some time—a correlation between age and BLW use, with BLW use declining as speakers become more conservative with age.

Stroh-Wollin (2010: abstract), who compared the results of two similar surveys on swear word use and perception carried out in the 1970s and in the 2000s, seemed to confirm the widespread view of younger generations being more liberal as well, as they said that:

> The tolerance towards traditional swear words as well as towards the practice of swearing in general has increased considerably since the 1970s. People in the older survey often argued that the use of swear words is a characteristic of poor language. This view was rare in the new investigation, where people were more apt to state that swear words are a natural part of the lexicon and even a usable resource.

These studies then seem to confirm that younger generations are both more tolerant regarding swear words, and also that they use it more often. This shift can also be felt in Rathje (2014: 59), who mentions that, in her study evaluating the perceptions of two generations of Danes regarding swear words, "most elderly people do not like swear words – they destroy the

language – while half of young people do not see a big problem with ugly words – they are part of the language".

In his study of swear word usage on MySpace, Thelwall (2008: 101) reported that "[y]ounger users had more swearing in their MySpaces than older users", which indicates that this trend seems to also be relevant on social media. Although not applying to both genders, Oliver and Rubin (1975: 191) reported that "younger women seemed generally to be much freer with their use of the "stronger" expletives while older women (over 55) seemed to fit the model Lakoff suggests, namely eschewing usage of these "stronger" expletives even in the more intimate situations".

The literature on the topic then seems to widely acknowledge the fact that older generations are less tolerant vis-à-vis swearing than younger ones, but as Baker (above) noted, this does not necessarily mean that people are overall more tolerant now than before. Indeed, this could simply be the sign that younger generations use more swear words than older ones, regardless of the period under study. Bailey and Timm (1976: 448) provide an interesting insight as they state that:

> [a]ge was of less importance for men in determining expletive usage, though the group aged 28 to 32 did show a notably higher average frequency per questionnaire (14.7) than the other 3 age groups (9.0 for the youngest men, and 9.7 and 9.3 for the 2 groups of older men).

Here, the fact that the youngest men were the ones swearing to a lesser degree seems to indicate that young people (and furthermore, men) do not necessarily swear more when they are influenced by their teenager speech patterns. This may then be the sign that change is indeed in the air, and that the contrast between younger generations and older ones reflects an actual linguistic change in progress making swear words more casual than ever. However, objectivity must prevail and I must acknowledge the fact that 1) Bailey and Timm's study was carried out on a "modest scale"[39] 2) we have provided much more evidence of the fact that younger generations' attitudes are more likely to be evolving than not, and 3) this study is rather old, so no matter the tendencies observed here, they may not necessarily accurately reflect more recent ones.

[39] They acknowledged it themselves (see 1976: 442).

As Margolis (see above) mentioned, swear words as a whole are much more common, and *fuck* has been particularly sensitive to this phenomenon as it is probably the swear word which has evolved the fastest. This word went from a word one would barely whisper a few decades ago, to one which can be heard several hundred times in recent movies. Both men and women are involved in this evolution, and both use this word to a much greater degree than before, so we may wonder why this word evolved so much.

1.2.2 The "F-Bomb": still as "explosive" as before?

As mentioned before, one sign reflecting the evolution of the status of *fuck* in public speech is people's increasing tolerance regarding this word on television shows. Indeed, it is nowadays fairly common to hear what was called "the F- Bomb" a few decades ago. The first person to utter this word on British television was Kenneth Tynan on 13 November 1965, which caused quite a scandal which pushed the BBC to formally apologize, and the House of Commons to sign four motions[40]. As can be seen in the table below[41], *fuck* is far from being as exceptional in cinemas and on TV as it was in 1965:

Movie	Year	Fuck count	Minutes	Fuck/minute
Swearnet: The Movie	2014	935	112	8.35
Fuck – a documentary on the word	2005	857	93	9.21
The Wolf of Wall Street	2013	569	180	3.16
Summer of Sam	1999	435	142	3.06
Nil by Mouth	1997	428	128	3.34
Casino	1995	422	178	2.40
Straight Outta Compton	2015	392	167	2.35
Alpha Dog	2007	367	118	3.11
End of Watch	2012	326	109	2.99
Twin Town	1997	318	99	3.21

Table 2.1: Movies in which *fuck* occurs exceptionally often

[40] See "Has swearing lost its power to shock?", *The Guardian*, February 5th, 2004. Last seen on January 4th, 2017. URL: https://www.theguardian.com/media/2004/feb/05/broadcasting.britishidentityandsociety

[41] Inspired from Wikipedia. Last seen on December 14th, 2016. URL: https://en.wikipedia.org/wiki/List_of_films_that_most_frequently_use_the_word_%22fuck%22#

Another sign showing the growing casualness of *fuck* is the fact that this word is more and more used in languages other than English. Indeed, evidence is now given that *fuck* is regularly used in casual conversations between native speakers of at least French (see Jaffe, forthcoming), Danish (Rathje, 2016), Swedish (Beers Fägersten, 2014), Finnish (Hjort, 2015). These studies showed that *fuck* was completely integrated to the linguistic arsenal of the people using it, and that they could say *fuck* in situations when another swear word from their native language could have been used. Also, the report of the survey from the BSA on what New Zealanders found acceptable and unacceptable to hear in the media revealed that *fuck* is the word which experienced the greatest decrease in unacceptability, going from 70% of the informants rating it as unacceptable in 1999, to 50% in 2013. We can thus wonder what causes this growing acceptance. Is it the fact that more and more profanity is broadcast, so people hear it more often and thus use it more? Or is it the fact that people use it more in their daily lives, which makes television and radio producers freer to use it? According to Lakoff (1973: 144), "[t]he speech heard in commercials or situation comedies mirrors the speech of the television-watching community: if it did not, it would not succeed". According to this, both theories are probably valid; an increase of profanity in people's speech probably influenced the media in the amount of swearing they used, which had people hear it even more in their daily lives, rendering it more casual as it was and so on...

To better understand this phenomenon with actual numbers, I will now take a closer look at a study from McEnery and Xiao (2004). This study investigated the use of *fuck* in the written section of the BNC, and they focused on two periods in order to understand the way *fuck* and its derivatives were used. The table below is inspired from their study[42]:

[42] "RF" referring to the raw frequency, "NF" referring to the normalized frequency per million words, "LL" being the log-likelihood score, and "sig. level" being the p-value obtained.

Form	Date	Words	RF	NF	LL	Sig. level
Fuck	1975-93	75 501 632	762	10.09	5.241	0.022
	1960-74	2 036 939	11	5.4		
Fucked	1975-93	75 501 632	128	1.7	6.815	0.009
	1960-74	2 036 939	0	0		
Fucks	1975-93	75 501 632	18	0.24	0.958	1.000
	1960-74	2 036 939	0	0		
Fucking	1975-93	75 501 632	937	12.41	0.020	0.888
	1960-74	2 036 939	26	12.76		
Fucker(s)	1975-93	75 501 632	47	0.62	1.642	0.200
	1960-74	2 036 939	3	1.47		
All forms	1975-93	75 501 632	1892	25.06	2.520	0.112
	1960-74	2 036 939	40	19.64		

Table 2.2: Occurrences of *fuck* in the written BNC between 1960 and 1993, from McEnery and Xiao (2004)

As can be seen, there is no statistically significant difference of use of the various forms of the word between the two periods[43]. Even if the observed frequency of the word in the BNC (RF) may sometimes seem to be contrasting between the two periods, the normalized frequency per million words of the variable (NF) shows no real disparity. The only noticeable difference is for the form *fuck*, where the NF shows a 5% gap between the two. At first then, it would seem that, apart from a slight increase in the most recent period, the word has not really been used more. However, the data which can be the most interesting for this investigation is the one which McEnery and Xiao pointed out as well, that is, the fact that between 1960 and 1974, absolutely no occurrence of the derivatives *fucks* and *fucked* has been recorded. Then, this would mean that *fuck* started to be massively used as a verb from 1975 only, thus limiting its use in everyday speech before this period, and maybe explaining why it started to expand so massively. Therefore, even if we cannot speak of a huge increase in the use of the word *fuck* itself, what triggered its casualness in our everyday lives might be the proliferation of its derivatives, rendering the word a lot more accessible, and enabling it to be used in various

[43] The figures would be considered as statistically significant if the p-values were below the 0.001 threshold.

situations. But what about gendered uses of *fuck*?

In their study, McEnery and Xiao also investigated this aspect by analyzing the spoken part of the BNC this time. They additionally analyzed it in the written part of the BNC, but since so far I have mainly focused on what can be considered as "natural speech" patterns, or spontaneous speech, the audio recordings composing the spoken part of the BNC are closer to this kind of speech than the written part, which includes material taken from newspapers, science, business etc, which are more informal kinds of discourse[44]. The table below allows to have a better idea of the distribution of *fuck* and its derivatives according to gender in conversations:

Form	**Gender**	**Words**	**RF**	**NF**	**LL**	**Sig. level**
Fuck	Male	4 918 075	337	68.52	50.025	< 0.001
	Female	3 255 533	106	32.56		
Fucked	Male	4 918 075	25	5.08	0.510	0.475
	Female	3 255 533	13	3.99		
Fucks	Male	4 918 075	5	1.02	0.386	0.534
	Female	3 255 533	2	0.61		
Fucking	Male	4 918 075	1394	283.44	353.624	< 0.001
	Female	3 255 533	321	98.6		
Fucker(s)	Male	4 918 075	18	3.66	8.967	0.003
	Female	3 255 533	2	0.61		
All forms	Male	4 918 075	1779	361.73	401.668	< 0.001
	Female	3 255 533	444	136.38		

Table 2.3: Gendered uses of *fuck* and its derivatives in the spoken part of the BNC, reproduced from McEnery and Xiao (2004)

[44] Note however, that despite this gap in register, the findings of McEnery and Xiao for gendered uses of *fuck* in the written and spoken BNC are relatively close.

As can be seen simply by looking at the contrast between the gendered uses of all forms of *fuck*, the difference corresponds to what could be expected when considering traditional and stereotypical assumptions on women and men: overall, men use the word and its derivatives almost three times as often as women. Even if the difference in use between women and men is statistically significant for two variants of *fuck* only (namely *fuck* and *fucking*), when considering all forms of the word, the difference between genders is statistically significant. Then, we may be tempted to assert that *fuck* is a gender-specific feature of spoken speech for British people (at least in the 1990s), since McEnery and Xiao (2004: 511) mentioned that, according to their results, *fuck* is "a marker of male readership/authorship as it is a marker of male speakers".

However, as we will see in greater detail later, aggregate data do not mean much, and often tend to blur the statistical reality in some regards. So, to go beyond the results presented by McEnery and Xiao I will analyze the dispersion of the word *fuck* and its derivatives throughout the corpus[45]. First of all, it should be noted that in the corpus, the two men (out of 2448) using *fuck* or variations of it most frequently in the corpus account for 32.5% of all occurrences of the word for men, while the two women (out of 1360) using it the most account for 26.4% of its occurrences for women. The word and its variations are thus clearly not dispersed evenly, and a small minority of speakers account for a large proportion of the occurrences of the word for both genders. Also, of the 1360 women recorded in the spoken part of the BNC, 5.2% (71 women) used the word *fuck* or one of its derivatives, and of the 2448 men recorded, 5.2% (129 men) used it. Therefore, in the spoken part of the BNC women and men use *fuck* at exactly the same rate. In this example, what varies, and could have given the wrong impression that *fuck* is a male word, is the individual use of it, and the fact that some men used the word at a much greater rate than individual women, thus biasing the overall data in favor of men. Once more, this is the proof that:

1) We must be careful when analyzing aggregate data, and that higher-level statistics may not always reflect the sample accurately.

[45] The data was retrieved using the BNCWeb tool. Last accessed on December 15th, 2016. URL: http://bncweb.lancs.ac.uk/cgi-binbncXML/dlogs2.pl?selected=Location%3A+%2Fcgi-binbncXML%2FBNCquery.pl%3FtheQuery%3Dsearch%26urlTest%3Dyes

2) Women and men actually use swear words in a way which is very similar, at least much more similar than stereotypes predict.

So, we have seen so far that swear words are continually rated as less offensive with time. However, despite this overall growing acceptance, as we have seen with the survey from the BSA, and the rating from the BBC, *fuck* and other swear words are still considered by a majority of people as being unacceptable (in the case of the BSA), or as being part of offensive, or very offensive language (in the case of the BBC). However, studies seem to confirm that people are using these "very offensive" words more than before in their everyday lives, and also in various kinds of media. We may wonder then, what triggers people and the media to use with a greater frequency words which are considered by a majority to be unacceptable? An explanation to this can be found in Beers Fägersten's (2007: 16) comments on offensiveness rating data, suggesting:

> [...] the unlikelihood that any participant, when presented with a list of isolated swear words void of context and asked to rate their offensiveness, would consider swearing from an alternative perspective. Consequently, offensiveness ratings are traditionally high, which, when juxtaposed with the similarly high frequency counts of swear words, contributes to a 'swearing paradox', representing the question of how this highly offensive behavior (according to ratings studies) can also enjoy such a high rate of occurrence (according to frequency studies).

In other words, offensiveness ratings would encourage the participants to virtually over-estimate their evaluation of the strength of swear words by presenting them as offensive to begin with. Thus, as mentioned before, participants would not consider swear words as being what they often can be, discourse markers of affection, bonding factors, expressions of pain, love, which may not necessarily be perceived as negative in these contexts. This observation is directly related to the example given earlier of "fast-food clerk" which, when void of any context, will probably not be considered as something offensive. So, what Beers Fägersten argues is that if the whole array of meanings and uses that swear words can have was taken into account, offensiveness ratings would probably result in a much lower unacceptability of these words, thus conforming to the fact that they are more and more used and going against the dichotomy of a high usage/high offensiveness ratio.

I have shown that swear words have been used increasingly in face to face conversations and in the media, but some studies indicate that one of the contexts triggering the most frequent use of swear words is probably that of social media.

1.2.3 "Overswearing" on the Internet

Online communities are often plagued with negative content – user-generated content that is negative in tone, hurtful in intent, mean, profane, and/or insulting.
(Sood, Antin and Churchill, 2012: 1481)

The Internet thus seems to be a particularly interesting place to study swear words, since as mentioned in the quotation above, swear words appear to be very frequent in some places. According to Herring and Stoerger (2014), the website 4Chan would be one of those places where "the discourse is notoriously profane and sexist". Some researchers try to associate the frequency of swear words reported in corpora of naturally occurring conversations to the overall proportion that swear words represent in everyday speech. For example, basing their claims on such studies[46], Wang et al. (2014) concluded that "[p]rior studies have found that 0.5% to 0.7% of all the words we speak in our daily lives are curse words". Such claims are somewhat ambitious, as the corpora used in the studies on which Wang et al. base their conclusions are relatively limited, both in terms of the representativeness of their sample, and in terms of the amount of data they have[47]. Thus, although very interesting, these studies cannot be claimed to be an accurate representation of everyone's speech patterns regarding the use of swearing.

However, when comparing Wang et al.'s results to studies of a similar scope and scale, the amount of swear words they found in their sample still seems unusually high. Indeed, according to Wang et al. (2014), about 8% of all tweets emitted would contain a swear word, which is substantially more than a study carried out on online chatrooms and which found that 3% of the utterances in their sample contained swear words (see Subrahmanyam et al., 2006).

Although comparisons between the amount of swear words present online, and in face to face interactions are hard to make because such measurements will vary a lot from person to person and from context to context, it still seems that swearing is common online, and it may even be the case that swear words are favored in certain online communities. Indeed, as Sood et al. (2012: 1489) found out in their study of online comments, "profane comments are more popular or more widely read than non-profane comments". Online hostility is often referred to as

[46] Notably, Jay (1992) and Mehl and Pennebaker (2003).

[47] Jay's corpus was based on elementary school and college students, and was composed of a total of 11 609 words in total. Mehl and Pennebaker recorded 4% of the conversations of 52 undergraduate students for four days.

flaming, which Kiesler, Siegel and McGuire (1984: 1129) describe as "remarks containing swearing, insults, name calling, and hostile comments", and this raises the question of the need for (self) censorship online. Very often, swear words are censored when mentioned on TV[48], or in print media like newspapers articles for example, as in this example from the Independant[49]:

> For anxious Republicans, it might seem bad enough that Donald Trump has been caught on tape saying that when you are a star you can grab women "by the p***y … You can do anything".

Thus, despite the growing acceptance of swear words that I have dealt with so far, these words are still frequently censored, which indicates that they are believed to be likely to offend the audience. Although this remains rare, swear words are also sometimes self-censored on social media or online forums, as Sood, Antin and Churchill (2012: 1489) indicate when they report that "only 0.67% (11,092) of all comments in the data set (1,655,131) contain an '@' symbol. Within this set, 39.9% of '@' usage was within the context of a censored or disguised profane term". Wang et al. (2014) state that "the following symbols, ' ', '%', '-', '.', '#', '\', '", are frequently used to mask curse words: f ck, f%ck, f.ck, f#ck, f'ck ! fuck", but because of the vagueness of this assertion it is impossible to know the proportion of censored tweets. As can be seen from the previous example, punctuation signs or special characters are often used to censor swear words, as Sood, Antin and Churchill (2012: 1487) indicate when they mention that "[t]he @ symbol is just one of many punctuation marks that could be used to disguise profanity", although not providing any figure regarding the overall proportion of censored words either. Self-censorship, or even obfuscation[50], using special characters can occur for several reasons, one of them being to avoid censorship by automatic online profanity detection systems, or a manual one by community managers. Another reason is to simply mitigate the potential offensiveness of the word for readers, as is the case in the previous examples of newspapers articles using asterisks to censor swear words.

Swear words can be censored on the Internet then, but it does not seem to be as frequent on social media, where people seem to express themselves more freely, using swear words with a relatively high frequency, and without censoring themselves. Solely based on these grounds

[48] See The Daily Show broadcast on October 11th, 2016. Last seen on December 19th, 2016. URL: https://youtu.be/LiPjWUn-PUo?t=9m12s

[49] See the Independant, "Donald Trump: all the sexist things he said". Last seen on December 19th, 2016. URL: http://www.independent.co.uk/news/world/americas/us-elections/donald-trump-sexist-quotes-comments-tweets-grab-them-by-the-pussy-when-star-you-can-do-anything-a7353006.html

[50] As labelled this way by Laboreiro and Oliveira (2014).

and on what I have shown so far, the language of the Internet (and especially that of Twitter, as we will see in section 1.2.4) can be considered as a distinct one, and as something different from face to face interactions. In the examples mentioned earlier in this section, I have shown that swearing does seem to be frequent on the Internet and on social media, but I have not yet focused on gendered differences. Some studies did highlight gendered and age differences in the way swear words are used on social media, and some of these revealed interesting patterns. Thelwall (2008) investigated the use of swear words by MySpace users and mainly focused on age, gender, and the region where the users were from. The first thing he noticed, and which corresponds to what I have shown so far is that "[i]n all cases the younger users' MySpaces contained more swearing than average" (2008: 97). Younger generations are once more shown to use profanity more often that older ones. Concerning gendered differences, Thelwall (ibid) adds that "female use of swear words was greater than male use for younger users in two cases: moderate language in the US and strong language in the UK". This, on the other hand, seems to contrast with the results presented up to now. Although evidence has been given multiple times that stereotypes presenting men as swearing more than women are wrong, data showing that women can actually swear more than men, especially when using words considered as strong, would be an interesting pattern to look at. Thelwall (ibid) even goes as far as saying that "the findings for younger U.K. users strengthen the evidence that in the U.K., strong language is no longer dominated by males". Thelwall partly used the BBC standards regarding his evaluation of swear word offensiveness in order to classify swear words in five categories, going from very mild, to very strong. What is considered as strong swearing in his corpus are variations of *fuck* only, and very strong corresponds to variations of *motherfucker* and *cunt*, and their frequency of appearance in Thelwall's corpus are listed below:

	Moderate (as %)	Strong (as %)	Very strong (as %)	Sample size
US Males	8	32	1	4659
US Females	8	25	1	3950
UK Males	25	30	5	486
UK Females	14	25	2	281

Table 2.4: Percentage of MySpace profiles containing various categories of swear words[51]

[51] This is a reproduction from Thelwall (2008: 95).

As can be seen from these figures from the whole sample on which Thelwall focused, regional differences are more relevant than gendered ones, which is one more sign that at least as far as quantitative figures are concerned, gender is not the most relevant parameter to consider. In terms of gendered differences themselves, apart from UK users in the case of moderate swearing, the gap between men and women from the same region is not statistically significant when it even exists. Also, here again the most noticeable pattern is that people are a vast majority not to swear[52], although strong swearing (i.e. *fuck*) is quite frequent for all kinds of users. Concerning the youngest MySpace users, their use of swear words is reported in the table below:

	Moderate (as %)	Strong (as %)	Very strong (as %)	Sample size
US Males 16-19	10	47	2	1530
US Females 16-19	11	38	2	1287
UK Males 16-19	33	33	8	171
UK Females 16-19	18	38	3	130

Table 2.5: Percentage of MySpace profiles from younger users containing various categories of swear words[53]

In this case, it can indeed be observed that strong swearing (i.e. *fuck*) is more present among female MySpaces than among males', and that moderate swearing prevails for younger females. However, as Thelwall noted, I must point to the fact that these differences are not statistically significant, but two things are salient in this study:

1) Younger generations of users are here again more likely to use swear words than when considering the whole sample.
2) Even with no statistical significance, the figures presented here indicate that on this platform, the quantitative *and* qualitative gaps between younger women and men are inexistent, and that women may swear more than men overall.

About this last point, Thelwall (2008: 102) argued that:

> it seems likely 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. If so, this is an extremely important development for gender roles in the UK, especially because swearing is closely related to psychological development and

[52] However, note that very mild, and mild swearing are not taken into account in this table.
[53] This is a reproduction from Thelwall (2008: 96).

> hence probably reflects much more fundamental shifts in the social psychology of the population.

About evolutions of gender roles and social psychology mentioned by Thelwall, Oliver and Rubin (1975: 196) provide an interesting insight on the use of strong swearing by women by observing that "women who were working at being liberated are more inclined to use these [strong] expletives more frequently than those who feel completely liberated". This explanation of an increase in the use of (strong) swear words by women is also used by Bailey and Timm (1976), and by Stapleton (2003: 3) who explained that swearing "functions not only as a marker of (group) identity, but also as a means of negotiating and actively constituting [an] identity". Risch (1987) also explained that swearing could be used to denigrate outgroups (in particular men) as a linguistic device strengthening the cohesion of the group. This echoes a finding from McEnery (2004: 33) who found that in his study "the word *cunt* is directed exclusively at males by females. It is a pure intergender BLW for females". It would then seem that swear words are a way for women to gain social power and to assert themselves. Thelwall's findings could then be linked to the development of a more egalitarian state of mind among younger generations of women, who may have internalized the influence that swearing can have on their status recognition. Interestingly this phenomenon seems more marked in the UK. Thus, the pattern of overswearing on the Internet described above may be more characteristic of women (at least in the UK) than of the Internet. Here, I am using the term "overswearing" as a means of referring to the idea that swear words, according to many studies mentioned so far, seem to be more frequent on the Internet (at least in some parts of it) than in other modes of communication.

All these elements led me to believe that an up-to-date analysis of gendered patterns when swearing, while also taking different age groups into account, and focusing on the UK could potentially lead to a better understanding of how the sociolinguistic situation of British people evolved. Additionally, this would allow me to shed another light of the phenomenon observed by Thelwall and all the researchers quoted above, and confirm or refute the hypothesis Thelwall made in 2008, that the pattern he observed would become more widespread on social media. As mentioned previously, I turned to Twitter to carry this analysis out, for reasons which will be detailed in the next section.

1.2.4 Why Twitter

I have already explained that Twitter represents a very interesting way to collect data for several reasons. Indeed, with more than half a billion tweets emitted every day around the world[54], Twitter, and social media more generally, represent one of the most plentiful and contemporary digital forms of communication. Bontcheva et al. (2013: 83) described Twitter as "the largest source of microblog text, responsible for gigabytes of human discourse every day". The importance and time we devote to social media sites are growing every year according to a study from Ofcom (see the 2016 Ofcom report), and this phenomenon concerns people from all age groups, and all socioeconomic backgrounds, as "a majority of internet users in all four socio-economic groups have a social media profile (see 2016 Ofcom report: 82; see also Smith & Brewer, 2012). People from all age groups also use social media sites, as until age 54, most people declared using these sites (2016 Ofcom report: 75). The youngest generations are the greatest users of social media, as 91% and 89% of the 16-24 and 25-34 respectively mentioned using them at least weekly. On top of being more and more popular, these sites receive more and more attention in our daily lives, as "[t]wo-thirds of adults with a profile use social media more than once a day" (2016 Ofcom report: 85). Social media are then an integral part of our lives, especially for younger people. They represent a hobby or a way to communicate with others, but they are also legally present in our marital lives, as inside prenuptial agreements for example, it is now possible in certain countries to include clauses stating that it is not permitted for spouses to post certain kinds of pictures of one another on social media. Interestingly, researchers studied the impact of social media in people's marital lives and found that "[t]he individual-level analysis, thus, is consistent with the results of the state-level analysis: use of SNS such as Facebook is associated with lower marriage satisfaction and a higher likelihood of divorce or, conversely, respondents in troubled relationships use SNS, including Facebook, more often" (Valenzuela et al., 2014: 99). So, even if the causality between social media and happiness is not proved, this shows that nowadays these sites have an impact on every aspect of our lives.

This is why using Twitter data for research is more and more common, and is used in many different fields; in medicine (Freifeld et al., 2014), to detect earthquakes (Sakaki et al., 2010), predict the stock market (Bollen et al., 2011), study college students' grades (Junco et al., 2011),

[54] See Krikorian R., "New Tweets per second record, and how!", Twitter Official Blog, August 16, 2013. Last seen on January 7, 2017. URL: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

to study political debates (Diakopoulos and Shamma, 2010), etc. Also, the constraints imposed by Twitter, especially in terms of the limit of 140 characters for each tweet, lead users to focus on the most important information and triggers a greater sense of immediacy than in other sites. This makes it a popular tool to stay informed on a particular topic, and Zapavigna (2011: 804) mentioned that "Twitter is the place you go when you want to find out what people are saying about a topic right now and in order to involve yourself in communities of shared value that interest you in this given moment".

Given the evidence I provided of the inherent part that social media have in our lives, it would be fair to wonder about the influence that the way we express ourselves on these sites has on our "offline speech". Another thing to consider would be whether online speech, and more specifically on Twitter, is representative of face to face speech. Although I do not believe that this answer is a binary one, some elements have already been provided which can help to understand why this is a difficult problem to tackle. First, we saw earlier that anonymity on social media has been shown not to have much of an impact on a lot of speech patterns which are acknowledged in face to face contexts. On the other hand, swear words seem more present on Twitter than in other online communities, and maybe even more present than in face to face interactions[55]. So, as far as swearing is considered, Twitter does not seem to be fully representative of offline patterns. Tagliamonte and Denis (2008: 4) present the language on the Internet as a whole as "complete with its own lexicon, graphology, grammar, and usage conditions". Also, the character limit present on Twitter which does not exist in face to face contexts makes it a very specific mode of communication, and thus distinguishes it from other forms of online, as well as offline speech.

However, online forms of expression are now being used in offline communication, and their meaning and pragmatic functions are recognized by most people. For example, the acronym *lol* (for "laughing out loud") has been used in court[56] and as the title of a French movie[57], which shows how popular it became. Also, let's now consider the expression "to troll", which originally means: "make a deliberately offensive or provocative online post with the aim of

[55] Although as mentioned previously, the quoted studies evaluating the amount of swear words people utter daily cannot be said to be representative of general trends.

[56] See "Judge's 'LOL' as he jails online boaster who ducked sentence". Last seen on January 9, 2017. URL: http://www.bbc.com/news/uk-scotland-glasgow-west-37195836

[57] See http://www.imdb.com/title/tt1194616/. Last seen on January 9, 2017.

upsetting someone or eliciting an angry response from them[58]". This verb, which was initially limited to online contexts, is now used to refer to offline attitudes[59]. Thus, although different from one another, online and offline speech seem to overlap and to influence one another to some extent. So, as Thelwall did (see earlier), I will assume that studying the speech patterns of British women and men on Twitter may give us an insight into the evolution of online, as well as offline linguistic patterns.

[58] Definition from the New Oxford American Dictionary.
[59] See http://www.bbc.com/sport/american-football/38405373. Last seen on January 9, 2017.

Conclusion

In this chapter, I have shown that swear words are both more common and more accepted than before, and this also includes the "new contexts" that social media (and thus Twitter) represent. This has been observed in various kinds of media, as well as in recordings, and is valid for both women and men. Social media seem to be particularly sensitive to this phenomenon, and women have been shown to use strong swear words more than men in certain contexts. The most important result of this review of previous studies and reports is that once more, figures proved that women do not swear less than men, and that sometimes the contrary may even be closer to the reality. This is the sign of a lessening of linguistic barriers, but even more crucially, this shows that a reversal of linguistic expectations may be on going. This is reinforced by the fact that in 2004, "[a]n investigation of the attitudes of Swedish men and women towards swearing women, reveals that most people in the ages 23-50 find it equally acceptable or non-acceptable for men and women to use bad language[60]". As mentioned previously, the use of swear words by women has been shown to be linked with a more equalitarian awareness, so another link may exist between an increase in the use of swear words among women and the development of a more neutral attitude regarding women who swear. This thesis aims at shedding new light on this phenomenon, and the next chapter will be the opportunity to be more specific about the approach I chose to conduct this study.

[60] See Svensson (2004: abstract).

Chapter 3: Corpus linguistics as an adaptive set of tools

> The year 1961, which more famously saw the first manned space flight, is the date to which corpus linguists can look back as the date when the enterprise now known as corpus linguistics (or more precisely computer corpus linguistics) came into being.
> (Garside, Leech and McEnery, 1997: 1)

Corpus linguistics (CL) represents a set of methods to approach and analyze corpora, and these methods are some of which I will use in this study. Although the term "corpus linguistics" is a fairly recent one, the use of corpora for linguistic studies can be traced back several decades ago, as the quotation above indicates, and the developments of this methodology, as well as technological progress shaped what CL is now. Understanding the evolution of this discipline, how previous studies have defined CL, and its strengths and weaknesses, will in turn help us understand how I have applied CL for this particular project.

This chapter thus aims at describing what CL is. In section 1.3.1, I will review how CL evolved, in order to clarify the way I approach the data, and also in order for this study to be firmly anchored within the set of existing branches of linguistics.

In 1.3.2, I will review some of the potential weaknesses of corpus approaches which have been identified by other researchers, so that I can limit these as much as possible in my own work. This section will also be an opportunity to talk about the importance of size when collecting corpus data, and the limits one should impose when building a corpus.

In 1.3.3, I will finally explain what I will consider as a swear word for this thesis. After describing how other researchers have considered swear words in the previous sections, I will have to build on these studies to determine my own definition of what swearing is in the context of British women and men on Twitter. This is based on a review of the existing literature, but also on methodological choices which have been made, so this chapter will also serve as a transition between Part 1 and Part 2 which will describe in greater details the methodology I used.

1.3.1 "Is this a sociolinguistic study? Or is it corpus linguistics? Computational linguistics?"

The title of this section refers to the kind of questions I have often been asked when talking to people willing to better understand how I approach the data. However, although this question may seem easy to answer, as it would appear natural to know precisely where one's research fits, I find I have a hard time clearly answering this. The research questions which motivated this project are indeed based on sociolinguistic concerns, but I approach the data with methods derived from what is labelled as corpus linguistics, and the way I sort the data requires some form of computer-programming as I will explain later, so this may then be labelled as computational linguistics. Moreover, the research questions raised in this thesis have often been addressed using one discipline (i.e. most of the time, sociolinguistics), whereas this work combines the methodological tools from different disciplines to try to have a more dynamic approach. Thus, the answer I give is usually that my approach is interdisciplinary. I do not like the idea of choosing labels, as it may sometimes be restrictive and obscure some aspects of the research, but understanding what these refer to is important as this can directly influence the direction one's research takes. I use computer programming to sort the data[61] only, so a computational approach, or one involving my own development of programs or algorithms, is not in direct relation with the analyses derived from it, so I will mainly focus on describing the corpus linguistic approach here. However, being able to determine where one field begins and where another ends is not an easy task. Several branches of linguistics often overlap and the barriers between them are sometimes blurred, which adds to the possible misunderstandings. Let's take the example of corpus linguistics:

> What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon a set of procedures, or methods, for studying language (although, as we will see, at least one major school of corpus linguists does not agree with the characterisation of corpus linguistics as a methodology). The procedures themselves are still developing, and remain an unclearly delineated set – though some of them, such as concordancing, are well established and are viewed as central to the approach. Given these procedures, we can take a corpus-based approach to many areas of linguistics.
> (McEnery and Hardie, 2012: 1)

So, corpus linguistics could apply to almost any kind of linguistic study based on a corpus (hence the name of the area…), which may make things even more blurry, especially as

[61] See Chapter 6 for more details regarding this.

McEnery and Hardie (ibid) go on to define corpus linguistics as "dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions". From there, we can define a corpus as being a set of machine-readable texts meant to be "representative of a given aspect or aspects of language" (Nelson, 2010: 56). Thus, according to these definitions, any linguistic study based on a digitized corpus could be said to be part of corpus linguistics, which in my opinion does not make things clearer. To add to this blurriness, let us consider the following quotation:

> The close connection between corpus linguistics and sociolinguistics can be underlined by the fact that, even in Labov's (1972) pioneering sociolinguistic study in New York City, he used a data collection method which has later been widely applied in corpus linguistics, that of random sampling. His informants were chosen by random sample on the basis of the gender, age, social classes and ethnicities represented in the city, much as a present-day corpus builder would do.
> (Andersen, 2010: 548)

Techniques used in one domain may of course freely be used in another, and according to the methodology he used, Labov could be considered a corpus linguist, which he generally is not. Labels do not mean much then, which is why I will not attach too much importance to detailing which branches of linguistics this work fits into. It can be seen, as I usually say people, as being at the crossroads of several disciplines, and understanding the orientations of these disciplines and how they evolved can be much more influential than definitions.

Corpus linguistics is nowadays fairly popular, and the emergence of corpora used for linguistic purposes can be traced back to long before the emergence of digitized corpora as we know them today. It is from the 1950s onwards however, that the development of technology increased the possibilities offered by CL, and "[i]t was the revolution in hardware and software in the 1980s and 1990s which really allowed corpus linguistics as we know it to emerge" (McCarthy and O'Keeffe, 2010: 5). Originally then, CL was not created to specifically answer sociolinguistic questions. We could thus wonder why I did not adopt a methodology specific to sociolinguistics, as I mentioned before that the main research questions of this thesis are directly inspired from sociolinguistic considerations. What are the advantages of CL in my case, and why turn to this to help me in my sociolinguistic investigation? As mentioned earlier, CL is a methodology rather than a branch of linguistics per say. Thus, this methodology can be applied to any discipline based on corpora. CL cannot be said to have an end in itself, as it *is* an end in itself. In other words, CL represents a toolkit which it is possible to apply to a variety of contexts. In order to explore this further, providing information regarding various aspects of

CL, and the different kinds of corpora there are, will prove useful for this work. Although doing a thorough review of every principle of CL is not my aim[62], a review of the key elements I had to take into account will allow for a better understanding of the way my corpus has been built, and as a landmark from where to base the interpretations I will later form.

Computers have been increasingly able to handle large amounts of data, and developments in computer programming allowed researchers to create better adapted tools to explore corpora and analyze them. From there, the emergence of larger and larger corpora led to much more diverse ranges of applications of CL. It is now considered to be a methodology rather than a discipline, and CL techniques can be used in phonology, sociolinguistics, but also geography or political science. Corpora can be as varied as one needs, and one of the advantages of CL is that its methods are not rigid, and can adapt to various conditions. However, although the tools used to analyze a corpus are important, the construction of the corpus plays at least as important of a role, and this is another aspect which is core to CL.

Two kinds of corpora can be distinguished then, the monitor corpus, and the sample corpus. The monitor corpus aims at continually expanding to add more content over time and thus always be up-to-date (see Sinclair, 1991: 24). The NOW (News On the Web) is an example of a monitor corpus[63], composed of about 4 billion words from Web-based newspapers and magazines. These kinds of corpora are particularly useful when carrying out diachronic studies investigating the way specific features evolved. Even if the goal of the monitor corpus is to keep renewing itself to stay contemporary, it is not necessarily preferable to a sample corpus, both just have different purposes. Monitor corpora generally seek to be representative of one variety of a language, or of a whole period. Sample corpora are much more focused and target a specific sub-category of the language or of the population. The sample corpus aims at being representative of a language or sub-group at a given point in time, and is usually composed by carefully selecting the sources/informants to maximize representativeness (see Biber, 1993). The BNC is an example of what a sample corpus is and provides a snapshot of what British English was in the 1990s. However, even if it is generally large and expanding, a monitor corpus (like most corpora) cannot be said to be representative of the language focused on, as "the corpus *itself* is necessarily a finite subset of a much larger (and in principle non-finite) entity,

[62] Other people have already focused on that in much greater details than I, see for example McEnery and Hardie (2012).
[63] See http://corpus.byu.edu/now/ for more details on the corpus. Last accessed on January 18th, 2017.

language" (McEnery and Hardie, 2012: 15). Thus, claims based on corpora which do not represent the entirety of the language or sub-group focused on must be made with care, no matter how thorough the composition of the corpus has been.

In my case, the corpus of tweets I collected is obviously a sample corpus aiming at giving an idea of what the speech of women and men is in the UK during the specified period, and on Twitter. In terms of their size, the first large-scale corpora compiled, like the Brown Corpus, were composed of about a million words, and with the advances in computer technology and storage, it is not uncommon now to see corpora reaching several billion words, such as the NOW corpus mentioned above. However, although size is important if one wants to have a suitable representativeness of the sample, "the value of a corpus as a research tool cannot be measured in terms of brute size. The diversity of the corpus, in terms of the variety of registers of text types it represents, can be an equally important (or even more important) criterion" (Garside et al., 1997: 2). Thus, corpus and sub-corpus design should prevail over the sheer amount of data, but I will come back to the importance of size later in this chapter. In other words, the corpus has to be compiled in order for it to be both representative of the population/context under study and suitable for providing answers to the research questions, if any. Indeed, a corpus is not necessarily composed in order to answer specific questions, as it is generally the case with monitor corpora. These are generally made to serve as a resource to study many aspects of the language/population, but are not created with a view to focusing on a specific aspect.

This brings us to another distinction commonly made within corpus studies, namely, the difference between corpus-based, and corpus-driven research (see Tognini-Bonelli, 2001). Corpus-based research relies on the building of a corpus in order to answer pre-existing research questions. This is the approach I had, and my corpus of tweets was composed to answer the questions and hypotheses mentioned before. Corpus-driven research refers to the fact that no preliminary hypothesis is formulated, and the corpus itself should be used as a source from which observations will be made. It can be argued, however, that these distinctions are somewhat superficial, as building a corpus almost necessarily begins with, if not research questions, at least ideas regarding what the corpus will represent. Although vast and varied, the BNC represents British English as it was spoken in the 1990s, which even void of any specific interrogation, still was a goal in itself before the very building of this corpus.

These distinctions once more point to the fact that proper sampling is key for a corpus to be reliable, and this principle will be at the basis of this study.

1.3.2 The limits of corpora: Quantitative Vs Qualitative?

CL mainly revolves around quantitative approaches. Thus, word frequency, keywords, collocations, and many tools for analysis are based on counts. Because modern corpora tend to be larger and larger, a manual checking and reading of the entire corpus is impossible, and this often leads to the opposition between quantitative and qualitative approaches, CL being sometimes thought to be quantitative only. This is not the case at all. Although frequencies and statistics are a big part of any CL-oriented study, mere figures do not mean much, and can even be misleading, which is why we cannot rely on quantity only. For example, as Brezina and Meyerhoff (2014) showed, when doing sociolinguistic analyses, researchers often tend to base their statistical tests and conclusions on aggregate data. In other words, the observations made are based on groups of texts/users/speakers, and do not take into account variation which may occur at the individual level. The problem is that such a methodology 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. In this case, the authors (i.e. Brezina and Meyerhoff) were referring to statistical tests which do not take dispersion[64] into account, and which may provide statistically significant results whereas the tendencies may merely be triggered by a few outliers, i.e. individuals using the variable under study far more than the average. This is true of two of the most popular significance tests, the log-likelihood and chi squared test. Brezina and Meyerhoff (2014) showed that, because dispersion is not taken into account with these tests, it is possible to have results of significance tests indicating that the definite article "the" is significantly more used by a social group than by another. In the case of a study on the speech patterns of women and men, such results would imply that the use of "the" would be socially motivated, if for example we compare the use of "the" in the male and in the female corpus. This seems very unlikely, and highlights the fact that we need to be careful when interpreting the data. Let us now take a practical example to make this clearer[65]. Imagine a researcher who wants to study the use of "cheese" by women and men to see whether using this word may have social motivations. The

[64] Dispersion being the degree of internal variation there may exist inside each document/author/speaker present inside the corpus.
[65] Note that this example is directly inspired from Brezina (forthcoming).

researcher has two different corpora of male and female speech. I explained earlier that having corpora which are well-balanced is key in having reliable results, and our researcher is fully aware of this. He took all possible measures to have an equal representation of the informants of his corpus and has a perfect balance in terms of the representation of women and men, socio-economic status, age and ethnicity, and his corpora of male and female speech are composed of exactly 10 000 words each. His results are summed up in the following table:

	Women	Men
Occurrences of "cheese"	200	270
Sub-corpus size	10 000	10 000

Table 3.1: Use of "cheese" by women and men in an artificial corpus

According to this data, it seems clear that "cheese" is more used by men, but our researcher wants to go beyond a mere visual evaluation of the figures, and wants to know whether the difference between the two groups is statistically significant. He then decides to apply a log-likelihood significance test, which is fairly common in sociolinguistics. The figures he obtains comfort him in his observations, as the log-likelihood obtained for such data is of 10.46, with corresponds to a p-value of $p < .001$[66]. Without going to great lengths on the calculations and specific thresholds regarding the log-likelihood, these scores are generally interpreted as being very significant, which may have our researcher assert that the difference between the two groups is socially motivated. However, what this person would miss if not looking at the data in detail, and especially if not looking at the dispersion inside his sub-corpora, is what can be observed in the table below:

[66] Figures obtained via the Ucrel online calculator. See http://ucrel.lancs.ac.uk/llwizard.html. Last accessed on February 1st, 2017.

	Women	Men
Speaker 1	20	5
Speaker 2	19	12
Speaker 3	21	11
Speaker 4	18	10
Speaker 5	22	11
Speaker 6	20	10
Speaker 7	20	180
Speaker 8	25	11
Speaker 9	15	8
Speaker 10	20	12
TOTAL	200	270

Table 3.2: Individual use of "cheese" by women and men in an artificial corpus

As can be seen in this table, individual women actually use "cheese" far more than individual men, with the exception of one male outlier, speaker n°7, who used it 180 times, thus biasing the entire male sub-corpus.

As Brezina and Meyerhoff (2014) argued, these examples illustrate how aggregating data can completely obscure key aspects of the corpora and often produces results which are misinterpreted. A solution to this problem is simply to use statistical tests taking dispersion into account. These examples aim at showing that we have to be careful when using and interpreting figures, and that indeed quantitative approaches are not necessarily more reliable and have to be used with care. This may be considered as a limit of CL, but in fact it would be more accurate to say that this is a limit in researchers' thoroughness and knowledge in the use of quantitative techniques, and not a flaw of CL itself.

However, it does not mean that quantitative figures cannot be analyzed qualitatively, and actually, not doing so would not be advisable, as manually looking at the data in context is what can allow for a better understanding of the figures, and can prevent errors of interpretation, as we have just seen. To illustrate this, I will take one example from Bamman et al:

> Men mention named entities about 30 percent more often than women do, and women use emoticons and abbreviations 40 percent more often than men do. The contrast of named entities versus emoticons may seem to offer evidence for proposed high-level distinctions such as information versus involvement. However, we urge caution. The 'involvement' dimension is characterized by the engagement between the writer/speaker and the audience, which is why involvement is often measured by first and second person pronoun frequency (e.g. Biber 1988). Named entities describe concrete referents, and thus may be thought of as informational, rather than involved; on this view, they are not used to reveal the self or to engage with others. But many of the named entities in our list refer to sports figures and teams, and are thus key components of identity and engagement for their fans (Meân 2001). While it is undeniable that many words have strong statistical associations with gender, the direct association of word types with high-level dimensions remains problematic.
> (Bamman, Eisenstein and Schnoebelen, 2014: 145)

Thus, without a manual checking of the data by looking at the named entities in context, they may have stuck to the traditional view of men as being "informative" and women as being "involved". This brings us back to the idea that words alone do not mean much, and that context plays a major role on the interpretation of these words, and as explained earlier, swear words are no exception. In order to avoid such biased interpretations, parts of the corpus should be consulted manually and visually as much as possible. Thanks to manual checking and the creation of sub-corpora then, it is possible with a corpus linguistic approach to have access to both quantitative and qualitative aspects of the analysis. The focus on the corpus as a whole gives overall figures regarding a linguistic feature which may, or may not, be generalizable to the language or variety in question. Focusing on a specific part of the corpus, which may represent a sub-category of the language or of the users/informants/speakers, gives more detailed information about this sub-group. This information may conform to the patterns of the whole variety, or it may differ from it, and using specific examples or a much more focused form of analysis like Critical Discourse Analysis (see Bloomaert and Bulcaen, 2000; Weiss and Wodak, 2003) may provide even more depth to the study. CL provides a framework to approach the data, but any branch of linguistics (but not only) can support the analysis and interpretation of the data, as long as the sub-part of the corpus is carefully selected to be representative of the population or feature under study. This brings us to an important factor to take into account for any corpus: how do we know if we have enough data? In other words: how much is enough?

Concerning corpus size, Sinclair (1991: 18) said: "The only guidance I would give is that a corpus should be as large as possible and keep on growing". With the advent of technology, and corpora being bigger and bigger, not having enough data may seem stressful, or on the

contrary, having enough data may seem difficult. Sinclair's assertion seems fair then, but does not help in determining the size from which the corpus can be said to be acceptable. Being more specific about what corpora should tend to, Reppen (2010: 32) said that "the question of size is resolved by two factors: representativeness (have I collected enough texts (words) to accurately represent the type of language under investigation?) and practicality (time constraints)". A corpus then must be representative of the population or language under study, and it should contain enough material to provide evidence of the linguistic principles at work. From there, the focus of the study will be what determines when to stop collecting data (if ever), and obviously, the larger the focus group or context, the more data one will need. Let us imagine that one wants to study polite forms of address in the staff meetings of the city council of a small city; several filters will probably prevent the collection of millions of words. First, city council meetings may occur once a week at most, and may not gather more than a dozen people, especially in a small city. The amount of speech which will be recorded will thus not represent a high volume. On top of that, if one wants to focus on polite forms of address in particular, the amount of evidence one will be able to focus on will be very limited. However, in this case the reduced size of the potential corpus would not be a problem, as it would still be fully representative of all the interactions which took place in the meetings of this city for a given period. If, on the other hand, one wants to study the speech patterns of British people in multiple contexts, the population will be much larger, and the sample corpus will have to be much larger, and much more varied to account for all the contexts at play. Because it would not be possible, as in the case of the council meetings, to have access to all the linguistic productions of British people, a sample corpus will have to be built in the most thorough way possible to provide meaningful evidence of actual patterns. This is why the BNC, for example, is divided into several sections of equal size. Even with the best sampling, however, a sample corpus will never be able to account for every aspect of a language, and this is where size comes into play. The bigger the corpus, the more likely it is to provide evidence of rare forms, like neologisms. Lexicographers then, because they study the whole language and want to record new forms, need huge corpora in order to spot a few occurrences of the rarer words they are interested in (Nelson, 2010). This is something to keep in mind when using a sample corpus; as it is only a snapshot of the language/population under study, absence of evidence is not an evidence of absence, and the feature may not be present simply because there was not enough material in the corpus to account for it.

Practicality is also of key importance when building a corpus, and although bigger corpora are almost always preferred, access to more data is not always possible, and solutions must be found to adapt to the situation.

In my case, the question of representativeness of my sample presents pros and cons. As will be discussed in 2.4.1, my corpus represents the entire set of geolocated tweets emitted within the UK for the entire duration of the collection. In this sense then, the corpus can be said to be fully representative of geolocalized British tweets for that period. On the other hand, if we consider that only about 1% of tweets are geolocalized, this may be viewed as a limit, but I will come back to the limits of my sample later. What is sure on the other hand, is that my corpus aims at representing the way British women and men from various age groups swear on Twitter. From there, the focus of the study has to be adapted to the sample, and in my case, I also had to adapt the swear words taken into account, as every geographical region may not necessarily consider swear words in the same way.

1.3.3 What I consider a swear word to be for this research

As demonstrated earlier, it is difficult to define a list of swear words which would be acknowledged as such by all the informants represented in a corpus. However, for the purpose of this study, it remains crucial to determine a set of words to investigate. In previous chapters I have reviewed how other people have defined swearing before, and for the purpose of this work, I also need to define a list of words I am going to consider as swear words. Without such a list, important contrasts highlighting specific contexts in which swear words are used or not will be impossible to make, thus preventing any comparison or analysis. This study is partly based on Thelwall's observations, so using his methodology and list of words considered swear words could have been envisaged. However, I did not wish to fully base this study on Thelwall's list of swear words for several reasons. The first one is that his study was carried out nine years prior to the current study, and although not outdated, the list he used may not be as representative of recent patterns as possible. The second reason is that, as mentioned before, he based his classification of swear words on degrees of offensiveness. Although I will refer to such groupings to illustrate and compare the differences and similarities of my results to other studies, I do not wish to base my methodology and analysis on such a categorization of swear words. I mentioned Beers Fägersten's "swearing paradox" earlier, and even if offensiveness ratings are important, I want to avoid this potential bias by not grounding my research on these

aspects. The last reason to avoid using Thelwall's classification is that he took into account several words which are probably more likely to be used in a context other than that of swearing (e.g. *bird*, *pig*, *jug*, *jew*), and which would require considerable disambiguating to be able to isolate only the cases in which the word is used in an offensive manner. As mentioned before, I realize that a lot of swear words can be used offensively or not, but because of the amount of data I have, and because of the impossibility of disambiguating the occurrences of *bird* used as a reference to a woman or when talking about the animal, I had to make a choice. As I explain later, this choice focused on what would more likely apply to the context I am basing this study on, namely British tweets. Thus, I choose to only select the words which are, in the UK, likely to be used most of the time as swear words, and be considered as such by a majority of people.

In order to compile a representative list of swear words fitting into the context of British tweets, I focused on words recognized as swear words by most speakers of British English. To this end, I first made use of Wang et al. (2014), who used a list of 788 English swear words from existing swear word lists and their variations. In their study, the swear words were manually and independently annotated by two native speakers of English, who both agreed that these words are "mostly used for cursing" (2014: 418). This final list of swear words on which both annotators agreed was what Wang et al. used to identify swearing in tweets. I decided to use the Wang et al. (2014) study as a standard on which I would base certain aspects of my 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 in English and 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 swear word repertoires of Twitter users in their sample. However, I chose to include the 20 most frequent swear words in the Wang et al. study, in order to increase the scope of my 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 19 words is due to the fact that I have excluded the non-English word *puta*. This wordlist should be reliably recognized as English swear words, but because swear word status tends to vary among people, and the list was determined by only two native English speakers, I believed there was reason to consider even more possible candidates. Furthermore, I wish 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, my goal is to analyze a much more localized corpus, and thus swear word usage may reflect a geographical bias. In order to account for this, I used all 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 Wang et al. The BBC can be considered 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[67]. This represents a reliable addition I can use to create a comprehensive list of words widely recognized as offensive and applicable to a British sample. In the end, my list of 26 swear words reflects a selected compilation of Wang et al.'s study (2014) 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*. These words are at the basis of my methodology for determining what a swearing tweet will be. On top of that, abbreviations (e.g. *fuk*, *wtf*), derivatives (e.g. *fucking*, *fucker*) and plural forms (e.g. *cunts*) of these words were also taken into account. In order to make things as clear and transparent as possible, the full list of all abbreviations which are additionally taken into account for each swear word is presented below (if applicable, i.e. if any variant/abbreviation of the word has been taken into account):

- *Fuck* (*fk, fking, fked, fks, fck, fuk, wtf, omfg, ffs, stfu, ftw, fml, lmfao, af*)
- *Shit* (*shite*)
- *Ass* (*arse, arsed*)
- *Nigger* (*nigga*)
- *Fag* (*faggot*)

In order to be able to account for the high creativity of Twitter users in their use of swear words, for some of them, the computer program that I wrote to detect vulgar tweets[68] matches the swear words wherever they are in the tweet, whether in the middle of a word or not. For example then, *fuck* will be matched whether written as a distinct word, or if present in the word *motherfucker*, and *shit* will be taken into account as such or as under the form *shitty*. In other words then, the computer program will not strictly match the words present in the list of swear words it has been provided. In some cases however, not strictly matching the swear words can give rise to other, unwanted words, to be taken into account. In the case of the swear word *ass* then, if the

[67] 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

[68] See Chapter 6 for more details regarding my use of programming to sort the data.

program is not told to strictly match it, it would also recognize words like *associate* or *glasses* as matching the query. In the case of *hell*, this implies that *hello* would be taken into account as well. For these potentially ambiguous words then, the program was specified to strictly match them, that is to say that these words had to appear as distinct words followed and preceded by spaces. The words which have been strictly matched then are the abbreviations of *fuck*, *ass* and its derivatives, *hell*, *dick* and its plural, *tits*, *fag* and its derivatives, and *cum*.
The program also accounts for expressive lengthenings (e.g. *fuuuuccccckkkk*) and capital letters in any position of the words (e.g. fUcK).

Again, this list is not meant to be exhaustive, nor to be considered the only one on which it was possible to base my study. Instead, this is one example of an empirically constructed list of words which can be considered swear words in the context of British tweets, although it may have been possible to add or delete some words. This implies that every time one of these words will appear in the corpus, it will be considered a swear word. It may be argued that depending on the context, some words will not be considered swear words, like *bitch*, which can also designate a female dog, or *cock*, which can refer to a rooster. This is true, and only a manual checking of each tweet would allow for this factor to be accounted for. However, because of the vast amount of data flowing on Twitter, this manual checking was not an option. It could also be argued that computational techniques may disambiguate these potentially problematic words, but I do not have the resources nor the knowledge to put such a system in place. So, I will assume, as Wang et al. (2014) and their two human annotators did (see above), that all the words considered as swear words in my list are mostly used for swearing. Although in some cases some words may be used to refer to something other than what the swear word implies (e.g. a female dog for *bitch*), I will assume that the vast amount of data will limit the interference of these cases as much as possible.

This selection of words which are to be considered swear words in the context of this study should ensure an accurate representation of the words which are the most likely to be considered as such by British Twitter users. From there, distinctions can be made between the contexts in which swearing and non-swearing tweets occur, which will allow for a more thorough analysis of it. However, before mentioning the results themselves, an in-depth description of the methodology used to carry out this study and compose the corpus has to be laid out, which will be the focus of Part 2.

Conclusion

This chapter aimed at providing some background information regarding the origins and the development of corpus linguistics, which is the main method I used to analyze the data presented in this thesis. Thanks to the description of the advantages and potential weaknesses of the approach, I will be able to build on that for my own analysis. This will provide me with a better awareness of the errors made in previous studies so that I do not reproduce them. It will also provide me with better foundations to reproduce what has successfully been attested before.

I have also given detailed information, based both on previous studies and on methodological choices, regarding which words I am going to consider as swear words for this study. This latter section, and this chapter as a whole, serve as a transition between this first part, which aims at providing a theoretical background from which to base my methodology, and the second part which will be more technical and detail at length the applied methodology. Only thanks to a review of previous related studies can a methodology be appropriately laid out, hence the importance of the first part of this thesis.

Part 2: Methodology

> The ability to replicate a result, whether experimental or observational, is, nonetheless, still clearly central to scientific practice.
> (McEnery and Hardie, 2012: 16)

As the quotation above illustrates, proper research should be as transparent as possible for many reasons. One of the most important one is replicability. Any project should provide all the information necessary to enable anyone to reproduce what has been done in order to verify what has been asserted and modify, improve or discard what has been said in the original study; to me, this is what any research endeavor should tend towards. Being empirically proved wrong by a later study is something that we should not be afraid of because this represents a step forward towards a better comprehension of the practices at work. This is the unending cycle of academic research and what enables it to evolve.

The accurate description of one's methodology is therefore crucial in that it is the foundation which will allow for the study to be carried out in a relevant manner, and have results which can be considered as reliable. This will then enable others to build on that to go beyond what has been done so far. The account of methods used is thus both an essential part of any study to show that the chosen methodology is the best suited to answer the research questions. It is also an instructional part of any study and should thus be made thoroughly. In this part, I will focus on the reasons which led me to choose Twitter as a source of data, and I will present the central tools used for data collection and exploration.

In Chapter 4, I will present the barriers I faced, and how I overcame them by collaborating to create an online interface for me and other researchers to use.

In Chapter 5, I will detail how I used this interface, as it is a crucial component of my data collection and the methodology used to collect the corpus.

In Chapter, 6 I will explain how I sorted the data to filter it further in order to isolate the information that was relevant to my research goals.

Chapter 4: From IRL to API

These acronyms, and thus this title, may look foreign to someone who is not familiar with them. IRL (In Real Life) is a slang form often used to make a distinction between what happens on the Internet (online forums, video games etc…) and what happens elsewhere, in "the real world". As mentioned before, API stands for Application Programming Interface and refers to processes making interactions with websites and databases possible, and is one kind of protocol I used to collect Twitter data. This title then refers to the relationship between data collection in the physical world, with questionnaires, surveys and recordings carried out face to face, and data collection of exclusively online content collected on the Internet only.

Section 2.4.1 thus aims at presenting how I came to use Twitter data for my research, after originally wanting to use data taken from recordings and ready-made corpora. This change in the collection method led me to meet a lot of people with whom I have collaborated in order to provide tools which would help other researchers.

Section 2.4.2 presents the main reasons why Twitter is an increasingly popular tool among researchers.

Finally, section 2.4.3 provides a technical description of Twitter's APIs will help in understanding the methodological choices I made, and also in understanding what is and is not possible to do with Twitter.

2.4.1 What to do when no ready-made corpus fits your needs?

In Chapter 3, I explained how I came to be interested in the way British people have been using swear words lately. When I knew what I wanted to investigate, I started exploring the possibilities I had in terms of corpora that I could use to answer my questions. I then began looking for a ready-made corpus which could fit all my requirements, which were:

- The corpus had to be annotated for age and gender
- It had to be compiled with informants from the UK
- It had to be recent enough (compiled after 2000) so that the observations made in previous research could be verified.
- Of course, it had to contain forms of speech in which swear words were present, so very formal recordings, and semi-scripted interviews were not ideal, as people are more likely to monitor their use of swear words. Thus, I wanted to favor naturally occurring conversations to maximize the likelihood of swear words appearing.

After my search, I was left with a certain number of corpora which seemed interesting in some aspects, which were:

- The British National Corpus (BNC)
- The Collins Corpus
- Bergen Corpus of London Teenage Language (COLT)
- Cambridge and Nottingham Corpus of Discourse in English (CANCODE)
- Longman British Spoken Corpus (part of the BNC)
- Limerick Corpus of Irish English
- Scottish Corpus of Texts and Speech
- Intonational Variation in English (IViE)
- Cambridge English Corpus
- Corpus of English Conversation
- International Corpus of English (ICE)
- British English Speech Data
- Scottish Corpus of Texts and Speech
- Glowbe Corpus

All the afore-mentioned corpora had features which could potentially be used for my study. However, even with such a large amount of options, none of them could fit all of my basic

requirements, at least at the time I looked into them (early 2014). These corpora all lacked some crucial information, which made them incompatible with the type of research I wanted to carry out. The problem was that they were either not recent enough (BNC, Bergen Corpus of London Teenage Language…), not available to the public despite my attempts to contact the creators of those corpora (Collins Corpus, Cambridge English Corpus), or not interesting for a study on swear words because of the methodology adopted for the collection of some corpora. Indeed, concerning the IViE for example, there is no instance of swearing in the written part of the corpus, and among the thirty-six hours of recordings, the only informal conversations which have been transcribed are not usable because they are too short. Moreover, there is no instance of swear words in these informal conversations either. The reason why I am mainly focusing on the informal conversations in spoken parts of any corpus is that naturally occurring instances of swear words are hard to record to begin with; the only moments when an informant would utter such a word in an interview, or a context in which the informants know they are being recorded, would most likely be in informal conversations. Another example of a corpus which could have been interesting, but which eventually did not prove usable is the CANCODE. Indeed, as Carter (2004) said:

> [t]here are about five million words in the CANCODE corpus, and it's a very rich resource for researchers of spoken English. However, the data does have some limitations. Most people knew they were being recorded, and are chatting in informal situations such as while relaxing at home, with others of fairly equal social status. This means the interactions are generally consensual and collaborative, so the corpus has minimal evidence of conflict or adversarial exchanges.

Although it has been shown earlier that swearing does not necessarily occur in conflicts or adversarial exchanges only, this corpus does not contain anything which could be used for a study on swearing. Overall, the main issue with my search for a ready-made corpus which would fit my needs is that as previously mentioned, swear words occurring in natural interactions are very hard to record, which considerably limits the amount of useful data I would be able to use for such a project.

Thus, I was left with no viable option to explore and which would be substantial enough for a PhD thesis. I was then faced with the option of either compiling my own corpus, or modifying my research questions. Therefore, I had to find ways of collecting data which would fit my needs, and be large enough to try to draw results which would be representative of more than my own handful of informants. One other option I had come across was the use of Twitter data. Twitter is a means of collecting data which is increasingly popular because of the huge

amounts and the variety of data that one can collect as we will see later. However, collecting tweets is not as straightforward as it may seem, and a relatively strong programming background is required to collect data from Twitter. This makes it a problem for people who do not have that knowledge, and most of the time researchers in social sciences, like me, do not have these skills. This was the first problem I encountered. After a few months of trying to contact academics, programming groups and associations and trying to learn programming on my own, I was still faced with a problem that I could not overcome. What I needed was either too complicated for the people I met to help me, or it would require too much of their time, which they could not give me. The option of learning programming did not prove to be much more fruitful as it is an area of expertise totally different from what I had been used to in my studies, and it would take much longer for me to acquire that knowledge on my own. At this point, I started to question my choice of data collection. However, I knew that when the technical aspects were overcome, the data would fit every one of my requirements, while also originating from social media. I would thus be able to compare my results to some of the surveys I based my research questions upon. Moreover, I had already gone through most of the ready-made corpora that were available to me, and this brought me back to the fact that even if Twitter was at this point not the ideal solution for me, I had no other real alternative if I wanted to carry on with this research project. I then started all over again, and I once more contacted many people in the computer science departments of the different universities in Lyon, and this time I got an answer from Adrien Guille, who was a PhD student at the time, and who agreed to help me. We discussed my project, and after originally providing a program enabling me to collect tweets according to keywords, we discussed it further, and Fabien Rico, associate professor of computer science in Lyon 1, also saw my email and proposed to help as well.

From this point on a stream of collaboration started, which enabled me to reach my research goals, collect corpora of tweets. This meant that despite these few months of emptiness in terms of the results I got from my various attempts at collecting tweets, it eventually turned out far better than I originally expected. In other words, one should not be afraid to invest some time to go in a direction that one knows could be fruitful, even if it means getting involved in an area of research that is not familiar to them to begin with. Indeed, programming was something that was completely foreign to me, and the implications this had, i.e. my later involvement in computational techniques to explore and analyze corpora, were something new as well. Even if acquiring new skills during the research phase itself may appear scary, or worse, as something that should be avoided, it should not be ruled out simply because it can be considered a risk. As

I explained earlier, one has to be aware of the situation, and of the possibilities that are available at that time. In my case, I knew that no corpus was available to me, so investing some time to try something which I knew could solve my problems was worth it. Learning new skills is part of a researcher's life, and allocating time and resources in this regard should not necessarily be viewed as something to avoid, especially during one's PhD, which is often a turning point during which one's methodology evolves and strengthens.
The reason why I presented the previous section as an autobiographical monologue is to introduce an idea which is extremely important to me (and many other people), and which is at the basis of this thesis: the concept of collaboration in research.

One of the most important group projects I have been able to experience is probably the one involving CATS[69]. This is the name of a project born out of a common enterprise between researchers (computer scientists and linguists) from the universities of Lyon 1 and 2. It started in 2014, some time after I had the idea to use Twitter for my research and after I got in touch with computer scientists who agreed to help me. I knew that in social sciences researchers could find such a program useful too, so we were willing to make it available to others as well, so that other people could save time and not spend six months looking for others willing to cooperate with them as I did. Such tools already existed, and programs, or online interfaces allowing people to collect tweets were available, but the problem I faced with these was that they were either not easy to use, not reliable enough, or simply not free. Our idea was then to create something which would bridge that gap by providing a tool which would be free and easy to use for researchers who have no programming knowledge. The goal of CATS is to be used for research purposes, as Twitter is being a more and more popular way to collect data for studies in various fields.

2.4.2 About Twitter

As mentioned earlier, Twitter is a formidable source of textual data because of the sheer volume of tweets that you can collect in a short period of time. Apart from the social, demographic and linguistic aspects mentioned earlier and which make Twitter a very appealing way to collect corpora, some technical aspects make this social medium one of the most used in research despite the fact that there are more active users on Facebook (1.71 billion users

[69] Collection and Analysis of Tweets made Simple, see Chapter 5 for more details about what CATS is.

during the second quarter of 2016[70]) than on Twitter (313 millions[71]) for example. Indeed, some methods of collection enable one to literally stream tweets, meaning that it is possible to collect them as they are emitted, thus making one's corpus a most contemporary one. It is therefore an excellent way to study language use synchronically, i.e. at a given point in time. If one collects a corpus of tweets from a specific region during a short period of time, this corpus is a snapshot of language use of Twitter users from this region. This snapshot can then be analyzed in depth, which is what I will do, but many other options are possible. The corpus can for example be compared to other regions, or can be compared to a collection of tweets from this very region at another moment in time, several months or years later. In this case, Twitter would not only be a means of studying language synchronically, but diachronically, i.e. at different points in time. Monitoring language (change) through a constant collection of tweets then becomes possible. Also, if we take the apparent-time hypothesis into account (see Labov (1972) or Magué (2006)), studying age groups on Twitter can be another way to carry out diachronic analyses on this medium, but I will come back to the question of age on Twitter later.

Another aspect of Twitter which makes it popular among researchers is the fact that it is possible to collect corpora focusing only on the aspects one is interested in thanks to the different kinds of metadata associated with each tweet. Metadata are additional information such as timestamp, geolocation, username of the author, the description the user provided and much more, collected along with the tweets themselves. Actually, among all the metadata collected, the content of the tweet itself only represents one metadata, and is thus only a small portion of all that is collected with each tweet. At the moment of this thesis, there are 43 different metadata available for each tweet[72]. This plethora of information makes it possible to have an extremely fine-grained view of one's corpus by focusing on the relevant criteria only. As a result, even if millions, or even billions of tweets can seem like too much to handle, focusing on the relevant features of one's corpus/informants/population can be a first step towards organizing the data. Of course, this all depends on the purpose of one's research; I am focusing on swearing, gender, age, language, and the location of Twitter users, so I am going to focus on the metadata linked with this information, but not everyone will want to make use of such a specific focalization on

[70] Statista website. Last seen on October 12th, 2016. URL: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[71] Statista website. Last seen on October 12th, 2016. URL: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[72] Twitter developer documentation. Last visited on 31st October, 2016. URL: https://dev.twitter.com/overview/api/users

a corpus (i.e. not everyone will use the same metadata). Some researchers may only want to count how many times the first person pronoun "I" is used in a corpus compared to how many times the second person pronoun "you" is used, and may thus only be interested in the language in which the tweets are written. In this example then, focusing on the metadata determining in which language the tweets are written (i.e. English) could be one way to sample the tweets while collecting. Then, focusing on the various personal pronouns one is interested, selecting those as search terms in the corpus could be one way to carry out such a study for example. `

This raises issues that one has to take into account every time a corpus has to be filtered further before actually analyzing it in depth; the greater the number of filters one applies, the bigger the corpus needs to be in order to still have a substantial number of words/tweets/texts after "cleaning" the corpus. I will come back in greater details on the question of "how much is enough?" later (see Chapter 6 for more details) but a rule of thumb and general advice would be not to discard anything during the collection of a corpus if one can afford it, and this is what I did by collecting all the geolocated tweets which were in English and around the UK, to only filter that further later. It is indeed always more comfortable to later realize that one has too much data rather than not enough, and most of the time collecting the missing data afterwards is not possible. In fact, although it is possible to collect vast amounts of tweets by streaming them as they are emitted, it is not possible to stream tweets which have already been tweeted. Having access to these tweets requires using a different method of collection, the REST API, but makes it much more difficult to have access to as much data (see the next section for a more thorough description of the different kinds of APIs); hence I focused on the Streaming API only (see below for an in-depth description of the different kinds of APIs).

2.4.3 Some aspects of Twitter's APIs

Several methods can be used to collect tweets, but all of them involve connecting to Twitter's API in various ways, each of them serving different purposes. An API is a set of resources enabling people (most of the time developers) to create applications and interact and collect data from various sources (websites, databases etc…). At the time of this research, there are two major sorts of APIs which can be used to collect tweets: the Streaming APIs, and the REST APIs.

The Streaming APIs:

"The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data[73]". In other words, this means that these different types of Streaming APIs all enable data collection directly from the global stream of tweets, with certain variations in the type of data according to which API is used. There are four kinds of Streaming APIs:

- The Public API
- The Users Streams
- The Site Streams
- The Firehose

I will mainly focus on the Public API as it is the method I used to collect my corpora (and it is the API used by CATS). Also, to put it simply, the Site Streams and the Firehose (which is the name given to the total stream of public tweets) are either not open to the public (Site Streams) as this thesis is being written, or require special permission and infrastructure to use (Firehose). The Users Streams on the other hand is a more specific method of collection which provides "a stream of data and events specific to the authenticated user"[74]. This is thus much more specific as it focuses only on data related to the Twitter account and parameters of the user (i.e. the researcher in our case) willing to collect Twitter data. In my case, my goal was to collect as much data from as many people as possible as long as they were from the United Kingdom, in order to have a wide panel of informants and data, and not to collect data related to my own Twitter account. Therefore, I used the Public API which offers samples of all the public data being emitted in real time and for as long as the connection is maintained. In this case, it is important to note that this Public API provides a *non-random sample* of the public data. More precisely, this sample is a 1% sample of the firehose. The way Twitter selects the tweets which are part of that 1% sample (as it is "non-random") has been discovered by Kergl et al. (2014), and it was later discovered that "tweets are chosen for the 1% sample based on their timestamp, with all tweets made in a specific 10ms interval of each second appearing in the sample" (Yates et al., 2016: 2998). In other words, when collecting tweets via the 1% sample of the Streaming API without any filter[75], only the tweets emitted during the first 10 milliseconds (i.e. 1% of a second) of each second are collected. This is extremely important to know this for many

[73] Twitter developer website. Last seen on October 13th, 2016. URL : https://dev.twitter.com/streaming/overview

[74] Twitter developer website. Last seen on October 13th, 2016. URL : https://dev.twitter.com/streaming/userstreams

[75] See later for the reason why taking filters into account is important when collecting via the Streaming API.

reasons. First, because it means that researchers who want to monitor certain (linguistic) factors by collecting a sample of all the tweets which are emitted for an extended period of time will not have these tweets sampled according to their content. This information is crucial, especially for people who are interested in swearing, as it could have been the case that vulgar tweets falling into that sample were censored[76]. It is also worth noticing that even that 1% sample can allow for an interesting and accurate representation of the global stream of tweets if people are willing to study specific and limited events. The assurance of collecting all the tweets emitted globally (even if it is for 10ms only) each second means that one is assured to have a reliable, and more importantly, a consistent snapshot of the discussions going on globally about an event currently happening on a large scale. Without this information, it could have been supposed that the collection stopped for a moment when reaching the 1% limit before resuming, which would have given a much more chronologically unstable flow of tweets, thus potentially biasing the representativeness of the sample for certain events and discussions. Of course, although this sampling is revealed to be well-balanced, it may not be adapted to everyone depending on how fine-grained one's corpus needs to be, and in some cases, one may need everything that has been tweeted during a very short period of time, in which case the Streaming API may not be enough.

However, even if this 1% sample can seem small, it should be remembered that every day on average, about half a billion tweets are sent all over the world (Haustein et al., 2016), which means that this 1% sample still represents about 5 million tweets every day. In fact, it would not be advisable for any researcher to have access to the firehose, as collecting, storing and analyzing this much data every day would require an infrastructure which is rarely available to individuals and most often used by companies or computer science labs. However, it has to be noted that when applying filters, i.e. when collecting tweets according to the geolocation for example, as long as the number of tweets matching the query does not exceed 1% of the firehose (again, roughly 5 million tweets a day), all the corresponding tweets are collected as "the Streaming API returns public tweets that match a query; it returns up to 1% of all public tweets, which can be more than 1% of tweets matching the query" (Yates et al., 2016: 3002). In other words, there is a chance to collect everything that is relevant to a query if that query corresponds to a volume that is below the 1% limit. To give a practical example, in my case the only original

[76] Swear words are censored in certain media, so they might have been on Twitter, which would have been problematic for the data collection of this thesis.

criteria were that the tweets had to be geotagged inside the UK and be written in English; thus, I would only start to "lose" tweets if the number of tweets corresponding to these requirements went above 5 million per day. Taking into account the fact that geotagged tweets represent about 1% of the total number of tweets (Valkanas and Gunopulos, 2012) this means that more than 500 million tweets would need to be sent every day from the UK for me not to collect everything. As this is obviously not the case, it can safely be concluded that my corpus corresponds to the complete set of geotagged tweets emitted from the UK during the collection phase. Moreover, the average number of tweets I collected per day is of 334 102, and is thus far below the 5 million limit, which only confirms that the corpus did not reach the 1% limit.

Consequently, the Streaming APIs are to this day the most efficient method for collecting vast amounts of Twitter data in real time. It is not possible however to collect historical tweets with this method. So, if one wants to study tweets related to an event which occurred a year ago, there is no immediate way to have access to large amounts of data matching this requirement. The only way to have access to such historical tweets is to purchase them through companies like Gnip[77], who store absolutely everything emitted on Twitter since day one and who offer to sell exactly the data one needs. It is thus a very convenient way of having access to a personalized corpus to fit one's needs, but this method can be quite costly. Another possibility, although somewhat limited, would be to use another kind of API, the REST APIs.

The REST APIs:

"The REST APIs provide programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more.[78]"

The REST APIs are thus a set of APIs enabling users to interact with data in a much more focused manner than the Streaming APIs, as these provide ways to automatically tweet, retweet, send direct messages to other users for example; in other words, these APIs can help in automating the management of one's account. I am not going to go at length on the REST APIs, as these are not the kinds of data collection methods I used, but I am still going to briefly present one method which could be useful in this regard: the Search API. This API is a much more accurate and specific data collection method, but it is interesting because although it does not focus on completeness, it is one of the only (free) ways of having access to some historical data

[77] See https://gnip.com/

[78] Twitter developer website. Last seen on October 13th, 2016. URL : https://dev.twitter.com/rest/public

by enabling tweets to be collected and which are either considered as recent, popular, or both, and which have been published in the past 7 days[79]. It is thus interesting in that it provides a way to have access to tweets which are not accessible through the Streaming API[80], however the amount of data one can collect via this method is limited to a maximum of 450 queries/requests per fifteen minutes, which represents fewer tweets than what is possible to do via the Streaming API. However, these different APIs have different purposes, and the Search API cannot be discarded solely because the amount of data is more restricted. It can have extremely interesting applications if used creatively, and can allow the collection of recent tweets mentioning a specific hashtag for example, or it can also allow the collection of many tweets from a specific user who may for example display linguistic patterns one is interested in. However, despite the potential advantages of this API, in my case it was not useful so I will let the reader refer to the relevant Twitter documentation should they need more information about this.

[79] Twitter developer website. Last seen on October 13th, 2016. URL : https://dev.twitter.com/rest/public/search

[80] That is, if one's collection was not running the week before already, or if one missed these tweets because they were emitted during a time frame which did not correspond to that of the 1% sample for example.

Conclusion

This chapter showed how and why I came to consider using Twitter data for my research, as well as some of the difficulties I faced when looking for suitable datasets. I have shown how important collaboration has been all along my PhD, and especially when looking for ways to have access to tweets, and I described some technical aspects of Twitter APIs and what made it so interesting for research. This chapter is a most important one as it is the one laying the foundations of what this thesis is about in terms of methodological choices, but also in terms of philosophy and view of research.

Chapter 5: How CATS helped me get a grasp on the blue bird

As explained before, CATS is the result of the cooperation between researchers from social and computer sciences, and partly bridges one of the gaps there is between these two disciplines. The end goal of CATS is to represent a common platform for researchers from any background, so that anyone can take advantage of the ever-growing resources that social media have to offer.

CATS is a web interface hosted at the university of Lyon 2, in France. It is free to use, open, and accessible by anyone. Its goal is to provide a way for researchers from various domains to easily have access to Twitter's API to collect and analyze data for their own studies.

This chapter gives more details regarding CATS. The goal here is not for this chapter to serve as a user manual on how to use CATS, but rather as a presentation of what I did with the tool, and how I decided to use it and organize the data, so that this can be replicated and checked for future studies, or simply to improve the methodology I used here. I asserted earlier how important replication is in research, so giving technical details on a central tool used for a study, and how it was used is crucial.

Section 2.5.1 thus presents the different functions and tools implemented in CATS, and how I used them, starting with the collection phase.

In section 2.5.2 then, I present the various ways in which the metadata associated to each tweet work, and how useful they can be to sort the data according to various criteria.

2.5.1 Creating a new corpus

CATS is designed to be as user-friendly as possible. It should be obvious now that the two main goals of CATS are to enable researchers to collect and analyze tweets (although there is actually more to it…). Collecting tweets is obviously the first thing I needed to do before being able to carry out a study of any kind, and as I explained there are different ways of collecting tweets through the Streaming API (see previous chapter), and thus, although CATS only uses the Streaming API, there are different collection filters and parameters to take into account.

Duration

This is the first parameter that I had to set up, and this one is mandatory, no matter which filter collection I would go on to choose. What is important to note is that for this corpus, what I wanted to favor was representativeness and not the volume of data, and being able to assert that my corpus was a substantial enough snapshot of tweets emitted from the UK during the defined period is what I prioritized. I wanted to have a collection of tweets which would be representative of an extended and uninterrupted period of time. This was crucial to me because I wanted to prevent temporal bias as much as possible. As we will see (see Part 3), because Twitter is used a lot to react to certain events, these biases could emerge if one has a corpus that is focused on a very short period of time, like a few hours or a few days. It is indeed reasonable to assert that if one's corpus focused around a specific city during a major football match occurring in this very city during the collection phase, the composition of the corpus is likely to be influenced by the main event going on in this city when the corpus was collected (i.e. the football match). Of course in the case of a corpus collected precisely during a football match and focusing on tweets emitted from the hosting city the bias is obvious, and would be non-existent if the collection lasts for days, weeks or months. However, I realized that other types of events may influence the overall composition of the corpus, and in particular, events which last for more than a few hours. Although various kinds of events are constantly occurring around the world, locally and globally, and are an integral part of our society and one of the reasons why people tweet, and thus cannot be considered to bias a corpus of tweets, I realized that there may be exceptions. For example, my collection of tweets was running when the British had to choose whether they were in favor of remaining in the European Union or not. This vote and the discussions around it, and especially *after* it, lasted for several days. Such a major and historical event has of course been a central discussion on Twitter, and such a sample may thus

not be representative of how British people would normally tweet in the UK at another moment of the year. This is why I chose to collect tweets for an uninterrupted period of almost two months, so that the incidence of context (however exceptional it could have been) may be limited as much as possible in order to have a corpus which would be as homogenous, and thus as representative as possible of how UK users tweet.

As I mentioned earlier, I knew that, because of the way Twitter's limitations work, I would be able to collect the vast majority of geolocalized tweets emitted around the UK, so from there the volume of tweets I would have would only depend on how long I would collect tweets for, hence my main concern being related to the duration. Thus, I chose to enter a duration of 999 days, so that the collection would last until I decided to manually stop it, giving me all the flexibility I needed. However, because of intermittent problems[81], the collection of tweets stopped several times during the collection process. I watched the collection process very closely, and Adrien Guille, the administrator of the server hosting CATS, notified me when there was an issue, so every time a collection stopped for some reason, I was able to start another one hours, up to a maximum of a couple of days later. In the end, and despite the seven different collections I had to carry out during these two months of collection because of various interruptions, I limited the number of tweets lost as much as possible. Eventually, the various corpora I collected covered the periods as shown below:

Corpus name	Start date	End date	N° of tweets
Corpus n°1	06/09	06/13	1 748 725
Corpus n°2	06/13	06/27	7 016 333
Corpus n°3	06/27	06/30	1 662 152
Corpus n°4	07/01	07/13	5 007 179
Corpus n°5	07/13	07/22	3 209 268
Corpus n°6	07/24	08/02	3 993 928
Corpus n°7	08/03	08/04	89 667

Table 5.1: Detailed information regarding the various collections of tweets carried out

In the end, the collection ran from June 9th to August 4th, 2016. After collecting these corpora, they were concatenated into one single file before being further filtered (see chapter 3). Thus,

[81] These were due to power cuts, the need to restart the server hosting CATS for maintenance issues, bugs etc…

the reader has to be aware that the figures presented in the table above are the total number of tweets, before any processing of the corpus has been applied to filter it according to the gender, the age, or any of the other processes which will be described later.

Collection filters

There are three collection methods which are available through the Public API, and which is the API used by CATS, namely: keywords, location or users. Only one of these collection filters can be chosen for a single collection of tweets, and thus it is not possible to combine them while collecting. In other words, it is not possible to collect tweets from a specific region AND containing specific words. Thus, I had to make a choice. By now it should be obvious that the collection filter I chose was the location filter, and I made this choice for reasons that I am going to explain below.

Keywords

As mentioned before, the keyword filter enables one to collect in real time tweets containing at least one, or more words present in the list of words entered in this field. This keyword collection filter only focuses on the content of the tweets themselves, and does not take into account other parameters like the location, or the user tweeting this. I could have made the choice to collect tweets according to keywords and ensure that I would only collect tweets which I could consider as vulgar because they contained a swear word. This method would also have the advantage of ensuring that I would collect a much bigger corpus, because this would focus on tweets emitted all around the world. However, as mentioned above, the size of my corpus was not what I wanted to emphasize, and representativeness was my key focus, and in my case, the keyword filter was not the most appropriate choice for this. First, this would mean that I would not be able to focus on tweets emitted from the UK, implying that substantial work would need to be done afterwards to clean the data and isolate tweets associated to a geolocation corresponding to the UK. Secondly, even if we excluded the cleaning of the data, this would imply that my corpus would only be composed of vulgar tweets, which would completely prevent any evaluation of the frequency of use of swear words on Twitter. If non-vulgar tweets are not taken into account, it is impossible to calculate if men are more vulgar than not compared to women and vice versa. It would also make any objective observation of the proportion of vulgar tweets among the overall volume of tweets related to a specific topic/timeframe impossible. In other words, for me using the keyword filter would completely hide the parameters which would enable me to evaluate the use of swear words by women and men.

Moreover, with the volume of data collected with the keyword filter being very likely to be much higher than with geolocation, the 1% rate limit would probably be reached, meaning that in addition to having a lot of noise (data which would need to be filtered afterwards), a lot of tweets corresponding to my needs would be lost because they would not have been collected to begin with. Also, and as I wrote earlier, Twitter's limitations enabled me to collect all the geolocated tweets inside the UK, even when not just focusing on tweets containing swear words, which was why I chose this one.

Location

This filter enables one to only collect geolocated tweets which fall within a selected area. This collection method is very interesting if one wants to focus on users from a specific region/country, as this was the case for me. In order for CATS to determine this region, I defined a square around the region I was interested in (i.e. the United Kingdom) on a map, and CATS automatically retrieved the corresponding geolocation it had to take into account to filter the tweets as can be seen on the figure below[82]:

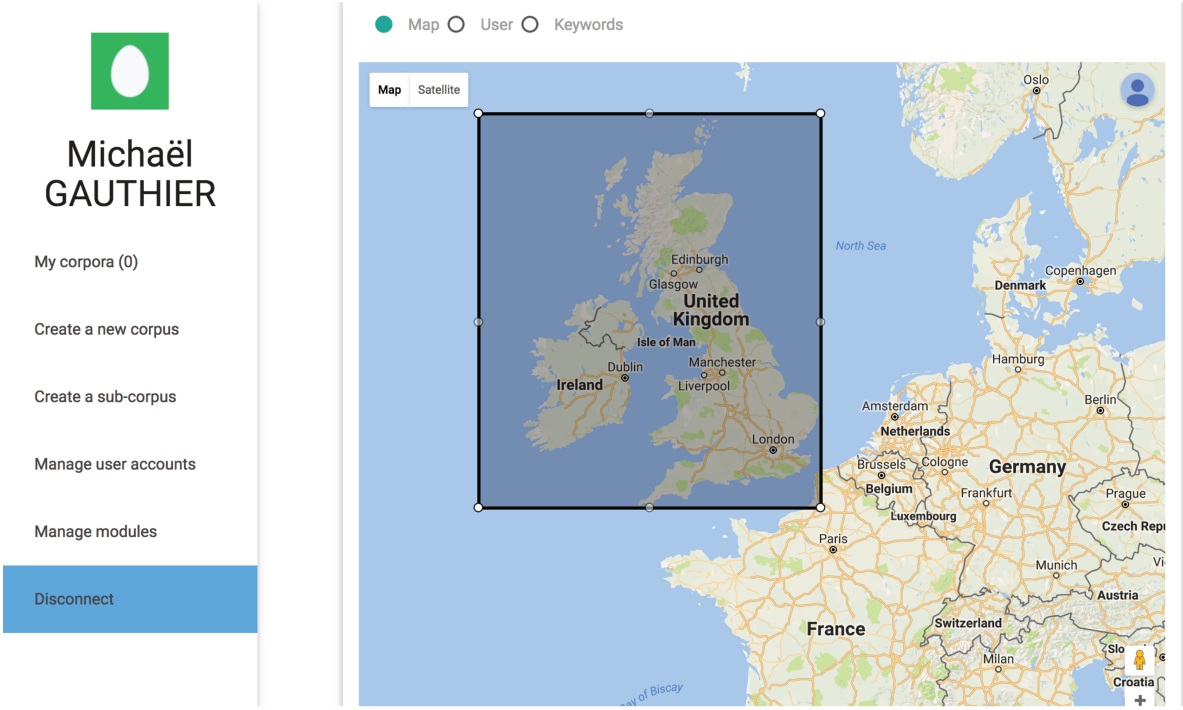


Figure 5.1: CATS' geolocation parameter

[82] Note that since the beginning of this thesis I have been referring to my corpus as being composed of tweets from "the UK", but this is actually not exact since as can be seen on Figure 5.1, the Republic of Ireland is also included in this collection. Out of convenience I thus refer to people from the selected region as "British people", also including those living in the Republic of Ireland, hoping not to offend anyone…

Only geolocated tweets falling within the requested bounding boxes will be collected by CATS. Actually Twitter's API uses two parameters to determine whether a tweet falls within the specified bounding boxes[83]:

- If the user activated the geolocation on their device, then the exact geolocation will be retrieved.
- If the user checked in a specific place (restaurant, street, airport etc…), the geolocation of the place will be retrieved, and if this matches the specified bounding boxes, the tweet will be collected.

In my case, it could be argued that one of the potential flaw of this collection method is that although I am sure that the tweets were emitted from the UK, I cannot be sure that the users are actually British, which is one of the requirements of my study. Indeed, if travelers from another country tweeted while they were visiting the UK and during the collection of my corpus, their tweets have been gathered. This will also take into account the tweets from foreigners living there. This is a true limitation, and I cannot be sure that 100% of the tweets present in my corpus are from native British people, and actually it is fairly obvious that some of the tweets collected are not from British people. However, it has to be noted that on top of the duration and the location filters I also added a language filter to only collect tweets which have been considered by Twitter as being in English (see next section for a more detailed description of the language filter). This ensures that at least non-British people not tweeting in English have been sampled out. On the other hand, this also ensured that *British* people tweeting in a language other than English have been sampled out, although this is not a problem as this thesis is about the use of swear words in English and not in other languages. So, even if some precautions have been taken to limit the inference of non-British people in the corpus, I had no way to completely prevent this phenomenon, and I have no efficient way of accounting for this. However, I trust that the language filter added to the sheer volume of tweets emitted from British people, which after all necessarily constitutes the vast majority of tweets sent from the UK, will make the bias of non-British people still tweeting in English reduced as much as possible.

Language

As mentioned above, for my collection I set up the optional language filter to only focus on tweets which have been labelled by Twitter's algorithms as being in English. Adding this filter

[83] See the Twitter developer documentation for more technical information on this. Last seen on October 28th, 2016. URL : https://dev.twitter.com/streaming/overview/request-parameters#locations

can be interesting when focusing for example on a single language when collecting tweets around a multilingual region (like Belgium for example). In my case this has been used as a way to only focus on English, as English swear words are the focus of this thesis. According to the Office for National Statistics, 92.3% of the population of England and Wales reported speaking English as their main language[84], 93% did in Scotland according to Scotland's Census[85], and 94% did in Ireland according to a 2006 census from the European Commission[86]. These statistics implied that focusing on English tweets only would ensure a homogenous representation of the population of the UK, while reducing the noise from non-British people tweeting in a language other than English inside the UK during the collection of my corpus. It has also been useful to limit the noise of potential French tweets, since as can be seen from Figure 5.1, the delimitation of the bounding box partly included the North of France.

As of today, the Twitter documentation regarding the techniques used to detect the languages in which tweets are written is not transparent[87], however members of the Twitter team reported that their "internal classifier at Twitter labels English Tweets with 99% precision, but on the recall-oriented dataset, its precision is 70%"[88]. This means that out of all the tweets which are labelled as being in English, the classifier is correct 99% of the time, but it fails to label 30% of the English tweets as actually being in English and labels them with another language. Trying to assess whether this recall is good or bad would be meaningless because in my case only precision is important as the labelling occurs during the very collection. So, I am guaranteed that at least 99% of the tweets present in my corpus are indeed in English, and I have no way of knowing how many tweets which have not been properly tagged as being in English have been emitted during the collection of my corpus. To try to limit the number of tweets lost due to improper language tagging, a possibility could have been to collect all the tweets which have been emitted around the UK, regardless of the languages used, and detect the language of all the tweets myself by using various existing techniques (e.g. machine learning). However, these skills are to this day far beyond what I am capable of, and I do not believe that I could have

[84] See the report from the ONS. Last seen on October 28th, 2016. URL : http://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/articles/languageinenglandandwales/2013-03-04

[85] See the report from Scotland's Census. Last seen on October 28th, 2016. URL : http://www.scotlandscensus.gov.uk/ethnicity-identity-language-and-religion

[86] See the 243th report from the Special Eurobarometer. Last accessed on October 28th, 2016. URL : http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf

[87] See the Twitter developer section on language. Last seen on October 28th, 2016. URL : https://dev.twitter.com/streaming/overview/request-parameters#language

[88] "Evaluating language identification performance". Last seen on October 28th, 2016. URL : https://blog.twitter.com/2015/evaluating-language-identification-performance

acquired an understanding of language detection techniques surpassing that of the Twitter research engineers in the time span covered by my PhD, thus I decided to limit myself to the tweets automatically labelled as being in English.

Users

This filter is the last method of collection which it is possible to use with the Public API. This filter allows the collection of all tweets from specific users. It may be interesting in the case of people studying public figures like politicians, actors etc… Note that retweets from other users will also be collected. Indeed, a tweet from someone retweeting one of the users mentioned will be collected as well. This detail is important, because this can potentially make one's corpus grow very fast if one focuses on public figures who are likely to be retweeted a lot. I have not used this filter at all for my research, but it may be interesting for future work to only focus on users displaying interesting swearing patterns, or to focus on users for whom specific information has been retrieved thanks to the description for example, in order to focus on the speech patterns of people from a specific city, or having specific hobbies etc…

Creating sub-corpora

Once my corpora had been collected, this feature enabled me to filter them according to various parameters, allowing me to focus on certain aspects only depending on which variables I took into account, as can be seen on the figure below:

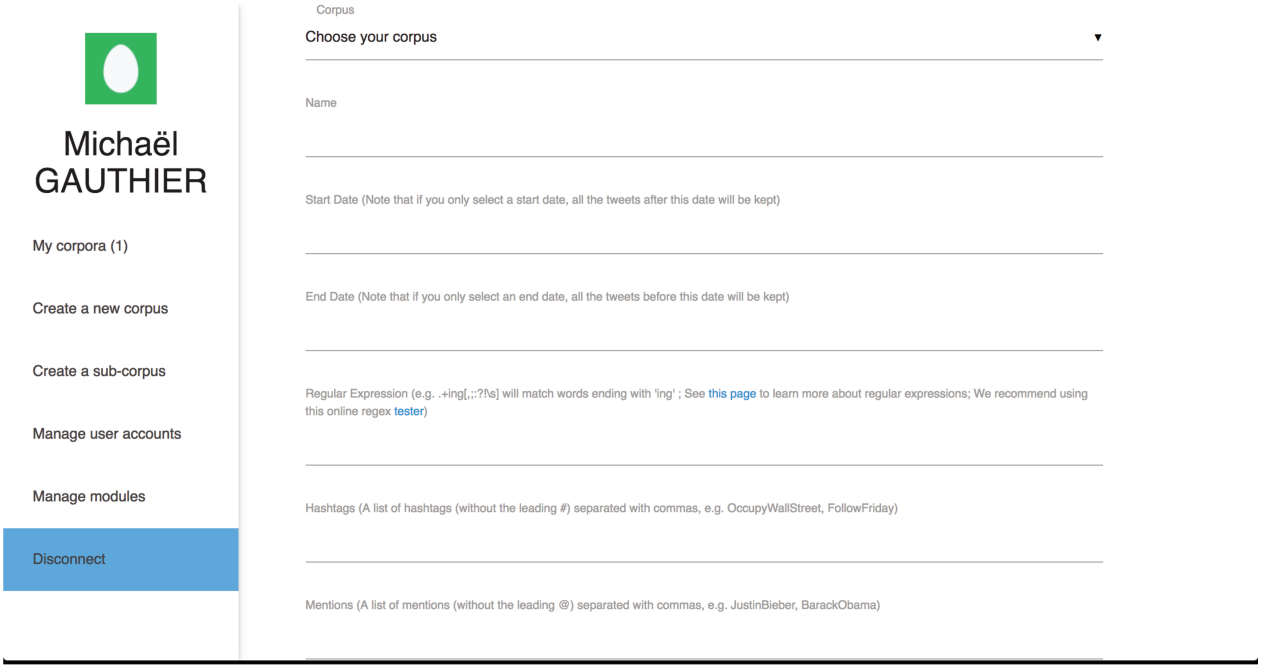


Figure 5.2: Sub-corpus creation interface

I was able to create sub-corpora of tweets containing each of the swear words I took into account for my study, as well as sort these sub-corpora according to each gender and age group[89]. Although regular expressions can allow for much more detailed queries (e.g. words having a particular ending and being at the beginning of a tweet for example), in my case I mainly focused on words themselves, while also taking the form of the words into account in certain cases. Thus, after creating a first sub-corpus of tweets mentioning *fuck* and its derivatives for men aged 18-25, I have been able to create a second sub-corpus of tweets from the same age group and gender but mentioning only *fucking*, and compare that to how the same form is used by women aged 18-25 for example.

Filtering by hashtags or mentions allows the researcher to focus only on tweets mentioning a specific person or hashtag. This was useful to me to see which people seem to trigger more profanity than others for example, or if there were recurring hashtags which were used as a means to swear.

Corpus mining

Once my corpora had been collected and sorted, the next logical step was for me to analyze them. The next section will present the various aspects which have to be taken into account

[89] Although CATS features several useful ways to create sub-corpora, I used other programming methods and software to create sub-corpora according to the gender and the age groups of the users, which I will detail in the next chapter.

when analyzing tweets (metadata, corpus size etc…), as well as the various analytical tools implemented in CATS, and how I used them.

2.5.2 Metadata available

As mentioned before, there are a certain number of metadata associated with each tweet when collecting them. These metadata give additional information and are one of the reasons why Twitter data is so useful to researchers and companies. Thanks to this data, it is possible to see how many followers users have, their location, if they changed their profile picture etc… These are these metadata I used to infer the gender and the age of my informants, but creativity in the triangulation of this data can lead to much broader application, and this is how companies use them in order to target a specific audience. Below is an example of the quantity of information collected with a single tweet and its associated metadata[90]:

"created_at": "Mon Oct 31 10:43:09 +0000 2016", "id": 793040607673475100, "id_str": "793040607673475072", "text": "This dirty timb wearing nigga gonna have the whole black nation delegation talking about meek all day... Fuck @TAXSTONE ", "truncated": false, "entities": {"hashtags": [], "symbols": [], "user_mentions": [{ "screen_name": "TAXSTONE", "name": "DADITO CALDERONE", "id": 70717194, "id_str": "70717194", "indices": [110, 119]}], "urls": [] }, "metadata": { "iso_language_code": "en", "result_type": "recent" }, "source": "Twitter for iPhone", "in_reply_to_status_id": null, "in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null, "in_reply_to_screen_name": null, "user": { "id": 18198440, "id_str": "18198440", "name": "YB", "screen_name": "YBizzle_", "location": "On your computer Screen", "description": "Liberation, Music, and Lubricities (In no specific order) A True Blue African American 🇺🇸🇺🇸 🇲🇧🇧🇷🇨🇦🇷🇺 e mail : yb.1music@gmail.com", "url": null, "entities": { "description": { "urls": [] } }, "protected": false, "followers_count": 472, "friends_count": 477, "listed_count": 17, "created_at": "Wed Dec 17 20:18:36 +0000 2008", "favourites_count": 856, "utc_offset": -14400, "time_zone": "Eastern Time (US & Canada)", "geo_enabled":

[90] This example tweet has been collected using the Apigee console. Last seen on October 31st, 2016. URL : https://apigee.com/console/twitter

false, "verified": false, "statuses_count": 14564, "lang": "en", "contributors_enabled": false, "is_translator": false, "is_translation_enabled": false, "profile_background_color": "008509", "profile_background_image_url": "http://pbs.twimg.com/profile_background_images/842981186/21cbe7d4ad7249f6cb88287e6964e0c1.jpeg", "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/842981186/21cbe7d4ad7249f6cb88287e6964e0c1.jpeg", "profile_background_tile": true, "profile_image_url": "http://pbs.twimg.com/profile_images/793003656224399360/VNhbSRfQ_normal.jpg", "profile_image_url_https": "https://pbs.twimg.com/profile_images/793003656224399360/VNhbSRfQ_normal.jpg", "profile_banner_url": "https://pbs.twimg.com/profile_banners/18198440/1477901781", "profile_link_color": "0084B4", "profile_sidebar_border_color": "FFFFFF", "profile_sidebar_fill_color": "000000", "profile_text_color": "0717BB", "profile_use_background_image": true, "has_extended_profile": true, "default_profile": false, "default_profile_image": false, "following": false, "follow_request_sent": false, "notifications": false, "translator_type": "none" }, "geo": null, "coordinates": null, "place": null, "contributors": null, "is_quote_status": false, "retweet_count": 0, "favorite_count": 0, "favorited": false, "retweeted": false, "lang": "en"}"

Thus, it seems obvious that when millions of tweets are collected, this amount of data leads to very heavy corpora in a short amount of time. In order to spare CATS' database and make it more efficient, not all the metadata are collected along with each tweet. This has several advantages, the most obvious being much lighter, but also data which are much more readable that the sample presented above. When viewing one's corpus in CATS, the metadata are sorted in columns, as can be seen in the figure below:

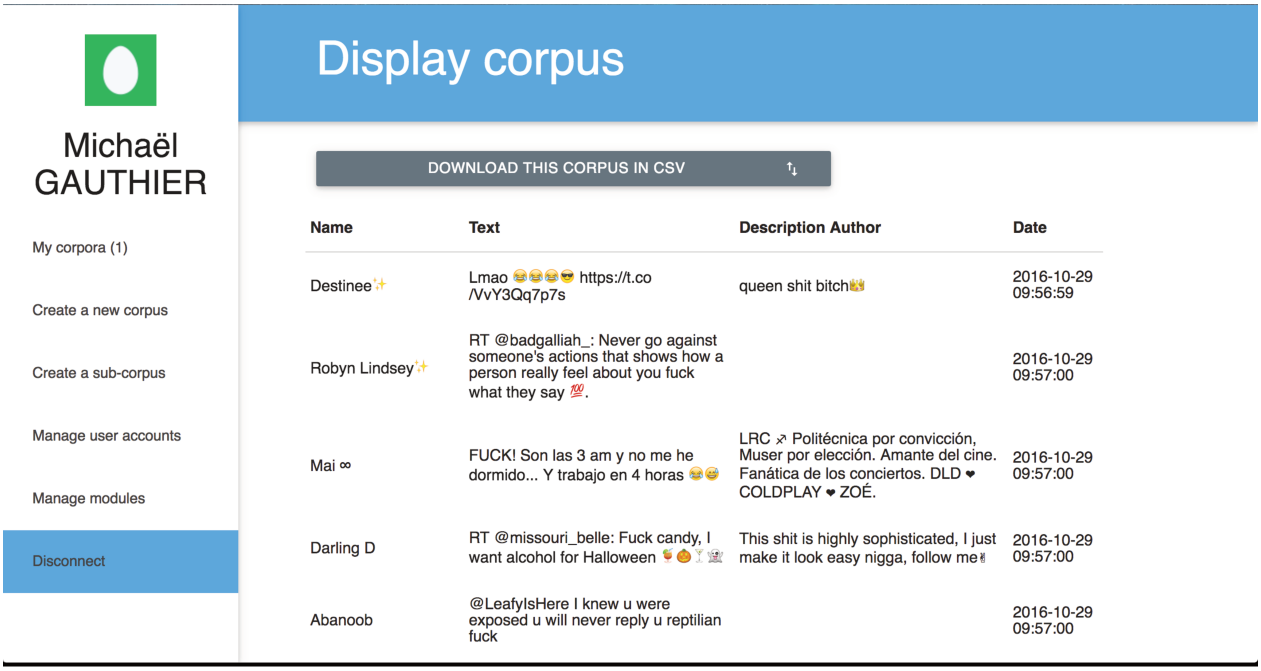


Figure 5.3: Corpus visualization interface

This made things considerably easier for me to read the data and focus on what is relevant to me. Thus, when viewing corpora using CATS, only the name of the user, the content of the tweet, the description the user provided (if any) and the timestamp. In my case, this is all the information I needed in order to determine whether the tweet was vulgar, emitted by a man or a woman of a specific age group. However, some people may benefit from other kinds of information, and more metadata are available when downloading the corpus as a .csv file, and one can also have access to the ID of the tweet, the ID of the author and the associated geolocation. This additional information is not included in CATS' interface out of a concern for readability. So, the metadata associated to each tweet and to which I have access to are: the tweet ID, the user ID, the timestamp, the tweet itself, the corresponding geolocation, the bio mentioned in the user's profile, the users' name. This information is what I use to sort the corpus and determine the gender or the age of the informants.

Conclusion

In this chapter, I reviewed some features of CATS, one of the key tools I used for the data collection phase of this thesis. I have explained what filters I took into account to collect tweets, and I have given a description of the various analytical tools which are implemented in the interface so that the reader knows what it is possible to do with it, and what the limitations are. This will prove useful in the next chapter, as I will further describe the content of my corpus, and especially how I sorted it. These aspects will also reveal helpful in the next part, as the description of the analyses and the results obtained will greatly be based on how I collected tweets, as well as on the technical limitations of CATS. Indeed, some of these limitations urged me to resort to other tools to sort or filter the corpus, and are thus key in better understanding the methodology, as well as the results which will be presented in Part 3.

Chapter 6: How old is he or she?

The problem for the linguist has shifted from accessing large enough quantities of data to elaborating a reliable methodology to describe and take into account this type of unprecedented evidence.
(Tognini Bonelli, in O'Keeffe and McCarthy, 2010: 18)

Indeed, many linguistic studies nowadays do not suffer from a lack of data. What is often more challenging than collecting data is its formatting, and this study is no exception. As mentioned earlier, the number of tweets I obtained after roughly two months of collection is 18 709 729, which represents 488 806 068 tokens[91]. Although the mere collection of tweets from a specific region and for an extended period of time represents a feat in itself for someone with originally so little understanding of programming like me, this was just the first of many other steps which I had to tackle before I could actually start analyzing this corpus. Indeed, many measures needed to be taken in order for this vast amount of data to be usable for an analysis of gender and age. As mentioned earlier, these two elements are not part of the personal information one can mention in their Twitter profile. Thus, they had to be inferred through various means I will present in this chapter. Additionally, I will give more details on the corpus after inferring gender and age, and discuss the differences which can be observed before, and after this operation, and the potential problems these may reveal.

In 2.6.1 then, I will detail how I managed to infer the gender of Twitter users, and the various parameters which were taken into account to carry this out, from the building of a repository of gendered names, to the problem of ambiguity of certain names which can be both female and male.

In 2.6.2, I will describe how I inferred the age of the users, the reasons why I chose to base this study on age groups, and how I chose these particular groups. I will also describe the distribution of tweets and users according to gender and age.

[91] Contrary to what is common practice in studies presenting corpora, I am not going to refer to the number of words present in the corpus as words per say, and I will be very specific about the use of tokens here, as tweets also frequently contain elements which may not be considered words, like URLs or emoticons.

2.6.1 Inferring gender

Unfortunately for researchers, Twitter profiles do not (yet?) have a gender category. If it did, the information provided by the user would be accessible as another form of metadata, and would make it much easier to know whether the user considers themselves to be a woman, a man, neither, or even another category. One thus has to be creative in order to have access to this information.

To infer the gender of the Twitter users, I chose to use the first name they provided in the field name of their profile, if any. Indeed, when setting up a Twitter account, there is no guideline stating that users must provide their real name, so not everyone will do it. Despite this potential limitation, focusing on the first name provided to infer age is commonly used in studies of online discourse (Mislove et al., 2011; Bamman et al., 2014; Sloan et al., 2013), mainly because of its relative simplicity. Indeed, other studies (Thomson and Murachaver, 2001; Cheng et al., 2011) showed that it is possible to determine the gender of people in online environments based on the content of the text itself. However, while this technique seems interesting because it could apply to every Twitter user and not just to those providing a valid first name, it requires a much greater knowledge of computational techniques than I have. Another drawback of this method is that it generally requires more context than what is usually available in 140 characters on Twitter, and may thus prove to be less accurate than in other online environments. Finally, this technique also implies the creation of a "gold standard", which is the corpus from which the patterns and tendencies will be retrieved. In the case of gender identification on Twitter, this would imply using a vast amount of tweets from users whose gender is already known, so that the latent trends can be extracted thanks to machine learning techniques, for example. Once a set of rules or patterns is observed, the program can be applied to a corpus of tweets from users whose gender is unknown, so that the patterns of the gold standard can be traced in this new corpus. The last problem with this method is that the building of this gold standard would require access to many Twitter users whose gender is verified, and putting this in place while taking into account enough parameters (that of the United Kingdom for example) so that the context can be reliable would be even more time-consuming, without being ensured that the lack of context in tweets would not be a problem.

Using first names, on the other hand, is relatively simple, although as mentioned before, this technique necessarily implies that a certain number of tweets will be lost. Indeed, as providing

a valid name is not mandatory on Twitter, a certain number of users do not provide a gendered name, or a realistic one, if any at all. Thus, it is not possible to infer the gender of a person whose "name" is "Fight till the end!", "**", or "SMOKE BREAK". Furthermore, it could be argued that we cannot be guaranteed that the first name provided is the user's actual first name. Indeed we cannot, but what is of interest to me is not whether users provide their real first name, but rather whether they provide a first name matching their actual gender. Previous studies indicated that they do most of the time (Huffaker and Calvert, 2005), making this method an efficient one despite the impossibility of taking all users into account, as mentioned before. So, a woman whose actual name is Jane Doe calling herself Paula Martins on Twitter would still be relevant to me. It must be acknowledged, however, that with this technique, I am operating from a binary gender perspective, and thus I am not able to account for other, non-binary categories. Further studies may be needed to go beyond these distinctions on social media.

The first thing needed to infer gender then is a list of gendered first names. These lists would be matched against the first names of users in my corpus, to attribute automatically a gender to each user depending on the presence of their first name in one list or the other. Other studies in which gender was inferred have used the lists of names from the US Social Security Administration (Mislove et al., 2011; Bamman et al., 2014) for example. However, although these lists are interesting, they are representative of names given in the United States. This project is focused on the United Kingdom, so I need to have lists of names representative of the region under study. Thus, I decided to turn to the data provided by three institutions, namely the Office for National Statistics (ONS), the National Records of Scotland (NRS), and the Central Statistics Office (CSO). These three organizations recorded the first names given to babies in each of the countries they represent: England and Wales for the ONS[92], Scotland for the NRS[93], and Ireland for the CSO[94]. These three entities are in charge of providing accurate and official numbers, and provide data back to 1965, which makes it even more interesting in order to have names which cover several generations to prevent a potential generational bias due to a too selective name-sampling. The ORS even provides the full lists of gendered names

[92] See the ONS website for more information regarding the lists of names. Last seen on February, 13th 2017. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datalist?filter=datasets

[93] See the NRS website for more information regarding the lists of names. Last seen on February, 13th 2017. URL: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/names/babies-first-names

[94] See the CSO website for more information regarding the lists of names. Last seen on February, 13th 2017. URL: http://www.cso.ie/en/releasesandpublications/er/ibn/irishbabiesnames2015/

given to babies every year since 1974, which makes this method even more exhaustive. The other two organizations are not as exhaustive depending on the period, but at least provide lists of the 100 most popular names for a given year and gender. Because of the amount of details given in certain cases (full lists of names, number of babies given these names, etc.), these lists are both exhaustive and representative of the region I focus on. This means that it is one more way to further exclude foreigners who would tweet from the United Kingdom, but who would not have a name included in this list of British names, in order to solely focus on British patterns. Of course, many non-British people will have first names which exist in their country of origin as well as in the UK, and will be taken into account in the corpus, but this still represents one more way to guarantee a better representativeness of the overall data. Thus, by using a combination of all the data available from these various sources and between 1965 and 2010[95], I have been able to build two lists of 32 825 female and 21 563 male first names.

One problem with this technique however is that many names are ambiguous and can be given to both women and men (like "Robin" for example). These names were detected because they were present in both lists, and there were 3169 of those. A choice had to be made concerning how I would consider these names, otherwise I would risk an outcome of some users being labelled with the wrong gender. I could, as other researchers have done (see Bamman et al., 2014), base myself on the majority gender assigned to each ambiguous first name. This means that I would need to consider the proportion of women or men given the ambiguous first name, and above a certain threshold the first name would be considered to be mainly male, or mainly female. The problem with this method is that there is, to my knowledge, no reliable and exhaustive source of data giving the detailed proportions of baby boys and girls given ambiguous first names around the UK. So, taking all these names into account would risk creating too much bias because of a more or less random attribution of a gender. On the other hand, discarding every ambiguous name would also mean discarding a lot of data, as some of these names are popular ones, like "Adam" or "Abbie", thus creating another form of bias. One solution was to do a compilation of the top 100 first names by gender over the 1965-2010 period mentioned above, and whenever an ambiguous name was present in the top 100 for one gender, this would be considered as this name's majority gender. The full list of different names present in the top 100s reached 205 different names for males, and 321 for females. The limited variety of names present in these top 100s shows that the most popular names have been relatively

[95] Although I do realize that none of the users I will collect data from will be born in 2010, this was one more way to have access to more data.

steady over the years, therefore implying that the ambiguous names present in these lists are strongly associated to their corresponding gender. However, even while taking into account these top 100 names, two first names were still present for both genders, namely "Charlie" and "Taylor", and were simply removed from both lists. Then, all the 3169 ambiguous names were deleted from both lists, and the ambiguous names present in the compiled lists of top 100 names were added back into the total list of first names of the corresponding gender. Deleting so many names may seem like a big loss, but as as the "top 100 boys' names accounted for 52% of all baby boys born in 2015, while the top 100 girls' names accounted for 43% of all baby girls born in 2015" [96], this reinforces the idea that focusing on the top 100 names alone ensures an appropriate representativeness of each gender, and reinforces the idea that still having roughly 50 000 names on which to base my methodology, an appropriate representativeness should still be reached. This should then limit the bias caused by the removal of the other ambiguous names as much as possible. Some more manual checking of the data revealed that some names were composed of one letter only, like "A" or "R". Such names have also been removed to prevent the wrong association of "A" used as the indefinite article in the name field of the Twitter profile with the actual name of the user, as in the case of someone calling themselves "A random person" for example.

After these removals, other elements have been manually added to these lists of male and female names. As explained earlier, manually checking and reading the data is crucial in understanding how it is organized, and to become aware of elements one may otherwise overlook. Although it is not possible to manually read the complete set of 22 million tweets, reading the name section of several thousands of them helped me take into account frequent indications of gender not based on first names only. Indeed, it is fairly frequent for Twitter users to name themselves "Mr" or "Miss" followed by their last name, or even "King" followed by any other name. If I was to strictly use lists of names, these markers would not be taken into account, and as we will see later, the program would fail to identify the gender of users which could easily be determined with such markers. Again, someone calling themselves "King of the North" is obviously the sign of a pseudonym being used, but as I am only interested in attributing a correct gender to users, and not in the faithfulness of the name provided, the particle "King" is relevant to me. Thus, I manually added "Miss", "Mrs", "Ms", "Mister", "Mr", "King", "Queen",

[96] See the Statistical bulletin from the ONS for the year 2015 in England and Wales. Last seen on February 13th, 2017. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/2015

"Prince", "Princess" to the respective lists of gendered names.

After both lists of names were created, cleaned and refined, the program to automatically assign a gender to users could be written. It was necessary to use programming to do that, as the size of the file could not be treated by more traditional software like Excel. Thus, I used the programming language Python[97] to write the programs allowing me to sort the corpus in whichever way I needed[98]. Because the vast majority of users who declare a first and last name, do in fact supply their first name as the first element of the "name" field[99], I will consider that the first word appearing in the name field of the user's profile will be the first name. This word will then be compared to the names appearing in the lists described earlier, and if the candidate name strictly matches one of the gendered names, then the corresponding gender will be assigned to this user and the tweet and its metadata will be written to a new file comprising all the tweets of a single gender. Two things need to be highlighted here. First, considering that the first word appearing in the name field will be the first name could be seen as a limitation, we could instead prefer to compare all the words present in the name field. Thus, we would successfully detect "Vador Mark" as being a male user, which is not the case with the program I wrote. The reason why I did not do that is simple; if I used this method, I may also consider "Jenny Thomas" to be a male user, as well as any user whose last name can also be a first name indicative of the opposite gender. To prevent any such confusion, and because a manual observation of the corpus seemed to indicate almost no user reversing their first and last name, I chose to stick to the method trying to match the first word of the field to a name present in the lists. The second point I want to highlight is the necessity of *strictly* matching a name present in the lists of names. If this was not the case, this would mean that the first name "Rob" would be detected in "Robert" as well. In this case, this is not a problem as both first names are male. However, if the user is called "Robin Davis", this user, who may be a woman, would be assigned to the male corpus. The advantage of strictly matching the names present in the list is that it avoids such confusions. However, because Twitter users' use of case is not always consistent, all names (those present in the lists of gendered names and the candidate names declared in the name field) are basically treated as lowercase, except in special situations detailed later.

[97] See https://www.python.org/about/ for more details about this programming language. Last seen on March 2nd, 2017.

[98] I need to once more thank Adrien 'Lopez' Guille, who helped me finalize some of the most complicated programs I will later describe!

[99] This was determined according to my manual checking of the data, although I must acknowledge that I do not have more large-scale, scientific evidence confirming that Twitter users who declare a first and last name state the first name first.

So, strictly matching names present in the list greatly improves accuracy, but also has other limitations. As mentioned earlier, there is no obligation for Twitter users to provide a valid name, and users are sometimes very creative with the names they choose, but also with the way they spell them, and many special characters are present in the name field. Thus, I had to proceed to another form of cleaning for the names to be properly detected as such. Indeed, in the case of "C/H/R/I/S" for example, it is easy for a human reader to determine what the actual name is, but as I explained, it is not as simple for the program I use. Creativity then, and the use of punctuation characters and other symbols is a barrier to the correct automated association of a first name to a user, and can thus be problematic for identifying gender. So, in order to maximize the chances of accurately determining users' first names, I chose to automatically delete a number of characters commonly present in the name field of the corpus, and which are not of any help in identifying the name of a user. So, the following characters were removed: . + / ° § , : ; ? ! "" ( ) { } [ ] @ # * $ € & = < > © ◊ ~ ≈ % £ ≠ ® † |
When these characters are deleted, they are not replaced by anything, thus linking the surrounding characters together. So, in the example used earlier, "C/H/R/I/S" would become "CHRIS". I added a different rule for the characters _ and – however. As these characters are frequently used to separate the first name from the last name, as in the case of "Paula_Harris" for example, I simply replaced these characters by a space, thus giving "Paula Harris". However, for cases where the user spelled their name "P-A-U-L-A" for example, the program will fail to identify the end result (i.e. "P A U L A") as a proper first name, and it will be discarded.

Also, I noticed that, as it is not uncommon for users to call themselves by their title (e.g. Ms, Mr etc…), some people also use the title "Dr". Because it is most of the time followed directly by the first name of the user[100], I also deleted all instances of "Dr" or "Dr.", as it does not give any indication regarding the user's gender.

Another challenge in identifying first names is that users do not always separate their first name from their last name with a space, as in the case of "BruceWayne" for example. Here again, a human reader easily makes the distinction between the two, mainly thanks to the capital letter at the beginning of the last name, but for a computer program, this would be considered as one

[100] Still according to my manual checking. It would, here again, be interesting to have more large-scale evidence supporting these observations.

single word and would thus fail to recognize it as a gendered first name. Users who do not separate their first name from their last name very often use capitalized letters to make a visual distinction between the two, so one solution is to write a program checking the number of words found in the name field. If the number of words is one, then the program looks for capital letters inside it, and if the capital letters are in the middle of the word (i.e. if they are not the very first letter of the word), then a space is added right before the capital letter. Then, "BruceWayne" would become "Bruce Wayne". However, users who do not capitalize anything (apart from the first letter of the field) and do not separate their names in any way (e.g. "Brucewayne") will not be identified and will thus not be taken into account because the program will not recognize the first name in it. Again, this may be seen as a limitation of the program strictly matching first names, but users not making any visual or graphical distinction between their first and last name seem to be a small minority[101], so this method is very likely to allow the identification of more users than the ones it fails to recognize.

Finally, some users use numbers or special characters to replace letters inside their name (e.g. "k8ie", "3LL107"), which prevents the identification of these names. As the numbers used in such cases can usually replace one single letter, it would be easy to automatically replace numbers by the letters they represent. So for example, "3" would be replaced by "e", "1" by "i" etc… However, this would also replace numbers which are used in other ways, and may create new kinds of confusions for the program, as in the case of "Richard33" for example, which would become "Richardee", and would not be recognized as a first name. Thus, numbers were not taken into account, and users having names like "k8ie" were simply discarded.

Taking into account all these parameters, and bearing in mind the methodological choices made, the program inferring gender could be written. It is presented in the figure below:

[101] Once again, it would be difficult to have detailed statistics regarding the number of users who provide a gendered first name *and* who do not use any sort of boundary between this and their last name. The same applies to users who spell their name "P-A-U-L-A", and who would not be taken into account, as in the example given earlier. Thus, I realize that some methodological choices are more based on manual and visual observations than on empirical evidence, and I concede that these may be seen as limitations.

```python
#!/usr/bin/env python3
import csv
import codecs
import re

def camel_case_split(identifier):
    matches = re.finditer('.+?(?:(?<=[a-z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])|$)', identifier)
    return [m.group(0) for m in matches]

list1 = set()

# read the list of male or female names to add them to a set
with codecs.open("full_list_female_names_UK.csv", encoding= "utf-8", errors= "ignore") as listnames:
    names = csv.reader(listnames, delimiter = ",")
    for row in names:
        list1.add(row[0].lower())

# format the name section of the corpus to remove all the unnecessary
# characters, properly format caps and add write all that to a new file
with open("UK_ultimate2_sorted.csv") as input:
    input_corpus = csv.reader(input, delimiter = ",")
    with open("gender_detection_corpus_female.csv", "w") as writer1:
        corpus_writer = csv.writer(writer1, delimiter=",")
        for row in input_corpus:
            row[6] = row[6].replace(",", "").replace(":", "").replace(";", "").replace("?", "").replace(":", "").replace(";", "").replace("!", "").replace("?", "").replace("\"
            caca = row[6].split(" ")
            if row[6] == "" or row[6] == " ":
                row.append("NoName")
                corpus_writer.writerow(row)
            elif len(caca) == 1:
                firstname1 = camel_case_split(row[6])
                row.append(firstname1[0]) # adds a column with the first name
                corpus_writer.writerow(row)
            else:
                row.append(caca[0])
                corpus_writer.writerow(row)

# read the name section of the corpus to detect the gender of the author of each tweet,
# and write each properly detected tweet to a new file
with open("gender_detection_corpus_female.csv") as last_reader:
    final_reader = csv.reader(last_reader, delimiter = ",")
    with open("female_tweets.csv", "w") as last_output:
        last_writer = csv.writer(last_output, delimiter = ",")
        for row in final_reader:
            firstname = row[7].lower()
            for n in list1:
                if n == firstname:
                    last_writer.writerow(row)
```

Figure 6.1: Python 3 gender inferring program

Figure 6.1 may not mean much to readers who do not have programming experience, and I am not going to comment on it in detail, but including this is still crucial for several reasons. The first one is, as I mentioned earlier, out of a concern for transparency about the methodology used in any study. By being transparent about the very program used to sort the data, any flaw in the methodology can be made apparent, and future studies may build on that to further enhance this methodology. Also, people who do not have a background in programming, or who are learning it and would like to carry out a similar study may use this program as a basis, and adapt it to their needs, potentially saving some time.

So, after running the program on the corpus in order to extract tweets and their metadata from users for whom a gender could be associated, a total number of 6 406 581 male and 4 442 194 female tweets were retrieved.

Now that a gender has been assigned to a certain number of tweets, it is necessary to infer the age of these users. This process is detailed in the next section.

2.6.2 Inferring age

Determining users' age, as for their gender, requires some form of inference as this is not

a piece of information which is required in users' profiles either. So, users do not need to provide their age, and not all of them do so, thus some sorting will also have to be done for users who actually provide their age. Those who declare an age do so in their "Bio", which is the description one can add to one's Twitter profile, in order to briefly describe oneself or one's motivations for being on Twitter. The bio part of one's profile is limited to 160 characters, so users can be slightly more descriptive than in tweets. Thus, the bio can be very different from user to user, and some may be very personal, as in the following example:

> (#001)[102] hi am a 12 year old girl in hartlepool i have adhd and dyspraxia and im just starting my first year in senior school

And some others may be much less personal, as in:

> (#002) Fermanagh / Liverpool

In the first example, the user's bio gives details regarding her age, where she lives, details concerning her health and information on her school curriculum. In the second example, we only have cities mentioned, so we may guess that this is the place where this user lives, but we cannot be sure of that, as it may also be the football teams she[103] supports, or the cities she loves. So, the information one can retrieve from the bio can be very diverse, if the user even has a bio, as this is optional, and left bank by some Twitter users.

For this study, the information I am interested in retrieving from the bio is the age only, and as we saw in the first example given above, some users clearly mention their age. However, as studies mentioned in Part 1 suggested, on the Internet, and on social media in particular, some conventions develop and ways of expressing emotions, or ways of providing information, can rise and be specific to certain media. So, we have to wonder if there are ways of giving one's age other than illustrated in the example above, which consists of giving one's age followed by "year old". This is important, since the process of retrieving age data has to be automated thanks to a computer program, which focuses on fixed patterns, as we saw earlier with names. At first, it may seem tempting to write a program so that any number provided by the users corresponds to their age. In the example given above, this would work, but what if the user mentions how

[102] All along this thesis, this convention will be used to refer to examples (of tweets, or parts of users' profiles) taken directly from the corpus.

[103] This bio is taken from a user who has been detected as being female.

many cats they have? What if they mention how long they have been playing football? What if they mention an address? A date? All these would be incorrectly associated to the user's age, so unfortunately the program has to be a bit more discerning.

The first thing I did then, was to manually read users' bios. Once again, getting to know one's corpus is crucial in understanding how to efficiently analyze it, and this allowed me to become aware of various patterns I did not notice before. Thus, I realized that there are 4 main ways for users to declare their age in their bio:

1) The age is given at the very beginning of the bio, as in the example below:

(#003) 24. London. Speech and Language Therapy student #dream. 'be useful and be kind'.

This way of providing one's age is by far the most popular one, as among the four ways taken into account, this one was used in 72.9% of all the cases I later detected. Thus, the way the program works is that it reads the bio part of the profile, and if this starts by two digits, then these two digits are to be considered as the user's age. However, by analyzing the results of the first tests, I quickly realized that this method had certain limits. Indeed, bios starting with a date, or an address for example, would be incorrectly associated to the user's age. Also, bios starting by something like "21st century guy…" would incorrectly match 21 as being the age of the user. To prevent this, I had to add exceptions to the rule so that bios starting with two digits, but not followed by anything which corresponds to the format of a date, or not followed by a third digit or by suffixes like "-st", "-nd", "-rd" or "-th", would be considered as the user's age. This solved the vast majority of the issues I first encountered, and allowed me to detect the age of many users thanks to that method alone.

2) The age is given in between two non-alphanumeric characters, as in the example below:

(#004) Shannon | 21 | Netball | Physio Student

This is the second most frequent way of providing one's age in my sample, and accounts for 19.8% of all the mentions of one's age I detected thanks to the methods I used. However here again the problem of dates remains. Indeed, most dates are represented under the form

“dd/mm/yyyy”, the month being necessarily enclosed between two of the characters the program is looking for. So here again, I had to add exceptions so that dates were ignored.

3) The age is clearly given as a number followed by “years old”, as in the following example:

(#005) Engineering student | 21 years old | MotoGP ❤ #93 #73 #19 #25 #32 #52 #50(58) #36 ❤ Music ❤ Rock&MetalMusic & PopPunk ❤ RedBull ❤ Instagram: @yolandaa_95

This way of declaring one’s age seems to be the most obvious, yet it only accounts for 5.3% of the age mentions detected. It has to be noted that not only “years old” as such was taken into account when detecting ages thanks to this method, and many variations of it were considered, like “yrs old”, “yo”, “y o”, “yr old” etc… Thus, here the program is looking for any sequence of two digits followed by one variant of “years old”. However, it is not uncommon for users to mention the fact that they have children, as well as their children’s ages. In this case, something like “I have a 30 year old daughter” would incorrectly be recognized as being a mention of the age of the user. So, to prevent this I added an exception saying that if the sequence in question was followed by “daughter”, “son”, “children” or “kid”, this should not be considered as the user’s age.

4) The age is clearly given, preceded by “I am”, as in:

(#006) I'm Alice, Im 26, I'm Mad On Cricket, US tv shows (too many) reading and music, also official sweetie girl to the Yorkshire boys!!!

This way of declaring one’s age is very close to the previous one in terms of the structure. Here again, “I am” is just one possible orthographic variation of it, and “I’m”, or “Im” were also taken into account. So, the program is looking for any of these followed by a two-digit number. Here again, testing the program revealed that some exceptions had to be put in place, as in the case of “Im”, the program would also match any word ending in “-im” and followed by a number[104], which caused some problems, so I had to add a rule saying that variations of “I am”

[104] Here again we realize the importance of strictly matching patterns…

were to be matched strictly, otherwise they were discarded. Here again, despite this way of declaring one's age being relatively obvious, this is far from the most popular as it only accounts for 1.9% of all the mentions of users' age.

Thus, after all these parameters were taken into account, the final program used to detect the age could be written, and is presented in the figure below:

```
#!/usr/bin/env python3
import csv
import re

with open("female_tweets.csv") as input_corpus:
    corpus_read = csv.reader(input_corpus, delimiter=",")
    with open("female_corpus.csv", "w") as output_corpus:
        corpus_write = csv.writer(output_corpus, delimiter = ",")
        for row in corpus_read:
            regex = re.compile(r"(\d{2})\s?(years old|yo|yr old|y o|yrs old|year old)\s*(son|daughter|kid|child)?")
            matches = regex.findall(row[5].lower())
            if len(row[5]) > 2 and row[5][0].isdigit() and row[5][1].isdigit() and not row[5][2].isdigit() and not re.findall(r"(\A(\d{2})[^a-zA-Z0-9_(\s)]+\s?\d{1,2}\s?[^a
                age = row[5][0] + row[5][1]
                row.append(age)
                corpus_write.writerow(row)
            elif len(row[5]) == 2 and row[5][0].isdigit() and row[5][1].isdigit():
                age = row[5][0] + row[5][1]
                row.append(age)
                corpus_write.writerow(row)
            elif len(matches) > 0 and matches[0][2] == "":
                age = matches[0][0]
                row.append(age)
                corpus_write.writerow(row)
            elif re.findall(r"([^a-zA-Z0-9_(\s)]+\s?\d{2}\s?[^a-zA-Z0-9_(\s)]+)", row[5].lower()) and not re.findall(r"((\d{1,2})[^a-zA-Z0-9_(\s)]+\s?\d{1,2}\s?[^a-zA-Z0-9
                age = re.findall(r"([^a-zA-Z0-9_(\s)]+\s?\d{2}\s?[^a-zA-Z0-9_(\s)]+)", row[5].lower())
                age = age[0]
                age = re.sub("[^a-zA-Z0-9_(\s)]+", "", age)
                row.append(age)
                corpus_write.writerow(row)
            elif re.findall(r"(i'm|i am|im)\s?\d{2}", row[5].lower()) and not re.findall(r"([a-zA-Z]+im)\s?(\d{2})", row[5].lower()):
                age = re.findall(r"(\d{2})", row[5].lower())
                age = age[0]
                row.append(age)
                corpus_write.writerow(row)
```

Figure 6.2: Python 3 age inferring program

Again, I wish to emphasize the importance of manually checking the data, as this played a major role in determining the exceptions which had to be handled. Had I not manually checked every step of the program as I wrote it, I would have missed a lot of cases incorrectly considered as the users' ages. Even after the program was written, manually comparing the files before, and after age was detected allowed me to quantify the precision and recall of the method. Indeed, after the detection of the age, each tweet (and its metadata) associated to a user for whom a valid age was discovered is written to a new file, so that for women for example, there is in the end one file for female tweets (with no age detected yet), and one file with female tweets to which an age has been attributed. In other words, the corpus is saved in two versions; one which represents the corpus before running the program, and another one after running it. By manually comparing these two versions, it is possible to realize what is missing, or what has incorrectly been included. I manually compared these two corpora for the female tweets,

and read through the bios of the first one thousand tweets of the input corpus[105] to assess how well the program behaves.

Among this sample of one thousand bios, three were incorrectly included in the final corpus, meaning that three ages were incorrectly attributed to users. These were:

(#007) BPSA Annual Conference Organiser 2015-16 // […]

(#008) 20 yrs from now […]

(#009) 1% humour, 99% terrible puns. […]

In the first case, the error is due to "16" being included between the two non-alphanumeric characters "-" and "/". In the second case, the error is due to "20" being present in the first two characters of the field, and not followed by any of the exceptions mentioned earlier. In the last case, the error is due to "99" being in-between "," and "%". These exceptions being somewhat specific and isolated (a precision of 99.7% being what can be considered as an acceptable one), I decided not to try to find ways to overcome them. I could have added an exception saying that for example, two digits enclosed inside two non-alphanumeric characters, but the second one being a "%" sign, this should not be considered as the age. Again, this sampling error being an isolated one, and the "%" sign being at other times used as a character enclosing actual ages for other users, I was afraid that I would slightly improve precision at the cost of recall[106]. In other words, I was afraid that I would lose as many, or even more tweets than I would gain by adding an exception. Precision is high then, which is good, but what about recall? In other words, among all the cases where Twitter users provide their age in their bio, how many has the program missed? Again, out of the one thousand bios I manually checked, 13 were missed by the program. These include cases like:

(#010) Worshipping yer da since 1989

(#011) nineteen || bristol

(#012) Just turned 34 work as a […]

[105] i.e. the corpus composed of the 4 442 194 female tweets before running the program detecting the age. Although one thousand tweets only represents a tiny fraction of the corpus (0.02% to be exact), and thus cannot be considered as accurately representative of the exact precision and recall of the method, this was meant to have an overall idea of the effectiveness of the program, and check for major issues which would have to be addressed before going further, as manually checking the whole corpus is not feasible. However, additional manual benchmarking would need to be done on a greater number of users, and also not just on women, in order to be more representative.

[106] In statistics, recall being the proportion of relevant matches following a query compared to the ones it misses. In other words, recall is considered as high if among all the patterns one is interested in, most of them are taken into account by the query. On the contrary, if the query misses a lot of the features one is interested in, recall will be low.

(#013) […] ’99 baby […]

In the first case, it can be assumed that 1989 is the year of birth of the user, and in this case, it would be very complicated for the program to automatically detect 1989 as a birth date, and then to convert that into an actual age. Concerning the other cases, they are either related to the first one, or are relevant to situations mentioned earlier for which there would be no real solution for me to efficiently detect those as actual ages (at least none that I am aware of). Again, it might be possible to improve recall, but I consider that achieving a recall of 98.7% is reasonable and that my programming skills would not allow me to increase this figure substantially, so I decided to keep this program as it is.

The figures regarding precision and recall indicate that Twitter users actually use a limited set of methods to indicate their age, and that being aware of these alone can ensure the collection of most of the material one is interested in. Beyond the mere statistical aspect of it, these figures also indicate that Twitter users developed their own linguistic standards as a way to overcome the limitations imposed by the medium while still conveying specific information. These standards seem to be recognized and understood by everyone, indicating that the online community has been able to adapt and create new modes of communication, whereas users do not have any obligation to display their age on Twitter. Actually, users mentioning their age are a minority, as can be seen in the following table:

	All users	Age only inferred	Gender only inferred	Age + gender inferred
Females			4 442 194 tweets (23.7%*)	395 709 tweets (2.1%*) (8.9%**)
			205 705 users (27.8%*)	13 536 users (1.8%*) (6.5%**)
Males			6 406 581 tweets (34.2%*)	418 127 tweets (2.2%*) (6.5%**)
			254 273 users (34.3%*)	11 689 users (1.5%*) (4.5%**)
Total	18 709 729 tweets (100%)	1 373 468 tweets (7.3%*)	10 848 775 tweets (57.9%*)	813 836 tweets (4.3%*) (7.5%**)
	739 221 users (100%)	39 379 users (5.3%*)	459 978 users (62.2%*)	25 225 users (3.4%*) (5.4%**)

Table 6.1: Social distribution of Twitter users in the corpus for age and gender (*compared to the whole corpus ; **compared to users for whom gender was inferred)

A majority of users declare a gendered first name, which reveals that although anonymity still seems important for a lot of people, most of them feel comfortable with revealing at least a part of their identity (i.e. their gender).

Comparatively to gender however, relatively few users reveal their age in their bio. It seems then that age is considered more personal information that Twitter users are more reluctant to reveal. However, it is interesting to note that those who do reveal their age are also very likely to provide a gendered first name, as out of the 39 379 users reporting their age, 25 225 (64%) also reported a gendered first name. It also seems that women are more likely than men (6.5% compared to 4.5%) to report their age.

As explained earlier in Part 1, age is a key factor determining how we express ourselves, be it in face to face interactions or on social media. In order to better account for the impact of age

on the speech patterns of Twitter users, I decided to split users into various age groups. Thus, I took into account four different age groups, namely 12-18, 19-30, 31-45, and 46-60, ending up with eight final sub-corpora, four sub-corpora for male users, and four sub-corpora for female users. The reason why I chose these groups in particular 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, it is assumed that 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.

Also, studies suggest that people who have children produce more standard forms than usual and tend to avoid the use of taboo language (Stapleton, 2003; Mercury, 1995). Thus, parenthood is likely to influence linguistic attitudes, hence the need to take that into account in my age classification as well. According to the Office for National Statistics, in 2014 the average ages of mothers and fathers were respectively 30 and 32 in England and Wales;[107] this is why age 30 was chosen as a delimiter for two age groups. I decided not to take into account users above 60, as people in the 46-60 group already represent a very small minority of users, so creating an age group which would take into account users above 60 years old would not be useful, as it would not be statistically relevant due to the extremely limited amount of data. Users below 12 have also been discarded because it seemed that Twitter users below 12 are very sparse, as only 7410 tweets from users detected as between 6 and 12 years old were present for both males and females. On top of the limited representativeness this little amount of data would represent, manually checking the data also indicated that the vast majority of the hits found were due to errors, such that the age data had incorrectly been associated to these users. These tweets were then discarded.

The remaining age groups should then allow the heterogeneity of the sub-corpora to be limited as much as possible. Such groupings 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

[107] See the 2015 report from the Office for National Statistics. Last accessed on March, 15th 2017. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsbyparentscharacteristicsinenglandandwales/2014

claim to be 22 in their descriptions, whereas they may now be 23 or 24. The age reported would in this case not be the actual age of the user, but they would still be associated to the correct age group. Even with such precautions however, errors are bound to happen with people whose age overlaps with two age groups and who did not keep their profile up to date.

Table 6.2 below summarizes the number of tweets and users populating each age group for both genders:

	12-18	19-30	31-45	46-60
Female	141 154 tweets	218 088 tweets	17 125 tweets	5 257 tweets
	4 845 users	7 898 users	342 users	151 users
Male	103 184 tweets	251 713 tweets	27 295 tweets	14 208 tweets
	3 229 users	6 839 users	661 users	335 users
Total	244 338 tweets (31.4%)	469 801 tweets (60.3%)	44 420 tweets (5.7%)	19 465 tweets (2.5%)
	8 074 users (33.2%)	14 737 users (60.6%)	1 003 users (4.1%)	486 users (2%)

Table 6.2: Number of tweets and users for each age group and gender

Unsurprisingly, this table reveals that although users are pretty evenly distributed across both genders, there are substantial imbalances in the representation of the different age groups, with 19-30 being a vast majority, both in terms of the volume of tweets emitted, but also in terms of the number of users. This was expected, because this corresponds to the data presented in the 2016 Ofcom report[108] on the demographics of social media sites users in the UK, as they state that in 2015, people aged 16-24 and 25-34 were the most likely to report using Twitter (39% and 40% respectively) (2016: 75). There is a difference however, between the number of people from a certain age group reporting using Twitter, and the number of Twitter users belonging to a certain age group, and although they may be related, the two are not necessarily interchangeable. Access to the demographics of Twitter users is not transparent though, and there are no openly accessible official records giving details regarding the age groups in which

[108] Last accessed on March, 16th 2017. URL: https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjgvJzZz9rSAhWHOBoKHetjAPIQFggpMAA&url=https%3A%2F%2Fwww.ofcom.org.uk%2F data%2Fassets%2Fpdf_file%2F0026%2F80828%2F2016-adults-media-use-and-attitudes.pdf&usg=AFQjCNFiiLW5GTmuV3fDYc5YA7dQEFkpWg&sig2=korKJDLiaJpg4Za0HZiwKQ

users belong, therefore figures from the Ofcom report are used as equivalent means of assessing the relevance of the figures derived from the corpus.

Table 6.2 also reveals that the proportion of users above 30 is very small, which confirms that taking into account users above 60 would not make sense, as the data would not be representative of anything, and thus not usable.

URLs and mentions:

A final measure taken to prepare the corpus for the analyses was to format the URLs and mentions so that I can better account for them. Indeed, URLs are frequently used in tweets to share photos, videos, articles, or any web-based content. The same holds for mentions, which is one way on Twitter to interact with others. Mentions are used by using the "@" symbol followed by the Twitter name of the user, so that the latter can be notified that another user mentioned them in a tweet, and create an exchange. Being able to account for these two aspects of content sharing on Twitter can potentially be important in determining gendered or age patterns. However, each URL and mention being unique (referring to a single web page or user), it can be difficult to easily take them all into account to calculate which gender uses URLs the most for example. So, to make things easier, I decided to replace all the URLs by **URL**, and to replace all mentions by **@mention**. The asterisks serve as a way to understand that "URL" or "@mention" has not been used as such inside a tweet, and that these refer to the standardized format I decided to adopt to spot these more easily. This format will later serve as a way to group all of these under the same labels and better account for them during the analyses.

Conclusion

This chapter presented all the steps which had to be undertaken for the corpus to be usable for in-depth analyses. In the end, roughly 4% of the tweets which were collected remain and fully correspond to the requirements I originally had, that is to have access to both the gender and the age of the authors. Although this figure may seem small, it must be remembered that figures indicate that this corpus probably corresponds to the full set of geolocated tweets emitted during the collection phase. So, although small compared to the original numbers, these roughly one million tweets represent the full repository of tweets matching the criteria I focus on, so this corpus can be considered to be highly representative of the population I am interested in.

With this in mind, and knowing that no more processing and formatting will need to be done, the actual analysis of the corpus can be carried out, and this will be the goal of the following and final part.

Part 3: Results

The first two parts of this thesis aimed respectively at providing enough background to understand the motivations of this study, and detailing the methodology used to build the corpus. In the previous chapter, I explained that all the tweets collected had been sorted in various sub-corpora according to the gender and the age of the users. The goal of this third part will be to analyze and compare how swear words are used *across* these corpora (i.e. comparing women to men for example), but also *inside* these corpora (i.e. do all men aged 19-30 use *fuck* uniformly for example?). To this end, various analytical strategies and tools will be used, going from the more general, or top-level ones, to the most specific ones, generally focusing on single words or tweets.

Chapter 7 will present various quantitative figures detailing the frequency of use of swear words in the corpus, by gender and age group. A particular emphasis will also be laid on the word fuck for reasons detailed later.

In Chapter 8, I will focus on swear words with a statistical approach to determine which swear words are the most significant among each gender and age group. To do this, I will use the Mann-Whitney U test, and the simple maths parameter to carry out a keyword analysis. This chapter will be an opportunity to study gendered and age tendencies regarding the use of swear words, but also regarding the topic which are favored by these sub-groups.

Chapter 9 will be the most qualitative of the three, and will focus in depth on patterns and (swear) words which have been shown as salient in some regards in the two previous chapters. These analyses will mainly be based on collocations and by isolating certain tweets which will be shown to be representative of certain trends.

Chapter 7: Some general figures

[A] corpus does not contain new information about language, but
the software offers us a new perspective on the familiar.
(Hunston, 2002: 3)

This concept is a familiar one by now, and the fact that linguists have been using computers to analyze corpora (as corpora are used to analyze language) has been extensively dealt with earlier. As we have seen, software and computers are very good at certain repetitive tasks often used in linguistic studies, like counting words, or generating statistical calculations. These processes are used in linguistic studies to analyze corpora with varying degrees of details. The goal of this chapter will be to provide an overall view of the corpus, so that we can proceed from there to further analyses and dig ever deeper to more focused aspects of the data.

In 3.7.1, I will present the distribution of every swear word in the corpus (i.e. the ones considered as such in this study and presented in Chapter 3), and proceed to preliminary observations on the most striking aspects of the data.

Section 3.7.2 will focus on the case of *fuck*, which is by far the most used swear word in the corpus, and I will compare its distribution here to the way it is used in the BNC to try to understand what may affect the differences and similarities observed between the two corpora.

Section 3.7.3 will present preliminary data regarding the number of vulgar tweets sent according to each gender and age group.

In section 3.7.4, I will analyze how evenly or unevenly distributed swear words are as a whole (and not just vulgar ones) between both genders and all age groups. This will help us gain a different perspective on the data, and to better understand how each gender and age group behaves.

In section 3.7.5, I will zoom in on one of the most striking aspects of the data, which is the generational gap there seems to be between two age groups in particular.

3.7.1 Swear word distribution according to gender and age

As explained earlier, corpora are most of the time so big that manually reading them entirely is an impossible task. Thus, an appropriate methodology has to be adopted in order to answer the research questions which originally motivated the investigation. Because of the size of most corpora, directly aiming for specific cases and isolated examples would be taking the risk to miss important aspects of the data, which in turn would render a proper interpretation of these specific cases impossible. It seems logical then to go from the broadest analyses, to gradually narrowing the scope in order to analytically zoom in on the most noticeable patterns. One of the first and easiest steps which can be taken when analyzing a corpus for linguistic purposes is to calculate word frequencies. This provides an overall idea of the distribution of the variables one is interested in, and word counts are also necessary for a lot of more fine-grained approaches like keywords or statistical analyses. However, the absolute frequency[109] (AF) rarely represents an efficient way of objectively comparing different corpora, which is ultimately what I aim at doing by comparing the way males (one corpus) use swear words compared to the way women (another corpus) use them. Indeed, unless the corpora are of the exact same size (in terms of the number of tokens), it is useless to know that a variable was used 150 times by men, and 200 times by women. It may at first seem like the variable is used more by women, but if the female corpus is composed of 1000 words, and the male one of only 500 words, then the situation is completely reversed. This is why most of the time when presenting the frequency of certain variables it is preferable to mention the normalized frequency (NF), which in other words is the average number of times a particular variable appears for a given number of words. In the previous example, we could say that the variable is used 300 times per 1000 words for men, and 200 times per 1000 words for women. Normalized frequencies provide an objective means of comparing two or more corpora, so we can confidently assert that the variable is used more often by men in our hypothetical example. In the case of Table 7.1 below, a NF of 19 for females aged 12-18 means that for every 1000 words present in tweets of this sub-group, 19 of them will be *shit*.

The same procedure has been applied for the number of occurrences of every one of the swear words taken into account among women and men from the various age groups, the results (NF per 1000 words) are presented in Table 7.1:

[109] The AF being the number of times this word appears in a corpus.

Age groups	All		12-18		19-30		31-45		46-60	
Gender	F	M	F	M	F	M	F	M	F	M
fuck (all)[110]	37.9	42.4	44	51.9	34.5	43.3	6.4	27	11.2	15.6
shit	16.8	18.2	19	21.3	16.2	18.4	3.6	13.5	4.9	8.5
bloody	3.9	2.8	3.4	2	4	2.9	4	4.1	4.1	3.1
piss	4.1	3.5	5.2	4.2	3.6	3.3	1.6	3.3	1.5	1.9
fuck (ab)[111]	4.1	2.9	5.2	4	3.6	2.6	1.1	3.8	2.8	1.4
cunt	1.9	4.7	2.3	5.7	1.7	4.6	0.2	3.6	0.5	2.3
bitch	3.8	1.7	4.2	1.9	3.7	1.7	0.9	1.2	1.1	0.9
hell	2.4	2.1	2	1.8	2.6	2.3	1.5	2.1	0.9	1.4
ass	2.2	1.8	2.2	1.8	2.2	1.9	1.8	1.5	0.1	1.2
crap	1.1	1.6	0.9	1	1.1	1.7	2.1	1.8	1.9	3
damn	1.7	1.7	1.4	1.7	1.8	1.8	1.4	0.9	0.3	1.6
bastard	0.6	1.5	0.4	1.4	0.7	1.6	0.6	1.3	0.7	0.8
prick	0.8	1.2	0.9	1.3	0.7	1.2	0.1	1	0.3	1.4
dick	0.6	0.6	0.8	0.6	0.6	0.6	0.5	0.5	None	0.3
bollock	0.2	0.6	0.1	0.4	0.2	0.6	0.4	1.1	0.3	1.2
wanker	0.3	0.5	0.2	0.5	0.3	0.5	0.2	1	0.1	0.5
tits	0.4	0.3	0.4	0.3	0.4	0.3	0.4	0.4	0.5	0.3
bugger	0.2	0.2	0.1	0.1	0.1	0.1	1.1	0.2	None	0.7
retard	0.1	0.4	0.1	0.8	0.1	0.2	None	0.2	None	0.3
cock	0.1	0.2	0.1	0.2	0.1	0.3	0.9	0.4	None	0.07
pussy	0.2	0.4	0.3	0.6	0.2	0.4	0.1	0.1	0.1	0.07
slut	0.3	0.2	0.3	0.2	0.3	0.1	0.3	0.3	0.1	0.07
fag	0.2	0.2	0.2	0.5	0.2	0.1	None	0.1	None	None
nigger	0.1	0.3	0.2	0.6	0.1	0.3	None	0.03	None	0.07
cum	0.05	0.1	0.06	0.1	0.02	0.09	0.4	0.1	None	None
whore	0.1	0.1	0.1	0.1	0.1	0.09	None	0.1	None	0.07
blowjob	0.01	0.01	0.007	0.01	0.01	0.02	None	None	None	None

Table 7.1: Normalized swear word frequency for all age groups

[110] This includes the word *fuck*, its derivatives, as well as the abbreviations derived from it and expressive lengthenings (e.g. *fuuuuuuuccckkkkkk*).

[111] This only includes the abbreviations derived from the word *fuck,* its abbreviated derivatives and expressive lengthenings (e.g. *fk*, *fck*, *fking*, *fukkkkkkk,* etc…).

In this table, the swear words have been classified in decreasing order of frequency based on their average overall frequency. Unsurprisingly, *fuck* and its variations is the most popular curse word, alone covering 39.8% of all the swear word occurrences, followed by *shit* (20.6%), *bloody* (5%), *piss* (4.7%), and the abbreviations of *fuck* (4.6%). The least popular swear word of this set is the word *blowjob*, which only represents 0.01% of all the occurrences of swear words. Note however that this data is not truly representative of actual figures, as the derivatives of *fuck* are counted twice in these statistics (once with the word itself, and once with the abbreviations only), so the actual statistics for each word are slightly lower than that[112].

There is a sharp drop then between the two most frequent swear words and the rest, with *fuck* and *shit* alone accounting for 60% of all instances of swear words. Although the distribution is a bit different, these observations are in line with Wang et al. (2014) who found that the words *fuck* and *shit* were also by far the two most frequent swear words in their sample of tweets, accounting for 33.5% and 15.4% of all the occurrences of swear words respectively. The distribution of swear words is therefore not balanced at all, and all users, no matter what their age or gender is, prefer using the two words *fuck* and *shit* (with the exception of women 31-45 maybe, who use *bloody* slightly more often than *shit*) a lot more than any other word. Apart from *fuck* and *shit*, which can be considered as outliers here, we observe that the 11 most frequent swear words break away from the others, and that the remaining 16 swear words are used much more infrequently, to eventually become marginal for the least frequent ones. This also is in line with what Wang et al. found, as they said that "the top seven curse words – *fuck*, *shit*, *ass*, *bitch*, *nigga*, *hell* and *whore* cover 90.40% of all the curse word occurrences" (Wang et al., 2014: 419). Although in the case of my corpus, the top twelve swear words (not counting the abbreviations of *fuck*) have to be taken into account to reach the 90% threshold, this shows that the distribution is very skewed, and that a minority of swear words account for the vast majority of swear word occurrences. In our case too, four out of the seven swear words reported by Wang et al. (2014) are present among the top seven swear words (still not counting the abbreviations of *fuck*) in this corpus. It must be recalled that in the case of Wang et al., their corpus was based on the global stream of tweets detected as being in English, and not focused on one particular region. The trends they observed can thus be considered as representative of general trends English speakers from all around the world displayed on Twitter. The fact that

[112] However, the difference between the two is so small (0.2 max) that I decided to leave the figures as they are in order to keep things simpler, and not confuse the reader with two different measurements.

there is an alignment between their top seven swear words and ours indicates that there is a global convergence regarding the use of English swear words, and that some words can generally be considered as the most widely used English swear words (at least on Twitter).

Also, we can observe that in Table 7.1, many swear words do not appear at all among the two older generations. This can partly be explained by the tendency for older generations to swear less, as mentioned earlier, but in this case it must once more be acknowledged that what probably plays a bigger role is the relative lack of data regarding these two generations. Indeed, some swear words do not occur at all among the two older generations. This is not necessarily a problem, as some swear words are still used fairly often for older users, which is the sign that if a swear word is adopted widely enough as part of the repertoire of a category of users, even restricted, this word will appear in the results. So, even if many swear words do not occur at all, without concluding that these words are never used by this age group[113], it can safely be assumed that they are not part of the most widely adopted ones either. Moreover, the vast majority of the words which do not appear at all are also part of the least represented ones for the younger generations of users as well. This is the sign that on top of not being massively used by older people, these words are also part of the least popular swear words for users as a whole (i.e. the bottom part of the table). I am not going to do a detailed review of the gender or age differences in swear word usage now, because as we have seen before, and as we will see later in greater detail, the data presented in Table 7.1 being based on aggregate data (not taking inter-speaker variation into account), more detailed analyses may lead to erroneous interpretations. The aim of this table is to first have an overall idea of the distribution of swear words, without paying attention to the variation between individual users.

In addition to this, Figure 7.1 and Figure 7.2 present the absolute frequencies of swear words for women and men as a whole:

[113] Let us not forget that absence of evidence is no evidence of absence.

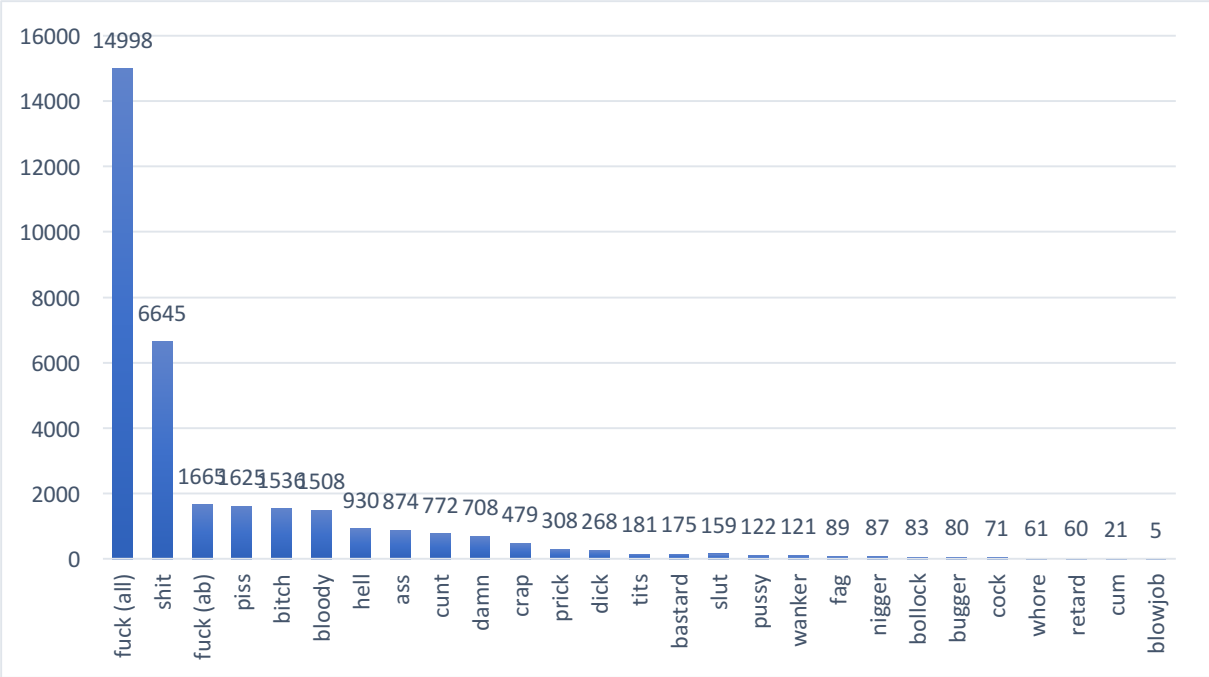


Figure 7.1: Raw frequencies of swear word usage for women

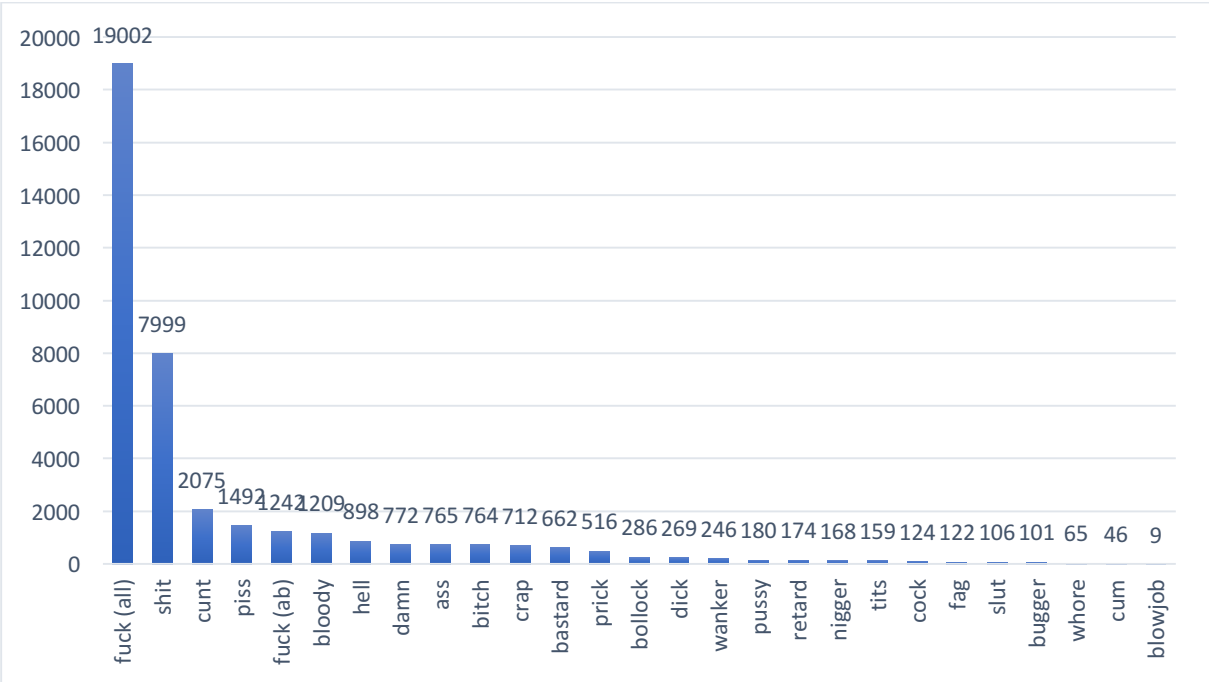


Figure 7.2: Raw frequencies of swear word usage for men

Compared to Table 7.1, these figures have the advantage of giving visual indications regarding the representation of each swear word, which in this case clearly reinforces the salience of the most used words compared to the least used ones. In addition to this, such graphs also allow to see the rank in which each word appears for both genders. In other words, we can better

apprehend the gendered preferences regarding swear words. Simply by visually comparing the ranks for both genders, we can notice that, apart from a few exceptions, most swear words appear at a similar rank (position in the graph). This would imply that there may be an inter-gender alignment regarding swear word preference, and that although frequencies may vary (men swearing more overall), both genders share the same perceptions concerning which words are to be used more often. However, bar charts are not the most adapted method to study the rank in which swear words appear, and on top of that, it would be interesting to take a look at ranks for each age group as well. To this end, Table 7.2 will help us compare these:

Age groups	All		12-18		19-30		31-45		46-60		Age diff	Gender diff
Gender	F	M	F	M	F	M	F	M	F	M		
fuck (all)	1	1	1	1	1	1	1	1	1	1	0	0
shit	2	2	2	2	2	2	3	2	2	2	1	1
fuck (ab)	3	5	3	5	5	6	9	4	4	9	7	3
piss	4	4	4	4	6	4	6	6	6	6	6	2
bitch	5	10	5	7	4	10	11	11	7	13	16	14
bloody	6	6	6	6	3	5	2	3	3	3	9	3
hell	7	7	9	9	7	7	7	7	8	8	2	0
ass	8	9	8	8	8	8	5	9	15	11	8	0
cunt	9	3	7	3	10	3	20	5	10	5	**17**	**31**
damn	10	8	10	10	9	9	8	15	12	7	4	2
crap	11	11	11	13	11	11	4	8	5	4	25	5
prick	12	13	12	12	12	13	21	13	12	9	**6**	**10**
dick	13	15	13	15	14	15	14	16	No	17	No	No
tits	14	20	15	21	15	18	16	18	10	17	**8**	**18**
bastard	15	12	14	11	13	12	13	10	9	14	4	2
slut	16	23	16	23	16	22	18	19	15	20	**5**	**19**
pussy	17	17	17	16	18	17	21	22	15	20	10	4
wanker	18	16	18	19	17	16	19	14	15	16	6	4
fag	19	22	19	18	19	23	No	24	No	No	No	No
nigger	20	19	20	16	21	19	No	26	No	20	No	No
bollock	21	14	25	20	20	14	15	12	12	11	29	15
bugger	22	24	24	24	22	24	9	21	No	15	No	No
cock	23	21	22	22	25	20	11	17	No	20	No	No
whore	24	25	23	25	24	25	No	24	No	20	No	No
retard	25	18	21	14	23	21	No	20	No	17	No	No
cum	26	26	26	26	26	26	16	22	No	No	No	No
blowjob	27	27	27	27	27	27	No	No	No	No	No	No

Table 7.2: Swear word frequency ranks according to gender and age

The first thing which can be noted is the fact that the main tendencies observed in Figures 7.1 and 7.2 are striking in Table 7.2 as well, notably the fact that *fuck* and *shit* rank 1 and 2 respectively for the vast majority of users from both genders. When taking a closer look at the table, we also notice that generally speaking, there is not much inter-gender variation inside each age group, and that what is the most marked is the variation between age groups, as there seems to be a particular degree of variation between the younger and the older age groups. In other words, users aged 12-18 and 19-30 seem to behave differently from users aged 31-45 and 46-60. In order to assess the degree of variation across gender and age more effectively than with visual verification only, the "Age diff" and "Gender diff" columns were added, each having a specific "score". In order to know the "Gender diff" score for the word *bloody* for example, the ranks for males in each age group have been added, as have the ranks for females in each group. The smaller sum is subtracted from the larger sum, yielding a variation score:
6 + 5 + 3 + 3 = 17 for males
6 + 3 + 2 + 3 = 14 for females
17 – 14 = a "variation score" of 3

The same procedure has been adopted for age groups, except that in this case we are focusing on the variation between the younger age groups (the 12-18 and 19-30) and the older ones (the 31-45 and 46-60). So, the totals for the 12-18 have been added to those of the 19-30, then the totals for the 31-45 and 46-60 and the smaller sum has been subtracted from the larger. This method is a simple one which, to my knowledge, has not been used by anyone else. This may not be the most statistically accurate measure of variation, but it still allows us to see the main tendencies here, and in particular to compare the effects of gender and age for each swear word. Thus, we can observe that indeed, inter-generational variation plays a much bigger role than inter-gender variation. Indeed, the age variation scores are higher in 12 out of 18 cases (the 9 cases where there was no occurrence of a swear word were not counted), whereas gender scores were higher in 4 cases only (these are in bold characters, and underlined in Table 7.2). Here, what I mean by "variation" is simply an assessment of how similar or different women and men are in their ranking of swear words.

Age seems to play a crucial role in the way users will swear then, but it does not mean that gender has no impact. Actually, the five words for which gender plays a great role are displayed in red in Table 7.2, and these words are *bitch*, *cunt*, *tits*, *slut*, *bollock*. For these words, there is always a clear gendered tendency which is very marked in most age groups. *Bitch* appears to

be favored by women, and the other four appear to be favored by men. Without going much deeper into the statistical analyses, we could hypothesize that these words are the most gendered swear words. Of course however, more thorough calculations will need to be carried out to confirm or refute this, which will be the object of Chapter 9.

3.7.2 About the word *fuck*

Concerning the sheer frequency of appearance of swear words, it may be interesting to compare how often certain words appear in this corpus to how often they appear in other reference corpora. *Fuck* is by far the most frequent swear word in this study, and it has been shown to be the most frequently used swear word in several other studies previously mentioned as well. We will take the example of the BNC to try and compare the frequencies of *fuck* in the two corpora (i.e. the BNC and the one under study here) to have an idea of how frequently or infrequently the word appears in one or the other. McEnery and Xiao (2003: 506) found that in the spoken section of the BNC, the word *fuck* and its derivatives were used about 3232 times per million words by users aged 0-59. I decided not to take into account the data provided by McEnery and Xiao for users aged 60+ because I did not take them into account myself. I did, however, count users aged 0-14 in the reference study, whereas I did not take into account users below age 12 in my own study. This may create a bias, but having no means to know how many times users aged 12 and above used *fuck* in the reference study, I decided to still take them into account. Furthermore, this comparison is just used as an illustration, and as a way to have an overall idea of how comparable the two datasets can be. For this comparison, I chose to base it on the data concerning the spoken part of the BNC, because as mentioned earlier, tweets are more comparable to spoken than written language. Concerning the frequency of *fuck* in my corpus of tweets then, the word and its derivatives occur 79.4 times per thousand words for (37 + 42.4 in Table 7.1 above), which corresponds to 79 400 times per million words, which means that it appears about 24 times more often than in the spoken part of the BNC. This may at first seem like a huge difference, but a few things need to be put in perspective to understand this.

First, the contexts of utterance are not the same at all; the BNC data is composed of people who knew they were being recorded, and a lot of them were at their workplace, which in itself may prohibit the use of swear words. Tweets on the other hand are produced on the personal profile of the user, and are primarily addressed to users the person chose to add to their network. Also, although every user is supposed to know that their tweets might be visible or collected by third

parties, this outcome is probably not made as consciously obvious to the user as when wearing a microphone, or having a recorded conversation with a researcher, so people may feel more free to swear on social media because of the way they work to begin with. The number of derivatives of *fuck* may also be one way to explain this increase. Indeed, as I pointed earlier (see Section 1.2.2), one of the elements which may have triggered an increase in the use of *fuck* in the 1970s could be the appearance of derivatives of the word, which created new ways in which it could be used, thus explaining why it became more popular. This is the same with written creativity which has been on the rise with the appearance of new digital media, and the fact that variants of the word *fuck* like *wtf*, *ffs*, *fk*, *ftw*, etc. are taken into account in this study probably plays a major role in the difference between the frequency of *fuck* in this corpus and in the BNC. Finally, roughly twenty years separate the two corpora, which in itself might be another reason for a greater acceptability of the word. Thus, such comparisons are to be taken with a pinch of salt and not as definite empirical proofs of a major shift in the way *fuck* is used and/or perceived. However, these figures seem to confirm the increase in the use of *fuck* observed earlier (Section 1.2.2), and may also be confirming that an increase in the number of derivatives of a swear word can play a role in its spread.

The second thing which can be noted about the frequency of *fuck* in this corpus is that there seems to be a major gap between how often the younger generations (i.e. the 12-18 and the 19-30) use it compared to the older ones (the 31-45 and the 46-60). We have seen in the previous section that age plays a bigger role in determining how often swear words will be used, so age may be one explanation, and we could argue that despite *fuck* being the most popular swear word, the effect of age still influences its use among the older generations. Gender may also explain the gap there is, especially among users aged 31-45 who display a particularly striking gendered difference regarding how often *fuck* is used, with it appearing 6.4 times per thousand words for women, and 27 times for men. This is the age group displaying the greatest gendered variation for this word, and this also plays a role in how big the gap between the 19-30 and the 31-45 seems to be. However, even without this gendered gap, the use of *fuck* for men aged 19-30 still decreases from 51.9 times per thousand words to 27 times for men aged 31-45, which is a striking difference in itself. One last parameter which has to be taken into account to explain this imbalance between younger and older generations is the evolution of the word. Indeed, we saw in section 1.2.2 that the derivatives of *fuck* started to be widespread in the BNC from 1975 onwards. Although they may have existed before, their increasing presence in the BNC at this period indicates that the variants of *fuck* started being more popular at this point. The tweets on

which these figures are based were collected in 2016, so roughly forty years after the beginning of this phenomenon. We can then hypothesize that people before forty in this corpus have always been accustomed to the widespread variants of *fuck*, which may explain why they are more comfortable using it. This would partly justify why younger generations use the word much more than others, and why the older ones (those aged 46-60) may be less comfortable with it, having not been used to the variants all along. Concerning users aged 31-45 however, they overlap with the period when the derivatives of *fuck* became more popular, so we could expect them to use them more, but this is not really the case here. We could hypothesize further that the phenomenon needed some time to be massively accepted, which is why the figures are still somewhat lower for users aged 31-45, but of course, without more data from this period, this idea will remain speculative, and other factors may be at play here.

3.7.3 How often do swear words appear among users?

So, according to the figures presented above, it can be concluded that *fuck* appears a lot more than in the BNC. However, does it mean that swearing as a whole has become standard practice and that most people do it on Twitter? Table 7.3 summarizes the proportion of vulgar tweets (tweets containing at least one swear word) for both genders and all age groups:

		12-18	19-30	31-45	46-60	All
Women	Vulgar tweets	8.4%	7.1%	2.8%	2.8%	7.5%
	Average n° tweets	29	28	50	35	29
	Median n° tweets	8	7	7	7	7
Men	Vulgar tweets	9.4%	8.2%	6.2%	4.2%	8.6%
	Average n° tweets	32	37	41	42	35
	Median n° tweets	7	8	8.5	7.5	7

Table 7.3: Proportion of (vulgar) tweets emitted according to gender and age

Concerning first of all the number of tweets sent, Table 7.3 shows that overall, men send more than women. This is true for all sub-groups, apart from ages 31-45, among which women tweet more than men of the same age. Perhaps surprisingly, women 31-45 are the sub-group sending the greatest number of tweets on average. This is unexpected, because the age groups reported as using Twitter (i.e. having a Twitter account) the most are the youngest ones, so it could be supposed that these users would also tweet more. However, this does not seem to be the case

here, and this is true for both women and men aged 31-45, who tweeted on average more than the two youngest generations in both cases. This might be explained by the fact that fewer people aged 30 and up use Twitter, but those who do tend to be more active. For both women and men, the top ten users being the most active (in terms of the number of tweets they sent during the collection phase) sent between 1157 and 4328[114] tweets each during the period of collection. In other words, the range of tweets sent by each Twitter user in the corpus goes from 1 to 4328. Because of this gap between the people who have been the most active and those who only tweet from time to time, the average does not necessarily mean much. Indeed, because a limited number of users may have extreme attitudes and send a lot more tweets than the vast majority of users, this may pull the resulting average up and give the wrong idea of the most representative average number of tweets sent. To prevent this, I also considered the median number of tweets sent, which in this case may actually be more representative of overall attitudes on Twitter because of the phenomena I just explained. The median is another kind of average, and "is the middle value in a series of values ordered from the smallest to the largest" (see Brezina, forthcoming). In other words, it means that in an ordered list of values, there are as many values which are greater than the median as ones which are lower. Because the median only looks at the middle value then, potential extreme values are discarded and in cases where the distribution may be very wide, it may give more representative results. When looking at the median number of tweets sent then, we realize that women and men are actually much closer to each other. The results are the same for women and men considered as a whole, meaning that what gives the impression that men tweet more than women when looking at the average value is probably greatly influenced by a smaller number of male outliers tweeting a lot more than the rest. Here again however, we observe that users aged 31-45 are (at least for men) tweeting more than the other age groups, but on the whole the results for the medians are very close to each other for both genders, indicating that there is not much overall difference between the number of tweets sent by women and men.

3.7.4 Overall swear word use by gender and age

One of the main research questions this work is based on is whether women (and younger women in particular) swear more than men. The overall results indicate that men swear more than women overall, and this is also verified when comparing individual age groups as well. The sub-group swearing the most is men aged 12-18, with 9.4% of their tweets being

[114] This is the greatest number of tweets sent by a single user in this corpus, and this user is a woman.

considered as vulgar. As could have been expected, we observe a gradual decrease in the proportion of vulgar tweets with age, older generations swearing less than younger ones. Both genders follow this pattern, however the drop in swear word use between the two younger generations and the two older ones is much sharper for women, going from 7.1% of vulgar tweets for ages 19-30 to 2.8% for both ages 31-45 and 46-60. The drop among men on the other hand is much more gradual. It seems then that age plays a major role in determining swear word use for women, and age 30 seems to be a landmark from which swear words seem to become more sparse.

More detailed and qualitative analyses would be needed, but it seems that at least in terms of frequency of swear word use, age may be a more influential factor than gender alone. Also, because the frequency of swear word use is exactly the same for women 31-45 and 46-60, it may be possible that gender is a determining factor for women only, which would explain why beyond a certain age women swear much less than younger ones but remain constant in their frequency of use. This gender effect would be less marked among men, which would explain why the only visible trend seems to be affected by age only. Again, these are all hypotheses which will need to be checked later, as at this stage the level of details of the data is far too low to allow for a clear understanding of the phenomena at play here.

When looking at the overall swear word use by women and men, women use swear words in 7.5% of their tweets, and men in 8.6%. Comparatively, it is interesting to note that for users for whom only the gender was inferred[115], 4.9% of the tweets for females are vulgar, and 5.8% of male tweets are. This is noteworthy because it shows that:

- Among users who do not reveal their age, men are still more likely than women to use swear words.
- Users who declare an age are overall more likely to swear than those who do not.

This gap between users who reveal their age and those who don't is probably due to younger generations (12-30) who are a vast majority in the corpus of users for whom I have been able to infer an age. So, it is possible that younger generations as a whole may be more willing to mention their age than others, and because they also tend to use swear words more than older generations, this imbalance does not push the average use of swear words upwards as much

[115] In other words, for users who did not declare their age in their bio.

among people who do not declare their age. Alternatively, it may also be likely that the age distribution is the same among users who do not report their age, but that these very users feel less comfortable with swearing as a whole. There would then be a link between the willingness to mention one's age and the likelihood of swearing. Despite that, we note that the proportion of vulgar tweets for women and men is steady, i.e. men swearing slightly more frequently than women, but overall the main tendency still is that people do not swear at all.

To come back to the use of swear words for users for whom both the age and the gender was inferred, the question which needs to be answered now is: is the difference relevant? Arguably, the difference may seem important, especially because a one percent difference applied to billions of tweets, or words of speech can quickly represent a substantial amount of variation. On the other hand, in some of the cases analyzed earlier when reviewing previous studies pointing to gendered differences, in some cases the gap between women and men was far beyond one percent, so the difference between women and men in our case may be considered small. Because the point of view on this may be relative, using a statistical test to determine whether the difference observed is statistically significant or not may help:

> Users of a corpus must be aware of its internal variations, and researchers sometimes use statistical techniques to examine the degree of variability within a given corpus before using it. […] The degree of homogeneity of a corpus is then another factor in determining how well matched that corpus is to particular research questions.
> (McEnery and Hardie, 2012: 2)

As seen earlier (see Section 1-1-5), it has been shown that aggregate data do not mean much, and in this case the difference could merely be due to a handful of users using some swear words a lot more than the others, and thus bias the overall data. Because of this, parametric statistical tests which take the mean into account like the log-likelihood or chi-squared tests have been shown to lead to erroneous interpretations (Brezina and Meyerhoff, 2014). To prevent such errors of interpretations, I chose to use a statistical test which is non-parametric (which does not take the mean into account), and which also takes dispersion into account to prevent the potential interference of inter-user differences. The Mann-Whitney U (or Wilcoxson rank-sum) test was then used through the Lancs Toolbox interface[116] by providing the sum of the normalized frequencies of use of every swear word for each user. In other words,

[116] See "Statistics in Corpus Linguistics: a Practical Guide". Last visited on 11th April, 2017. URL: http://corpora.lancs.ac.uk/stats/toolbox.php?panel=5&tab=0.

I added[117] how many times each swear word appeared for one user to obtain a total number, converted that number into a normalized frequency and repeated the process for every user. In addition to the MWU statistic, the effect size was also calculated. Brezina (forthcoming) gives the following definition of the effect size:

> Effect size in descriptive statistics is a standardised measure, that is a measure comparable across different studies […], that expresses the practical importance of the effect observed in the corpus or corpora. For example, if we establish by a statistical test (see above) that two groups of speakers (e.g. men and women) differ from each other in the use of a particular linguistic variable, i.e. there is a statistically significant difference between these two groups, we still need to see how large this difference is and whether it is practically important.

In other words then, the effect size is used to measure how strong the tendency confirmed thanks to the MWU tests is (if it is confirmed at all). The measure of the effect size chosen here is Pearson's correlation (r) (see Cohen, 1988), and this measure is usually interpreted according to three cut-off points: 0.1 (small effect), 0.3 (medium effect) and 0.5 (large effect). In order to keep things relatively simple however, the effect size is mainly provided out of a concern for transparency, but it will not influence the interpretation of the data, or whether I use it to question the validity of a significant MWU value.

MWU tests were then run to compare the overall results for women and men from every age group, and all the results were found to be at least statistically significant at the level of $p<.01$ with a small effect size[118]. Thus, it can be concluded that according to the MWU tests, men swear significantly more than women in every age group, as well as when taken as a whole.

Considering these results then, it might be tempting to conclude that swearing is "proved" to be a male thing, and that traditional stereotypes presenting men as vulgar and women as profanity eschewers are true. However, this would be disregarding the main trend here, which is that the vast majority of the tweets for both genders and for all age groups is *not* vulgar, despite the difference in swear word use between women and men. As Baker pointed out:

> [A] potential problem with some methods within Corpus Linguistics is that they put researchers in a 'difference' mindset, privileging findings that reveal differences while backgrounding similarities. […] An alternative way of

[117] The computer did it for me to be exact.

[118] Generally speaking, the smaller the p-value the better, $p<.05$ being generally considered as the threshold above which the results are not considered as being significant. The threshold of acceptability can be lowered by the researchers if they want to only focus on "very significant" results, however.

> looking at gender differences would be to ask to what extent do differences outweigh similarities?"
> (Baker, 2014: 24, 25)

Objectivity here forces us to realize that despite the significance of swear word use by males, this only concerns a maximum of 8.6% of the tweets (for males as a whole), whereas more than 90% of tweets for both genders contain no swear word at all, which actually is the major tendency to note in this case, which shows that women and men display patterns which are more similar than different in this regard. It could be argued that the sheer volume of vulgar/non-vulgar tweets does not mean much, and that what is important is to know the proportion of users who use swear words. Indeed, if we hypothesize that the majority of users from one gender or the other use swear words, then we may consider associating swearing as a whole to one gender as more justified. Table 7.4 gives the percentage of users using swear words at least once for both genders and all age groups:

	12-18	19-30	31-45	46-60	All
Women	1969 (40.6%)	2771 (35%)	81 (23.6%)	32 (21.1%)	4853 (35.8%)
Men	1321 (40.9%)	2895 (42.3%)	221 (33.4%)	107 (31.9%)	4544 (38.8%)

Table 7.4: Proportion of users who used at least one swear word

Here again, we can observe the same trend as in Table 7.3; the majority of women and men from all age do not use any swear word. What is also worth noting is the sharp contrast there is once again between the two younger generations and the two older ones. As with the overall proportion of vulgar tweets, age seems to be a more determining factor deciding the proportion of users who will swear, as the gap between users aged 19-30 and 31-45 is much greater for both genders than between any other pair of age groups. Still concerning age, the same pattern can be observed which consists in gradually swearing less as users get older, with the possible exception of men aged 19-30, who are slightly more likely to swear than men aged 12-18. What is interesting to note is that there is a 10% gap between the number of women and men who swear among those aged 46-60, whereas for the younger generation this gap is almost non-existent (reduced to 0.3% only). In other words, gendered differences are much more marked among older users, and are less marked among younger users. It seems then that there is a gradual lowering of the impact of gender in determining how much swear words are used by women and men, at least in the case of Twitter users in the UK. The youngest generation is in this corpus that in which gendered differences are the least noticeable, first in terms of how

many swear words are used, but especially in terms of how many people use them. Although much more detailed analyses are needed to have an idea of how women and men use swear words in various contexts, these preliminary results seem to indicate that Thelwall was right when he predicted that "it seems likely 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" (2008: 102). It would indeed seem that gender equality in swearing has already been attained, at least in terms of the proportion of Twitter users using swear words. Concerning how much these users swear (Table 7.3), although the gap between women and men has been shown to be statistically significant, ages 12-18 comprise the generation where this gap is the smallest, so it seems possible that gender equality is indeed on its way.
One question remains completely unanswered though: what could explain the gap between users aged 19-30 and 31-45?

3.7.5 A gendered generational gap?

First, we have seen earlier (see section 1.2.1) that although people as a whole are more accepting regarding swear word use, the younger generations are more often than not reported to use swear words more often than older ones. So, from there, it is not surprising to observe a gradual decrease of swear word use across age groups. However as mentioned before, the drop is much more sudden between users aged 19-30 and 31-45 than between any other pair. One explanation for this phenomenon could be the very reason why I chose these age groups. As detailed in section 2.6.2, age 30 was chosen as a delimiter between two age groups because this age has for several years been repeatedly shown to be the average age of mothers in the UK[119]. As having babies has been confirmed in the literature (see Chapter 6) as influencing swear word use for parents, it seemed logical to take that into account. It may be possible then, that the combined effects of the generation users belong to, and having babies causes this gap between the 19-30 and the 31-45 for both women and men.

In order to have a better idea of the impact of age on swear word use, it may be useful to isolate this parameter and focus on this only. In other words, it may be interesting to look at the way people use swear words and focus only on people whose age can be inferred, disregarding their gender. This would then of course include the users whose gender has been inferred, but would also include every user who mentioned their age in their bio, but who did not provide enough

[119] The average age of fathers being slightly higher on average.

information in the name field of their Twitter profile to determine their gender. The same could be done for gender only, and we could look at the way swear words are used by women and men whose age is not mentioned. This basically means isolating each parameter, the gender and the age, to observe their influence on swear word use. This may allow us to better determine the extent to which one parameter may be more influential than the other, or on the other hand, we may realize that what is the most influential factor is the interplay of both gender and age. Table 7.5 summarizes the findings for users belonging to each of those categories:

		12-18	19-30	31-45	46-60	All
Women	Gender + Age	8.4%	7.1%	2.8%	2.8%	7.5%
	Gender alone	5.5%				
Men	Gender + Age	9.4%	8.2%	6.2%	4.2%	8.6%
	Gender alone	6.4%				
Age only		10.1%	9%	5.2%	4.8%	8.9%

Table 7.5: Proportion of vulgar tweets according to age, gender, and age + gender

As can be seen in Table 7.5, the tendency for males in the gender-only corpus is the same as the one observed earlier, which is that they tend to swear more than women overall. We also notice what was observed before about the fact that declaring one's age is linked with swearing more frequently. Indeed, apart from users aged 31-45, the proportion of vulgar tweets for each age group of the age-only corpus is greater than that of men whose age is known. In other words, for almost all age groups the proportion of vulgar tweets in the age-only corpus outweighs that of the most vulgar sub-group in the gender + age corpus. This confirms the idea that revealing one's age is related to being more comfortable with swearing whether users are male or female. The fact that this is true for almost all age groups allows us to go beyond what I hypothesized earlier when I said that the overall tendency for users revealing their age to swear more could be influenced by the preponderance of younger generations on Twitter. Indeed, this may be true for overall results, but this tendency to swear more when revealing one's age is also observed among the 46-60 (age only), and partly among the 31-45 (age-only), although again, men 31-45 swear more than the 31-45 as a whole in the age-only corpus. Thus, there seems to be a definite link between revealing one's age and swearing, which may be explained later when we proceed to analyzing the corpora in greater depth.

What is probably even more worth noting however, is the fact that, what I am going to call the "generational gap" between the proportion of swear words between users aged 19-30 and 31-45, is once more present here in the age-only corpus. In Table 7.4 we noted that this generational gap was only present among women, and that the decrease in swearing with age was much more gradual for men, thus leading to the hypothesis that this phenomenon could mainly be influenced by gender. In Table 6.2 we noticed that in the gender-only corpus, men are both greater in the number of users present, but that they also tweeted a lot more than women, and they were additionally seen to tweet more than women in the age + gender corpus. We may therefore make the assumption that this tendency also exists in the age-only corpus, and because we observed that men swear more than women in every case we have analyzed so far, as well as being on the whole more active, their attitudes should prevail in the patterns observed in the age only corpus. However, this is not entirely the case, as we do observe a greater proportion of vulgar tweets in the age-only corpus (in this case the influence of men's tendencies would hold), but we also notice the marked presence of the generational gap. This is striking, because if this gap was indeed a female feature only, because men are both more present and active, this gap should be smoothed out and be almost unnoticeable, but in this case it is. There is a paradox then, between the greater proportion of swear words observed in the age-only corpus and the marked presence of the generational gap, which cannot be explained by the supposed preponderance of male tweets and them swearing more. One explanation could be that this generational gap actually exists for both women and men, but then why is it not present in the gender + age data presented in Table 7.4? For now we have no way of clearly answering these questions, and we will have to come back to this observation when we analyze the corpus in greater detail.

Overall then, both gender and age seem to be relevant factors influencing the use of swear words, in that gender will generally trigger a slightly greater use of swear words by men, and age will trigger a gradual decrease of swear word use as users get older. These effects have been confirmed both when analyzing these parameters independently, and when combining them. The impact of age seems to be bigger than that of gender however, in that it seems to be what is at the center of the generational gap observed, and also because the difference between the youngest age group and the oldest one is much greater than the gap between women and men in any situation.

On the matter of the relevance of age compared to gender in analyzing the use of swear words, Brezina and Meyerhoff's results (2014: 19) on the use of *fuck* and its variants in the BNC point to the same tendency, as they found that age was the only statistically significant factor when studying the social parameters playing the greatest role in its use. This does not mean that gender is not worth studying of course, as our preliminary results just proved the contrary, but it means that 1) women and men may not swear in a way which is different from one another and 2) age, as we have already seen, is of key importance in addressing these issues. In the case of Brezina and Meyerhoff (ibid), they realized the prevalence of age only when using a statistical test which was adapted to the parameters in question (i.e. the Mann-Whitney U test), as otherwise, as when using the log-likelihood test for example, a lot more parameters were (erroneously) found to influence the use of *fuck*. As explained earlier, this is one more sign that the interpretations made from aggregate data alone cannot be considered reliable, and can certainly not be the only means of analysis of the data. So far, although I have applied the Mann-Whitney U significance test to various (sub-)corpora, the detailed data presented regarding the use of each swear word was based on aggregate data (the NF per 1000 words). This is interesting to have a preliminary idea of the words which are the most frequent, but it is not detailed enough to allow for more relevant interpretations. To address this lack of detailed analyses, Chapter 2 will dig deeper into the data thanks to analyses based on non-aggregate data (among others).

Conclusion:

According to these results then, and especially because men have been shown to swear more than women overall, can we already conclude that swearing is a male thing, and that Thelwall (2008) was wrong in his predictions that strong swearing may become a female preference?
Again, what we have been reviewing so far is just the "big picture", and I previously gave numerous examples of cases where this big picture provided a distorted image of a much more complex reality. So indeed, we can definitely not conclude that, at least quantitatively speaking, "strong swear words[120]" are becoming a female thing only because *cunt* and *fuck* as a whole have been shown to mainly be used by males. However, we also observed earlier in Table 7.1 that when focusing on the abbreviations of *fuck* only, these were mainly used by women. So, this already nuances the picture as far as *fuck* is concerned, as it could be hypothesized that women are re-appropriating the word in new ways thanks to these abbreviations. Only a more qualitative approach can allow us to look at the contextual uses of the words for both genders, and these may reveal other forms of gendered preferences which are invisible at this stage. So, although a quantitative approach seems to partly present swearing as being a male tendency to some extent, much more analyses need to be carried out before validating such a dichotomy. Lastly, it must once more be recalled that although the difference in the proportion of swear word use by women and men has been shown to be statistically significant, the similarity between them, which is the fact that the vast majority of them do not swear, completely outweighs any gap there may be between their use of any swear word. Beyond the mere gender dichotomy, these results allowed us to realize that age may trigger more variation than gender alone, and that the interplay of both parameters needs to be taken into account for the full range of relevant contrasts to be visible.

[120] As a reminder, "strong swearing" was mainly composed of *fuck* and *cunt* for Thelwall.

Chapter 8: Who prefers what?

Instead of being conditioned by available CL [Corpus Linguistics] tools, the selection of a statistical measure should be dependent on the research question.
(Paquot and Bestgen, 2009: 246)

The quotation above illustrates the idea that before using a tool to carry out analyses, researchers should reflect on what they want to highlight, and whether the tool and the statistics behind it are appropriate in this regard. This is of prime importance here, as we will start to analyze the data more thoroughly, and the statistics will play a bigger role in figuring out which patterns are salient among women and men. So far, I have discussed why the MWU tests were more relevant in my case, and why they have been used before to determine whether men could be considered as swearing significantly more often than women. This statistical test will be used in more detailed analyses in this chapter as well, but I will also resort to other procedures in order to analyze keywords in every sub-corpus. This will allow us to have a more fine-grained picture of the swear words (but not only) which are the most salient among each age group and gender, and will help us to better understand the contexts in which Twitter users swear.

In section 3.8.1 then, MWU tests are applied to each individual swear word in every age group in order to see if certain swear words can be considered as being used significantly more by one gender or the other.

Section 3.8.2 provides a detailed analysis of the words (either vulgar or not) which are considered as keywords when comparing both genders. This will highlight the words and topics which are preferred by women and men inside each age group.

In section 3.8.3, I carry out the same analysis, this time focusing on intra-gender variation, in order to go beyond the mere opposition between males and females, and better understand the role played by age for both genders.

3.8.1 MWU tests

In the previous section, Mann-Whitney U (MWU) tests were applied to assess whether one gender, or one age group, could be said to swear more than the other. This procedure can also be applied to individual (swear) words, to determine whether some words are significantly more used by one sub-group than another. Using statistical tests in such a way is frequent in sociolinguistic studies, and it has been done in a lot of the studies mentioned previously and which also analyzed the use of swear words by certain social groups (McEnery and Xiao, 2003; Thelwall, 2008; Murphy, 2009). Again, the procedure itself is a relevant one, and being able to tell which swear words are disproportionately frequent by one gender can be key in understanding gendered patterns. What is problematic however, is when the statistical tests used to measure this are based on aggregate data (on the mean). As explained on several occasions now in this thesis, taking into account individual variation is crucial to be sure that a few outliers are not solely responsible for producing statistically significant results. Unfortunately, most of the studies analyzing social uses of swear words did not take into account individual variation. Indeed, McEnery and Xiao (2003) for example used the log-likelihood scores only, Thelwall (2008) used chi-squared tests, and Murphy (2009) used the frequencies per 1M words. These three procedures are based on the mean values, and are thus unable to take into account the potential interference of a handful of individuals using the variable a lot more than the others for example. This does not mean that previous results cannot be trusted, as the interpretations based on aggregate data can still be relevant of certain tendencies, but knowing that such tests can be misleading should encourage researchers to be more careful in their choice of statistical tests.

For this study, MWU tests were chosen for calculating which swear words could significantly be considered as gendered, so tests were run for every one of the swear words for each age group, each time comparing the data for women and men of the same age group. In other words, MWU tests were run to determine if among the 12-18-year-olds, certain swear words are used significantly more by one gender and can thus be considered as more male or female, and the same has been done for every other age group. The results are presented in Table 8.1:

Swear words	All	12-18	19-30	31-45	46-60
fuck (all)	Male	Male	Male	Male	Neutral
shit	Male	Neutral	Male	Male	Neutral
bloody	Female	Female	Neutral	Neutral	Neutral
bitch	Female	Female	Female	Neutral	Neutral
fuck (ab)	Female	Neutral	Neutral	Neutral	Neutral
cunt	Male	Male	Male	Male	Neutral
piss	Neutral	Neutral	Female	Male	Neutral
crap	Male	Neutral	Male	Neutral	Neutral
bastard	Male	Male	Male	Neutral	Neutral
prick	Male	Male	Male	Neutral	Neutral
bollock	Male	Male	Male	Male	Neutral
wanker	Male	Male	Male	Neutral	Neutral
retard	Male	Male	Male	Neutral	Neutral
cock	Male	Male	Male	Neutral	Neutral
pussy	Male	Neutral	Male	Neutral	Neutral
cum	Neutral	Neutral	Male	Neutral	None
bugger	Neutral	Neutral	Neutral	Neutral	Neutral
ass	Neutral	Neutral	Neutral	Neutral	Neutral
hell	Neutral	Neutral	Neutral	Neutral	Neutral
dick	Neutral	Neutral	Neutral	Neutral	Neutral
damn	Neutral	Neutral	Neutral	Neutral	Neutral
slut	Neutral	Neutral	Neutral	Neutral	Neutral
fag	Neutral	Neutral	Neutral	Neutral	None
nigger	Neutral	Neutral	Neutral	Neutral	Neutral
tits	Neutral	Neutral	Neutral	Neutral	Neutral
whore	Neutral	Neutral	Neutral	Neutral	Neutral
blowjob	Neutral	Neutral	Neutral	None	None

Table 8.1: Gendered tendencies for each swear word and age group[121]

[121] Out of a concern for readability, this table only features the main tendencies (female, male or neutral), but it does not present the MWU scores. A tendency was considered as male or female if the p-value was at least significant at the level of $p<.01$.

Table 8.1 presents the tendencies revealed thanks to the MWU tests. Before carrying out the analyses, it was decided that the p-value of $p<0.01$ would be the threshold below which a swear word would be considered as displaying a gendered tendency. If the MWU test indicated a gendered preference for the word in the age group in question thanks to a p-value below 0.01, then the corresponding gendered tendency is indicated. The colors are here to spot the tendencies more easily.

Looking at Table 8.1, and especially at the more salient colors, we may be tempted to conclude that males predominate, and thus that swearing appears to be a male activity according to the data. Indeed, overall there are a lot more cases in which the tendency is in favor of males rather than females. To be exact, a male tendency can be observed in 36 cases, and a female one in seven cases only. However, there are a total of 135 cases, which means that although male tendencies clearly outweigh female ones, we still cannot assert that swearing is mainly a male thing. This is true when taking all cases into account, but this tendency also holds when looking at specific age groups. Among the 19-30-year-olds, which is the age group in which the greatest number of male tendencies can be observed (males predominate in twelve cases), there are still thirteen cases which are neutral, and two being female. Table 8.1 once more shows that in most cases, no clear tendency emerges regarding a gendered preference for swear words.

Apart from the predominance of cases where swear words are neither male nor female, we observe that there is no word displaying a tendency for the 46-60. Again, this is probably due to the limited amount of data for this age group. Many swear words appear fewer than five times (or do not appear at all), so from there it is difficult to have statistically significant results, and to observe any relevant trend. Because of this lack of data, the 46-60-year-olds are not going to be taken into account for this part of the analysis. According to this data then, we can consider that *fuck*, *cunt* and *bollock* are the swear words which are preferred by males, as they appear as being significantly more used by them in every age group, as well as when taking all age groups into account. Similarly, *bitch* and *bloody* can be considered the swear words which are preferred by women, although *bloody* "only" appears as being significantly used among the 12-18-year-olds and when taking all age groups into account.

The fact that *fuck* appears as one of the most significant words for males is striking. Considering that, as we have seen earlier, it is the swear word which is by far the most used by Twitter users, we could have expected both genders to use it with a relatively similar high frequency. The

same applies to *shit*, which is also by far the second most used swear word, but here again, we nevertheless observe a male tendency to use it more than women in most cases (not among the 12-18-year-olds). However, despite this overall male preference, what was observed in our earlier observations is partly confirmed here: women prefer the abbreviations of *fuck*, and they use them significantly more than men. These are noticeable differences, but without more details regarding the contexts in which these words are used, we have for the moment no way of knowing the reason why men use these two words significantly more than women, and why women seem to prefer the abbreviated forms of *fuck* more than men.

Concerning inter-gender variation, here again, there is not much of it. As seen earlier, most words are neutral, and for words for which there is a gendered preference, apart from one exception (i.e. *piss*), the preference across age groups is unilateral. The fact that very few words appear as gendered among the two older age groups could give the impression that swear words are more equally distributed among these users, but it must be reminded that the lack of data may partly explain this here again. As we have seen, it probably plays a role in explaining why users aged 46-60 do not display any gendered preference, so we could wonder how much this affects users aged 31-45 too. However, even if the relative lack of data were to affect the MWU results for the 31-45-year-olds, it must be noted that a certain number of words still display a preference regardless, so the effect, if present, seems to be minimal compared to that present among users aged 46-60.

In Table 7.2, we estimated thanks to the rank scores that the words displaying the greatest gendered tendencies were *bitch*, *cunt*, *tits*, *slut* and *bollocks*, and that the most generational ones were *bollock*, *crap*, *cunt* and *bitch*. Here, with the more accurate individual MWU tests, only *bitch*, *cunt* and *bollock* appear as being gendered, *bitch* being preferred by females, the latter two by males.

So, thanks to this data we can confidently assert that, contrary to what Thelwall (2008: 97) predicted, strong language (i.e. *fuck* and *cunt*) in the UK is still dominated by males, to a certain extent at least. The swear words *fuck* and *cunt* are not used more by women on social media (at least not Twitter), as actually these results prove the contrary. Indeed, these two words are included among those which are the most representative of men as far as their spread across the different age groups is concerned. However, some of the results may nuance this conclusion to a certain extent. As mentioned earlier, we noticed that the word *cunt* is the one which has

evolved the most in terms of its frequency of use. What is the most striking is that males from the younger generations do not use the word so much more than men from older generations, as *cunt* is ranked as the fifth most frequent word for both the older generations, and as third among the younger ones. Women, on the other hand, display very different patterns, at least among the 31-45-year-olds. For women of this group, *cunt* is in rank 20, then in rank 10 for the 19-30, and in rank 7 for the 12-18. This is the greatest difference in rank for a swear word in this study. So, this could be the sign that Thelwall was wrong in saying that strong language is no longer dominated by males, but he was probably right when he said that "gender equality in swearing […] will slowly become more widespread, at least in social network sites" (2008: 102). Although these figures still support the idea that most "strong" swear words are used more by males, they also indicate that the gap between women and men is closing fast on all kinds of swear words, and maybe even more quickly on the strongest ones, as we have seen with *cunt*. Thelwall (2008: 97) acknowledged that the cases where he observed that women used strong swear words more than men were "not statistically significant in either case", but beyond the mere statistical evidence, what may be even more important here is the tendency he noticed, which this Twitter data seems to confirm.

However, it has to be noted that for a reason which seems to go against this hypothesis, *cunt* is ranked tenth in frequency among female users aged 46-60. This may be due to the lack of data preventing a more accurate ranking, which seems to be the more likely explanation, but this may also be due to a bad interpretation of the results. So, in order to be able to better interpret the figures and tendencies observed so far, we will dig deeper into the data via a keyword analysis of the various sub-corpora.

3.8.2 Keyword analysis: inter-gender variations

Keyword analysis is one of the first and most common means linguists resort to when analyzing a corpus. A widely accepted definition of a keyword is that it is "a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind" (Scott, 1997: 236). In order to determine what word is "key" in a corpus, one has to compare the frequency of a word in a corpus to what is commonly called a reference corpus. Thus, if a word occurs significantly more often in the corpus of interest than in the reference corpus, that word will be considered as a keyword. A widely used measure of significance for this is, here again, the log-likelihood (LL)

statistic. Most of the tools used by linguists (e.g. AntConc, WMatrix, WordSmith Tools) to generate a keyword analysis rely on the LL. Beyond the fact that this statistical test does not take dispersion into account, as pointed to earlier, this procedure has been criticized on other aspects in the context of keyword analysis. With regards to tests of significance (like the LL), Rosenfeld and Penrod (2011: 84) have observed that:

> Tests of statistical significance are dependent on the sample size used to calculate them. [...] With very large sample sizes, even very weak relationships can be significant. Conversely, with very small sample sizes, there may not be a significant relationship between the variables even when the actual relationship between the variables in the population is quite strong. Therefore, different conclusions may be drawn in different studies because of the size of the samples, if conclusions were drawn based only on statistical significance testing.

In other words, analyzing keywords with tests of significance on bigger corpora will lead to more keywords being discovered, many of which will merely be due to a great number of occurrences, and not necessarily to a meaningful difference between the two corpora.

Another problem directly derived from relying on the significance only is that the keyword list is generated in decreasing order of significance, with the most significant keywords topping the list. The problem is that the evidence we have with such lists, is simply that there is a significant difference between the two corpora, but we do not know how big this difference is in terms of its frequency. Gabrielatos and Marchi (2012[122]) summarize this clearly by saying that "LL measures statistical significance, not frequency difference". So, the risk is that the LL may highlight a contrast in a word appearing ten times in the corpus of interest and one time in the reference corpus. In itself the ratio of 10:1 is impressive, but would be much more so if the frequencies were of 10000 and 1000 in the corpus of interest and the reference corpus respectively[123]. This problem is partly why effect sizes are being more and more frequently implemented, in order to "measure […] the practical significance of a result, preventing us claiming a statistical significant result that has little consequence" (Ridge and Kudenko, 2010: 272).

For these reasons, I decided to avoid corpus tools using the LL. However, when reading about many other statistical tests used in keyword analysis (%DIFF, Log Ratio, simple maths

[122] See their 2012 conference presentation. Last accessed on April, 22nd 2017. URL: https://repository.edgehill.ac.uk/4196/

[123] See Kilgarriff (2012: 5) for more details regarding this and the solution he proposes (i.e. the simple maths parameter).

parameter, t-test, Cohen's D), I realized that there was no real agreement for one test which could be considered reliable in every situation. Instead of going on at length about each of these tests, I will quote Brezina (forthcoming), who noted that "[c]urrently, the question of which statistic best suits the identification of keywords is an open one". Each seems to have its advantages and drawbacks in certain situations, and in order to answer my research questions then, I decided to use the SketchEngine platform, which uses the simple maths parameter (Kilgarriff, 2012) as a method to define keywords. The calculation for the simple maths parameter is as follows:

$$simple\ maths\ parameter = \frac{\text{relative frequency of } w \text{ in corpus of interest} + c}{\text{relative frequency of } w \text{ in reference corpus} + c}$$

where *w* is the candidate keyword, and *c* is a constant. As described by Kilgarriff (2012: 9), the constant can be any positive number (but it is usually 1, 100 or 1000), and first serves "as a solution to a range of problems associated with low and zero frequency counts" (Kilgarriff: ibid). Indeed, one of the advantages of the simple maths parameter is that, contrary to other tests, it provides a result which is easy to read and comparable across corpora: the number of times that *w* appears in the corpus of interest compared to the reference corpus. If the result obtained is 2.0 for example, we can say that *w* appears roughly twice as much in the corpus of interest than in the reference corpus. The problem with low and zero frequency words however is that one cannot divide by zero, meaning that zero frequency words in the reference corpus would be automatically discarded, which would not be advisable. With this constant, we are guaranteed to end up with values greater than zero. Another advantage is that, as Brezina (forthcoming) explains, the constant:

> serves as a filter that allows focusing on words above certain relative frequencies in the corpus. For example, if we use 1 as the constant, we highlight low-frequency unique words, while 100 would filter out words that occur with the relative frequency smaller than 100 per million words.

It is up to the researcher then to decide the constant which is the most adapted to their needs, and as Kilgarriff (2012: 9) explains, this "model lets the user specify the keyword list they want by adjusting the parameter. The model provides a way of identifying keywords without unwarranted mathematical sophistication, and reflects the fact that there is no one-size-fits-all list and different lists are wanted for different research questions".

The simple maths parameter then allows for a focus on the actual frequency of the words, and not on the statistical significance, as explained earlier. It also allows for a focus on lockwords[124], as going down the list of keywords to words displaying a score closer to 1 will point to words appearing with similar frequencies in both the corpus of interest and the reference corpus. One potential drawback of this statistical test however is linked with what I have insisted on many times in this thesis now: dispersion. The simple maths parameter does not take dispersion into account, so it may be hypothesized that in some cases, some words could appear as salient keywords because of a handful of users using the word in question many times. This is a valid criticism, and there is no way for me to limit this bias. However, right now there is no existing tool providing ways of doing this, and time constraints prevent me from creating my own tool to extract keywords while taking dispersion into account. Also, even with a method for this, one would need extremely big corpora to be able to have enough occurrences of every word to reliably establish how evenly or unevenly distributed they are. In my case then, the relatively small sizes of the vulgar sub-corpora may not even be enough to have dependable figures. Now that a tool of analysis has been found, i.e. the simple maths parameter through the SketchEngine platform, I will carry on with the keyword analysis itself.

The various sub-corpora have thus been loaded into SketchEngine, so as to be able to extract various lists of keywords according to each gender and age group. I also made a distinction between sub-corpora containing vulgar tweets only from the sub-group in question, and sub-corpora containing all the tweets from this sub-group. In other words, for women aged 12-18, I loaded a corpus containing vulgar tweets only from this sub-group, and another corpus containing all the tweets from this group, vulgar or not. This should provide a clearer view of which keywords appear as salient in a context dominated by swear words among each age group and gender, and being able to compare these results to contexts taking everything into account (and not just swear words) will allow to make a distinction between "regular" and "vulgar" contexts. However, because of the limited sizes of sub-corpora for age groups 31-45 and 46-60, the keyword analysis did not reveal salient patterns, so the focus on these age groups will not be presented here. These age groups were taken into account however when comparing the keywords for males and females as a whole, and not just in vulgar tweets.

[124] Baker (2011: 66) described a lockword as "a word which may change in its meaning or context of usage when we compare a set of diachronic corpora together, yet appears to be relatively static in terms of frequency". The term lockword is now used to talk about words with similar frequencies in synchronic corpora.

Table 8.2 presents the results obtained for the first thirty keywords for all women and men whose age was known[125] in vulgar contexts. Thus, to obtain the keywords for females, the male corpus was used as the reference corpus and vice versa. The minimum frequency was set at 20, meaning that only words appearing at least 20 times were taken into account. This is to ensure that words for which we have little evidence due to a low frequency will not interfere with salient keywords. The constant used to calculate the simple maths parameter (see above) was set at 100, which represents an appropriate in-between in order to highlight high-frequency words without setting aside low-frequency words either. Everything is considered as lowercase, so that *Fuck* is not counted separately from *fuck*, for example. *AF* indicates the absolute frequency, *NF* the normalized frequency per one million words, *rc* refers to the reference corpus, and the score being the result obtained for the simple maths parameter[126].

[125] Taking into account every generation then.

[126] Again, this can be interpreted as roughly being the number of times the word in question appears in the corpus of interest compared to the reference corpus.

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
Females						Males					
loveisland	273	619.8	68	130.7	3.1	eng	386	742	43	97.6	4.3
mood	252	572.1	79	151.9	2.7	euro2016	408	784.3	59	134	3.8
boyfriend	92	208.9	10	19.2	2.6	tournament	188	361.4	10	22.7	3.8
girls	237	538.1	78	149.9	2.6	fans	299	574.7	43	97.6	3.4
mum	282	640.3	102	196.1	2.5	players	181	347.9	17	38.6	3.2
oitnb	113	256.6	23	44.2	2.5	roy	216	415.2	29	65.8	3.1
ur	349	792.4	139	267.2	2.4	team	344	661.2	65	147.6	3.1
omg	212	481.3	74	142.2	2.4	sterling	194	372.9	25	56.8	3
bitch	1093	2481.6	531	1020.7	2.3	goal	200	384.4	27	61.3	3
feel	797	1809.5	386	742	2.3	hodgson	150	288.3	14	31.8	2.9
makeup	70	158.9	8	15.4	2.2	kane	204	392.1	30	68.1	2.9
cute	125	283.8	37	71.1	2.2	england	781	1501.2	202	458.6	2.9
bc	81	183.9	14	26.9	2.2	iceland	194	372.9	30	68.1	2.8
feeling	218	495	87	167.2	2.2	wal	123	236.4	10	22.7	2.7
u	908	2061.5	460	884.2	2.2	hart	165	317.2	23	52.2	2.7
her	663	1505.3	341	655.5	2.1	wales	323	620.9	73	165.7	2.7
xxx	79	179.4	17	32.7	2.1	vardy	152	292.2	23	52.2	2.6
crying	130	295.2	50	96.1	2	play	306	588.2	74	168	2.6
terry	82	186.2	22	42.3	2	cunts	569	1093.7	165	374.6	2.5
am	929	2109.2	522	1003.4	2	league	108	207.6	10	22.7	2.5
bitchy	77	174.8	20	38.4	2	ronaldo	179	344.1	34	77.2	2.5
she	859	1950.3	486	934.2	2	game	576	1107.2	176	399.6	2.4
cry	102	231.6	35	67.3	2	russia	103	198	11	25	2.4
myself	360	817.4	189	363.3	2	france	192	369.1	43	97.6	2.4
n	463	1051.2	253	486.3	2	player	120	230.7	18	40.9	2.3
friends	182	413.2	84	161.5	2	welsh	126	242.2	21	47.7	2.3
omfg	145	329.2	62	119.2	2	club	125	240.3	22	49.9	2.3
my	5387	12230.8	3228	6204.8	2	portugal	162	311.4	36	81.7	2.3
girl	214	485.9	104	199.9	2	manager	89	171.1	9	20.4	2.3
angry	98	222.5	34	65.4	2	season	240	461.3	66	149.8	2.2

Table 8.2: Comparison of gendered keywords for all users in vulgar tweets

Topical differences

The most striking aspect of Table 8.2 is probably the topical differences there are between women and men when they swear. In this regard, men are particularly homogeneous, as every single one of the first thirty keywords is either directly related to sports or used in this context (as revealed by the concordance lines), and more particularly to football or people from the football sphere, whether they are players, club managers or national and local teams. Concerning women, the keywords seem to be more heterogeneous, and refer to television shows (i.e. *loveisland*, *oitnb*[127], *terry*[128]), other people, and women in particular (*girls*, *mum*, *ur*, *u*, *her*, *she*, *friends*, *girl*), themselves (*am*, *myself*, *my*), and emotional states (*mood*, *feel*, *feeling*, *crying*, *cry*, *angry*).

Thus, it would seem that men are more prompted to tweet about the Euro football competition than women, at least when users swear, as we are dealing with vulgar tweets only here. The fact that the competition started the day after the collection phase started (the competition started on the 10th of June then) may play a role in the overrepresentation of this topic among (male) keywords, as the collection probably started precisely when the expectation phenomenon was at its peak, and when people were eager for the event to begin. The event ended on the 10th of July, so a month before the end of the collection phase. It could be argued that the "topic" of football is unevenly represented, which may be seen as a bias. This objection, and the question of the representativeness of certain themes on social media is a recurring one. On the other hand, it could also be argued that a medium like Twitter being oriented towards immediate events and reactions for reasons detailed earlier, homogeneity of topics will never be fully reached, as users continuously react to more or less local or global events (Guille and Favre, 2015). Whatever the position we adopt, the main fact to focus on here is that women had as many possibilities as men to react to the Euro competition when swearing, which they did not, so the difference here is still relevant. However it should not be concluded that football as a whole is a "male topic", as so far we do not have enough detailed evidence to assert that. It could very well be that women do not tweet about football, or tweet about football without swearing (remember that right now we are only focusing on vulgar tweets). What these figures allow us to know here is simply that men mention the topic significantly more often than women when swearing on Twitter.

[127] Which is the abbreviation for the TV series "Orange Is the New Black".
[128] In this case "Terry" mainly refers to a contestant of the 2016 season of Love Island.

Swear words

The swear words present in the top keywords are *bitch*, *bitchy*, and the abbreviation *omfg* (i.e. "oh my fucking God") for women, and *cunts* for men. It is not surprising to find these swear words among the top keywords, as these were part of the most salient gendered swear words according to the Mann-Whitney U tests carried out in Chapter 7. What may be more surprising however, is to see that for men the variant of *cunt* which is considered as a keyword is the plural one, indicating that men use the word to refer to groups of people more often than women, and in particular they use it more often than they use the singular version compared to women, otherwise *cunt* (singular) would also be present as a top keyword. *Fuck*, which was the second swear word found to be the most strongly associated with men with the MWU tests only appears as the 371st most relevant keyword under the form *fuckers* (with a score of 1.3). This is interesting from two aspects. First, this means that although shown to be statistically more used by men thanks to the MWU tests, it is not as salient in the keyword analysis, implying that the quantitative difference of use between women and men may remain marginal when put in relation with the rest of the linguistic repertoire of Twitter users. In other words, the statistical difference of use of swear words taken separately from any other words, as with the MWU tests, may prove significant, but there may actually be many other words which could prove to be more significantly used by one gender or the other. Secondly, the fact that *fuckers* appears as the first variant of *fuck* to be considered a keyword in the vulgar corpus indicates that what may cause *fuck* to be considered as being overused by men according to the MWU tests is the way men use it to talk about, or talk to, groups of people. McEnery and Xiao (2003: 504) showed that women and men differed in their preferences of certain forms of *fuck* already, so it would not be surprising to observe this pattern here as well. However, the fact that both the plural forms of *cunt* and *fucker* appear as being favored by men compared to women indicates that the main difference of use between women and men for these two words may be in the way they address these swear words to other people, or to who they address them, as can be seen in the examples below:

> (#014) Not sure they could've picked 2 bigger cunts than Farage and Chris Grayling to be on the question time panel

> (#015) the majority of the idiots who voted leave are literally just racist fuckers who don't want people coming to take their imaginary jobs

More contextual details would be needed to know more about this however, and the collocation analyses I will later get into will surely help us understand this phenomenon.

Abbreviations

Still concerning the choice of forms of *fuck*, it is also interesting to see that the abbreviation *omfg* appears as a keyword, as this correlates with our earlier observation that abbreviated forms of *fuck* were favored by women. Beyond that, it may seem that abbreviations as a whole could be favored by women, and not just those related to swear words. Indeed, among the thirty most salient keywords, eight are abbreviations, three of which being abbreviated forms of very frequent words (*ur* for *your*, *u* for *you*, and *n* for *and*). Thus, women's preference for abbreviations does not seem to be limited to swear words only, and appears to include a wide range of word classes.

The fact that we are focusing here on all users probably implies that at least some of these trends will be present when we focus on specific age groups as well, but doing so may also reveal new tendencies, or may give additional information allowing to better understand these differences. Table 8.3 presents the results obtained for users aged 12-18. The minimum frequency has been set to 10, everything considered as lowercase, and the constant to 100:

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
Females						Males					
xx	59	353	6	44.4	3.1	eng	99	731.8	12	71.8	4.8
girls	112	670.1	22	162.6	2.9	euro2016	83	613.5	10	59.8	4.5
oitnb	46	275.2	4	29.6	2.9	kane	58	428.7	4	23.9	4.3
mistress*[129]	31	185.5	0	0	2.9	wales	98	724.4	17	101.7	4.1
darling*	31	185.5	0	0	2.9	white	81	598.8	13	77.8	3.9
mood	136	813.6	30	221.8	2.8	roy	70	517.4	11	65.8	3.7
ur	225	1346.1	57	421.3	2.8	goal	59	436.1	8	47.9	3.6
dnt	29	173.5	0	0	2.7	sterling	61	450.9	9	53.8	3.6
mum	141	843.6	35	258.7	2.6	fans	75	554.4	14	83.8	3.6
makeup	40	239.3	4	29.6	2.6	nigger	37	273.5	1	6	3.5
loveisland	68	406.8	13	96.1	2.6	tournament	35	258.7	1	6	3.4
x	170	1017.1	47	347.4	2.5	hart	46	340	7	41.9	3.1
n	320	1914.5	96	709.6	2.5	sexual	33	243.9	2	12	3.1
feel	374	2237.5	115	850.1	2.5	season	70	517.4	17	101.7	3.1
bitch	465	2781.9	149	1101.4	2.4	vardy	45	332.6	7	41.9	3
boyfriend	42	251.3	7	51.7	2.3	england	205	1515.4	72	430.8	3
babe	36	215.4	5	37	2.3	players	44	325.3	7	41.9	3
xxx	33	197.4	4	29.6	2.3	faggot	29	214.4	1	6	3
brother	50	299.1	10	73.9	2.3	beat	41	303.1	6	35.9	3
little	155	927.3	48	354.8	2.3	play	79	584	22	131.6	3
clothes	29	173.5	3	22.2	2.2	bale	31	229.2	2	12	2.9
im	218	1304.2	72	532.2	2.2	team	92	680.1	28	167.5	2.9
dad	94	562.4	27	199.6	2.2	wal	28	207	1	6	2.9
omg	96	574.3	28	207	2.2	iceland	44	325.3	8	47.9	2.9
u	544	3254.6	195	1441.5	2.2	hodgson	37	273.5	6	35.9	2.7
drunk	59	353	15	110.9	2.1	win	65	480.5	19	113.7	2.7
crying	61	364.9	16	118.3	2.1	game	148	1094	59	353	2.6
face	84	502.5	25	184.8	2.1	afro	23	170	1	6	2.5
boy	68	406.8	19	140.4	2.1	bastards	62	458.3	20	119.7	2.5
fckin	34	203.4	6	44.4	2.1	portugal	39	288.3	9	53.8	2.5

Table 8.3: Comparison of gendered keywords for the 12-18 in vulgar tweets

[129] These words are considered keywords because of spam accounts, as revealed by checking the concordance lines.

Overall, Table 8.3 shows the same gendered tendencies as Table 8.2. Most of the keywords for males revolve around football, and most keywords for females are related to other people, TV shows and emotions. We also note that there is one more abbreviation among the top thirty keywords here for females than when taken as a whole, which may suggest that using abbreviations could be more popular among younger females. The gender specificity of abbreviations seems to be confirmed, as males of the same age still do not use any of them significantly enough for them to appear as top keywords[130]. Additionally, we notice that *omfg* as an abbreviated form of *fuck* among women as a whole is not present among the 12-18-year-olds anymore, but instead *fckin* is present as a salient keyword. On top of strengthening the idea that women prefer the abbreviated forms of *fuck* more than men, this also appears as a confirmation of what McEnery and Xiao (2003: 504) observed about gendered preferences concerning the variants of *fuck*. Indeed, they found that males as a whole favored the root form, *fuck*, whereas women preferred the *–ing* form, *fucking*. Although the results were not statistically significant in their study, the fact that we notice the same pattern here seems to indicate that this tendency really exists. Still concerning female patterns, we noticed that when taken as a whole they seemed to mention other women more than men. Among the 12-18-year olds, this pattern is still noticeable (*girls*, *mum*[131]), but additionally, people from the opposite sex seem to have a better representation here (*boyfriend*, *brother*, *dad*, *boy*), to the point where females from this age group may seem to mention males more than other females. *Bitch* is still present as a salient keyword for females, as well as the TV shows mentioned previously.

For males, again, the major tendency is the same: the main topic here seems to be football. Again, this should not be interpreted as implying that men only tweet about football, but simply that in these cases, they do it significantly more than women for these words to appear in the top 30 keywords. Concerning the swear words appearing as keywords, *cunt* or its variants is not present anymore, but it is replaced by *nigger*, *faggot* and *bastards*. Again, the fact that *bastard* is used in its plural form indicates that males may direct swear words to groups of people more frequently than women. This is strengthened by the presence of *fans* and *players*

[130] Note that *eng* and *wal*, which are abbreviations of England and Wales, are in this case present as part of the hashags *#eng* and *#wal*, and are thus not counted as abbreviations per se, because they are an imposed means/format of communicating on Twitter instead of a deliberate choice of the users.

[131] Note that I did not take *mistress* into account, as in this case most of the occurrences of this word (as well as *darling*) come from the same user, which/who appears to either be a bot, or a spam account (probably both actually) broadcasting sexual content, so irrelevant in the case of the current discussion.

(both plural), but on the other hand, the presence of many singular nouns referring to single people (*kay*, *roy*, *stirling* etc…) indicates that they probably swear at individuals relatively frequently too. Indeed, this is confirmed by the very presence of *nigger* and *faggot* which are in the singular form. Males thus use swear words to address groups of people, as well as individuals.

Table 8.4 will now give the results obtained for the 19-30-year olds. The minimum frequency has been set to 10, everything considered as lowercase, and the constant to 100:

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
Females						Males					
loveisland	195	773.3	53	164.8	3.3	eng	207	643.7	24	95.2	3.8
boyfriend	50	198.3	3	9.3	2.7	euro2016	257	799.2	35	138.8	3.8
mood	115	456	41	127.5	2.4	players	112	348.3	5	19.8	3.7
cute	77	305.4	22	68.4	2.4	tournament	127	394.9	9	35.7	3.6
bc	44	174.5	5	15.5	2.4	fans	181	562.9	27	107.1	3.2
oitnb	64	253.8	16	49.8	2.4	roy	131	407.4	15	59.5	3.2
her	382	1514.9	198	615.7	2.3	team	205	637.5	34	134.8	3.1
omg	111	440.2	45	139.9	2.3	sterling	119	370.1	14	55.5	3
girls	118	467.9	49	152.4	2.3	goal	123	382.5	16	63.5	3
feeling	125	495.7	53	164.8	2.2	iceland	134	416.7	19	75.3	2.9
bitch	603	2391.3	329	1023.1	2.2	hodgson	92	286.1	8	31.7	2.9
am	502	1990.7	279	867.6	2.2	england	491	1526.9	119	471.9	2.8
mum	134	531.4	62	192.8	2.2	hart	99	307.9	12	47.6	2.8
myself	209	828.8	108	335.9	2.1	cunts	339	1054.2	83	329.1	2.7
bitchy	43	170.5	9	28	2.1	ronaldo	126	391.8	21	83.3	2.7
cry	62	245.9	21	65.3	2.1	kane	132	410.5	23	91.2	2.7
she	503	1994.7	300	932.9	2	league	72	223.9	6	23.8	2.6
wine	35	138.8	6	18.7	2	russia	65	202.1	5	19.8	2.5
rude	44	174.5	12	37.3	2	wal	74	230.1	8	31.7	2.5
shitty	207	820.9	117	363.8	2	vardy	91	283	14	55.5	2.5
excited	76	301.4	34	105.7	2	play	193	600.2	48	190.4	2.4
feel	401	1590.2	247	768.1	1.9	player	76	236.3	10	39.7	2.4
terry	46	182.4	15	46.6	1.9	wales	195	606.4	49	194.3	2.4
jason	52	206.2	19	59.1	1.9	pogba	48	149.3	1	4	2.4
angry	56	222.1	22	68.4	1.9	manager	57	177.3	4	15.9	2.4
makeup	29	115	4	12.4	1.9	game	372	1156.8	109	432.3	2.4
bigbrother	29	115	4	12.4	1.9	joe	97	301.7	18	71.4	2.3
cos	105	416.4	56	174.1	1.9	against	70	217.7	9	35.7	2.3
friends	96	380.7	50	155.5	1.9	france	120	373.2	28	111	2.2
girl	115	456	63	195.9	1.9	rooney	51	158.6	4	15.9	2.2

Table 8.4: Comparison of gendered keywords for the 19-30 in vulgar tweets

Here again, the main tendencies are the same, and apart from the presence of *shitty* as a keyword for women, it can actually be difficult to spot the differences with Table 8.3. This lack of

difference between the most salient keywords for both genders and for the two younger generations, as well as for genders taken as a whole, implies that the main differences between women and men are relatively steady across age groups. However, it must be acknowledged that the lack of data for older generations prevents an accurate evaluation of this "steadiness" among these age groups. As we have seen in Chapter 6, more varied gendered differences may be observed for keywords if we were able to compare the two younger generations to the two older ones. However, even without studying older generations, a lack of difference is still relevant in itself, and shows some key aspects for which women and men diverge.

As a reminder, the results observed so far only present keywords for the sub-corpora only composed of tweets containing at least one swear word. We could wonder then, if we would have observed different tendencies if, instead of focusing on vulgar tweets only, we focused on *every* tweet, vulgar or non-vulgar. Indeed, would we still observe keywords related to football only among men if we did not take into account vulgar tweets only? If not, would it mean that men only mention about football when they swear, or that women mention football as much as men when they do not swear? In order to try to answer these questions, and to analyze the corpora under different angles, Table 8.5 presents the top thirty keywords for women and men as a whole, this time using the corpora composed of every tweet, vulgar or not. The procedure is the same, to calculate female keywords, the male corpus was used as the reference corpus and vice versa. The minimum frequency was set at 20, everything considered as lowercase, and the constant set to 100:

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
Females						Males					
xxx	2766	464.1	600	91.5	2.9	euro2016	5171	788.6	937	157.2	3.5
loveisland	2358	395.7	563	85.9	2.7	eng	3098	472.4	498	83.6	3.1
excited	4080	684.6	1357	206.9	2.6	players	2001	305.2	231	38.8	2.9
omg	3396	569.9	1139	173.7	2.4	mate	6009	916.4	1506	252.7	2.9
cute	2277	382.1	647	98.7	2.4	player	2007	306.1	251	42.1	2.9
hair	2461	413	771	117.6	2.4	tournament	1561	238.1	148	24.8	2.7
makeup	983	164.9	90	13.7	2.3	league	1542	235.2	150	25.2	2.7
mum	3148	528.2	1140	173.8	2.3	mm*	1131	172.5	12	2	2.7
vinteduk*[132]	753	126.4	0	0	2.3	game	6580	1003.4	1868	313.5	2.7
xx	4242	711.8	1705	260	2.3	barometer*	1092	166.5	0	0	2.7
girls	2430	407.8	840	128.1	2.2	team	3948	602.1	986	165.5	2.6
fab	1419	238.1	378	57.6	2.1	mph*	1102	168.1	11	1.8	2.6
size	1122	188.3	235	35.8	2.1	humidity*	1103	168.2	24	4	2.6
literally	4031	676.4	1795	273.7	2.1	mb*	1097	167.3	28	4.7	2.6
my	68069	11422.1	35977	5486.5	2.1	goal	2311	352.4	490	82.2	2.5
u	8631	1448.3	4292	654.5	2.1	temperature*	1131	172.5	62	10.4	2.5
boyfriend	938	157.4	169	25.8	2	england	6567	1001.5	2067	346.8	2.5
love	18142	3044.3	9582	1461.3	2	fans	2529	385.7	605	101.5	2.4
xxxx	897	150.5	166	25.3	2	wind*	1242	189.4	133	22.3	2.4
girl	2652	445	1140	173.8	2	iceland	1673	255.1	312	52.4	2.3
crying	1445	242.5	477	72.7	2	wales	3365	513.2	1006	168.8	2.3
miss	3651	612.6	1703	259.7	2	play	3717	566.8	1242	208.4	2.2
oitnb	925	155.2	189	28.8	2	wal	1442	219.9	311	52.2	2.1
sleep	4425	742.5	2134	325.4	2	score	1213	185	214	35.9	2.1
holiday	3180	533.6	1446	220.5	2	ronaldo	1432	218.4	316	53	2.1
ur	2326	390.3	972	148.2	2	payet	865	131.9	69	11.6	2.1
grab	832	139.6	141	21.5	2	france	1972	300.7	557	93.5	2.1
bed	3759	630.8	1777	271	2	games	1482	226	350	58.7	2.1
mistress*	575	96.5	8	1.2	1.9	pogba	755	115.1	34	5.7	2
wanna	3151	528.7	1487	226.8	1.9	cheers	1689	257.6	460	77.2	2

Table 8.5: Comparison of gendered keywords for all users in all tweets

[132] These words are considered as keywords because of spam accounts, as revealed by checking the concordance lines.

Here again, the results are very similar to those obtained when focusing on vulgar tweets only. One of the most notable differences is the presence of words related to meteorology among salient male keywords (*barometer*, *mph*, *humidity*, *temperature*), which come from a bot account regularly tweeting information regarding weather. The same observations can be made when comparing gendered keywords for each age group in all tweets (and not just vulgar ones), (see Appendices 1 to 4[133]). It can also be observed that the keywords for the two older generations are also influenced by bots and automated messages in these results as well. It was hoped that the greater number of tweets in corpora composed of all the tweets would allow for more coherent results among these generations, but here again we see patterns which are either similar to those observed among younger generations (football, etc…), or the even greater influence of automated content like weather broadcasts, or tweets automatically emitted through online games, as examples #016 and #017 show:

> (#016) Wind 6.9 mph E. Barometer 1021.0 mb, Falling. Temperature 17.6 °C. Rain today 0.0 mm . Humidity 72%

> (#017) I just completed this puzzle in Jigsaw Puzzles Epic! *URL* # jigsawepic *URL*

From there it can be concluded that the main gendered differences concerning users' linguistic preferences are not affected by the presence of swear words. This is important, as this may imply that Twitter users do not swear in certain contexts only. They may, on the other hand, *not* swear in certain contexts, but the recurring patterns of keywords in vulgar and non-vulgar corpora indicate that women and men will swear in the contexts they usually mention. This is an indication that swear words are not bound to be used in restricted contexts only, and it may partly explain the growing tolerance regarding swear words mentioned earlier. The similarity between the results from the vulgar and non-vulgar corpora also indicates that the tendencies observed earlier, about the preference of abbreviations from women over men for example, are not due to the presence or the absence of swear words, and thus that these tendencies are purely gendered ones.

[133] Because the results are very similar to those presented earlier, it was decided to not add those here and put them in the appendixes as a way not to overwhelm the reader with tables and information which are not so different from what has been shown so far.

3.8.3 Keyword analysis: intra-gender variations

So far then, we have been focusing on various levels of inter-gender differences, and the main observation we made is that the differences observed are steady whatever the age group or the context we take into account (i.e. vulgar or not-vulgar-only). What could reveal more subtle differences now is to study intra-gender variation, by looking at the differences there are among men and also among women from various age groups in vulgar contexts. From what we have seen in our previous analyses, we can logically anticipate that the semantic field of football will disappear. Being steady across all sub-groups of males in our comparison with females, it is then fairly obvious that when comparing the differences there are between males of different age groups, these words will not be salient. Beyond mere gendered differences then, what is at stake here is a better understanding of the potential generational differences there may exist inside each age group, so as to better understand the different contextual preferences there may exist across generations when swearing. Table 8.6 then presents the results obtained when comparing women aged 12-18 to women aged 19-30 in vulgar tweets only. The minimum frequency being 10, everything being considered as lowercase, and the constant being 100:

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
12-18 females						19-30 females					
n	320	1914.5	136	539.3	3.2	bbuk	218	864.5	40	239.3	2.8
prom	38	227.3	3	11.9	2.9	gameofthrones	49	194.3	2	12	2.6
mistress*[134]	31	185.5	0	0	2.9	cbb	46	182.4	5	29.9	2.2
exam	73	436.7	24	95.2	2.8	e32016	29	115	0	0	2.2
rt	33	197.4	3	11.9	2.7	hughie	36	142.8	3	17.9	2.1
exams	44	263.2	11	43.6	2.5	jayne	28	111	2	12	1.9
ur	225	1346.1	119	471.9	2.5	flat	32	126.9	4	23.9	1.8
dnt	29	173.5	3	11.9	2.4	those	90	356.9	26	155.5	1.8
u	544	3254.6	322	1276.9	2.4	brexit	43	170.5	9	53.8	1.8
darling*	31	185.5	5	19.8	2.4	bigbrother	29	115	4	23.9	1.7
college	49	293.2	18	71.4	2.3	student	21	83.3	1	6	1.7
rn	50	299.1	21	83.3	2.2	loveisland	195	773.3	68	406.8	1.7
boys	82	490.6	48	190.4	2	jason	52	206.2	13	77.8	1.7
fckin	34	203.4	13	51.6	2	euref	49	194.3	12	71.8	1.7
school	57	341	31	122.9	2	change	49	194.3	12	71.8	1.7
x	170	1017.1	118	467.9	2	marco	23	91.2	2	12	1.7
cba	46	275.2	23	91.2	2	bale	23	91.2	2	12	1.7
ngl	15	89.7	0	0	1.9	georgina	28	111	4	23.9	1.7
pics	29	173.5	12	47.6	1.9	london	66	261.7	19	113.7	1.7
pet	25	149.6	9	35.7	1.8	hollyoaks	17	67.4	0	0	1.7
lol	176	1053	134	531.4	1.8	bbgeorgina	17	67.4	0	0	1.7
soz	18	107.7	4	15.9	1.8	also	101	400.5	34	203.4	1.6
puppy*	18	107.7	4	15.9	1.8	real	91	360.9	30	179.5	1.6
b	37	221.4	20	79.3	1.8	match	26	103.1	4	23.9	1.6
xxxx	20	119.7	6	23.8	1.8	towie	23	91.2	3	17.9	1.6
boy	68	406.8	47	186.4	1.8	vote	101	400.5	35	209.4	1.6
babe	36	215.4	20	79.3	1.8	wynonnaearp	15	59.5	0	0	1.6
dad	94	562.4	70	277.6	1.8	racist	41	162.6	11	65.8	1.6
im	218	1304.2	177	701.9	1.8	natalie	41	162.6	11	65.8	1.6
shittest	52	311.1	34	134.8	1.8	eastenders	29	115	6	35.9	1.6

Table 8.6: Comparison of keywords for 12-18 and 19-30 females in vulgar tweets

[134] These words are considered as keywords because of spam accounts, as revealed by checking the concordance lines.

Now that we are comparing two generations of women, some patterns observed before appear here as well, but they are less marked as with men, and new tendencies appear. As expected, the smoothing of patterns which were common to both generations of women limited the recurrence of the same keywords, and leaves room for new ones which are specific to age groups only.

TV shows

Here again, TV shows are present, but only among the 19-30-year-olds. This is interesting, because when we compared females and males aged 12-18 we observed that *loveisland* was a keyword for females, as well as *oitnb*. Of these two shows, only Love Island remains, and it is only present as a keyword for women aged 19-30. This is the sign that although it was mentioned significantly more by females aged 12-18 than by males (which explains its earlier presence as a keyword in this case), it is even more used by women aged 19-30. The TV show *Orange Is the New Black* on the other hand seems to be mentioned at comparable frequencies, thus explaining why it is not identified as a salient keyword[135]. Also, it is interesting to note that TV shows do not appear at all in the salient keywords for females aged 12-18 anymore, but are relatively frequent for those aged 19-30. Among the top 30 keywords of this group, 15 of them are related to TV shows, either by referring to the shows themselves (*bbuk*, *cbb*[136], *gameofthrones*, *bigbrother*, *loveisland*, *hollyoaks*, *towie*, *wynonnaearp*), or by referring to characters, or contestants present in the shows (*hughie*, *jayne*, *jason*, *marco*, *bale*, *georgina*, *bbgeorgina*, *natalie*). What can be interpreted from these observations is that females as a whole seem to mention TV shows more often than males when swearing, and that among females themselves, certain sub-groups (namely, the 19-30-year-olds) mention these shows significantly more than others in vulgar contexts. Also, the TV show Big Brother seems to be particularly salient for women aged 19-30, who already mentioned it significantly more than men of the same age in Table 8.4. Here, it appears a lot more for these women again compared to females aged 12-18[137], indicating that the show may be more successful for people from this age group.

[135] Indeed, *oitnb* appears in the keyword list of females aged 12-18 but only appears in 607th position, with a normalized frequency of 275.2, and 253.8 for women aged 19-30, with a score of 1.1. It can then be considered as a lockword, and not as a keyword anymore.

[136] *bbuk* being the abbreviation of Big Brother UK, and *cbb* being the abbreviation of Celebrity Big Brother.

[137] Most of the first names present are names of contestants present in this show.

Politics

This is a domain which did not appear among salient keywords so far. In this case, it seems to be present among women aged 19-30 only (*brexit*, *euref*[138], *vote*), at least as far as the most salient keywords are concerned. In the case of the keyword *vote*, it could be hypothesized that it may not be used in a political context, as it may very well be mainly used in the context of TV shows, when talking about voting for certain contestants for example. A manual reading of the concordance lines revealed that the vast majority of the time people actually talked about voting in or out of the European Union, as the other related keywords *brexit* and *euref* suggest. The examples below are typical of those "political" tweets:

> (#018) I really don't understand this referendum shite. I don't know whether to even vote *or not*
>
> (#019) fuck that 16 and 17 year olds aren't even allowed to vote but all the racists with an intellectual capacity of a banana get their say

It does not seem surprising to find more political references among women aged 19-30, because (as suggested in example #019 above), users from the age group 12-18 are not allowed to vote, so intuitively it seems reasonable to anticipate that they may talk about that less than older generations.

Abbreviations and swear words

Here again, abbreviations (which are not already part of a hashtag) are fairly frequent among the top 30 keywords, but seem to be more popular among females aged 12-18 (*n*, *ur*, *dnt*, *u*, *cba*, *fckin*, *ngl*, *pics*, *soz*, *lol*, *b*, im[139]) than among the 19-30-year-olds (none). When compared to men aged 19-30, three abbreviations were found among the top keywords for women of the same age (*bc*, *omg*, *cos*), so although the difference is not as salient as in other cases, as we have just seen with TV shows for example, this could indicate that women as a whole use more abbreviations than men, but that 12-18-year-old females are the ones using abbreviations the most between these two generations. About swear words and abbreviations, we notice the presence of the abbreviated form *fckin* as a keyword for the 12-18-year-olds, reinforcing the idea that abbreviations, and abbreviations of the swear word *fuck* in particular, are preferred by

[138] Abbreviation of European referendum.

[139] Abbreviations of: *and*, *your* (or *you're*), *don't*, *you*, *can't be arsed*, *not gonna lie*, *pictures*, *sorry*, *laughing out loud*, *be* (and/or *bee*?), *I am*.

females. The fact that two swear words appear as keywords for females aged 12-18 (*fckin* and *shittest*) strengthens what was observed in Table 7.3 already (see Chapter 7), which was that swear words are used more often by females aged 12-18 than by those aged 19-30. This difference is made apparent here with the absence of swear words in the top 30 keywords for women aged 19-30, and the presence of the two afore-mentioned ones among the 12-18-year-olds. The presence of *shit* as a top keyword is both surprising, and expected. It could have been expected, because it has been shown to be the second most frequently used swear word after *fuck* for both genders. However, its presence as a keyword indicates that it is used significantly more by those aged 12-18 than by the 19-30-year-olds. As a word being used a lot by everyone, we may have expected a relatively homogeneous distribution across age groups. But here, the form which appears is the superlative *shittest*, which may indicate that users from this age group will use it more often than the 19-30 to complain about or denigrate something, as in the examples below:

> (#020) The fact that Wales has lost to England has put me in the shittest mood going
> (#021) Don't miss prom one bit shittest night of my life and hated nearly every fucker there

Thus, what these results show primarily is that even if one group displays a tendency, either when compared to another group (as in the case of women and men) or when taken as a single entity (as when we looked at the overall use of swear words by women in Table 7.3), we should always look deeper into the corpus and look at patterns inside the very groups under study. This can be done either by looking at the dispersion, or as I just did, by looking at the same dataset from a different angle so that other aspects of the corpus can be made apparent. Doing this will prevent conclusions on the whole group to be drawn, whereas the tendency could be particularly salient in one sub-group, and not so much on the other. This is related to what we have seen several times now about aggregate data, and the need to go beyond that as much as possible. Although here I did not specifically focus on individual uses of the words, the need to take context into account is once more made obvious, and this shows that although major tendencies are important, looking at things in context allows to make the most out of the dataset. In this case, we realized that although when compared to men, women were shown to mention TV shows much more, in fact women aged 19-30 are the ones referring to these shows even more than those aged 12-18 when swearing. The same is true with swear words, and although *fuck* and *shit* have been shown to be significantly more frequently used by men according to the

MWU tests, here we realized that 12-18-year-old females use some variants of these swear words significantly more than the 19-30-year-olds. This once more points to the fact that generational differences may be more relevant than gendered ones and that the tendencies observed when looking at aggregate data may not be uniform at all when we go past that level. Additionally, looking at inter-generational variation among women in all tweets (not just vulgar ones) reveals the same patterns as the ones observed in vulgar tweets only, which once again demonstrates that swearing is not restricted to a pre-defined set of contexts.

The same analysis will be presented for men now, and Table 8.7 presents the top 30 keywords for males aged 12-18 and 19-30 in vulgar tweets. The minimum frequency being 10, everything being considered as lowercase, and the constant being set at 100:

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
12-18 males						19-30 males					
white	81	598.8	35	108.8	3.3	cbb	71	220.8	2	14.8	2.8
sexual*[140]	33	243.9	1	3.1	3.3	bbuk	138	429.2	17	125.7	2.3
nigger	37	273.5	4	12.4	3.3	corbyn	52	161.7	4	29.6	2
exam	36	266.1	11	34.2	2.7	also	112	348.3	18	133.1	1.9
afro	23	170	1	3.1	2.6	final	75	233.2	10	73.9	1.9
sex	28	207	10	31.1	2.3	ufc200	32	99.5	1	7.4	1.9
fuq	17	125.7	0	0	2.3	labour	46	143.1	5	37	1.8
y	34	251.3	18	56	2.3	fair	94	292.3	17	125.7	1.7
faggot	29	214.4	13	40.4	2.2	remain	68	211.5	11	81.3	1.7
retardation*	16	118.3	0	0	2.2	trident	22	68.4	0	0	1.7
faggotry*	16	118.3	0	0	2.2	though	184	572.2	41	303.1	1.7
nigguh	15	110.9	0	0	2.1	racist	41	127.5	5	37	1.7
aint	28	207	16	49.8	2	crap	242	752.6	56	414	1.7
u	195	1441.5	217	674.8	2	sounds	48	149.3	7	51.7	1.6
usa	16	118.3	4	12.4	1.9	nigel	32	99.5	3	22.2	1.6
swear	39	288.3	33	102.6	1.9	clubs	24	74.6	1	7.4	1.6
mint	18	133.1	7	21.8	1.9	times	89	276.8	18	133.1	1.6
span*	12	88.7	0	0	1.9	por	35	108.8	4	29.6	1.6
dopaminal*	12	88.7	0	0	1.9	bit	133	413.6	30	221.8	1.6
school	36	266.1	31	96.4	1.9	bbbots	19	59.1	0	0	1.6
exams	12	88.7	1	3.1	1.8	germany	53	164.8	9	66.5	1.6
energy*	15	110.9	5	15.5	1.8	tonight	234	727.7	57	421.3	1.6
ma	44	325.3	43	133.7	1.8	stuff	64	199	12	88.7	1.6
brain	24	177.4	17	52.9	1.8	during	22	68.4	1	7.4	1.6
gon	14	103.5	4	12.4	1.8	pogba	48	149.3	8	59.1	1.6
n	96	709.6	112	348.3	1.8	scraptrident	18	56	0	0	1.6
rn	21	155.2	14	43.5	1.8	nigga	58	180.4	11	81.3	1.5
maths	10	73.9	0	0	1.7	vote	137	426	33	243.9	1.5
wtf	137	1012.7	177	550.4	1.7	farage	57	177.3	11	81.3	1.5
m8	17	125.7	11	34.2	1.7	crowd	17	52.9	0	0	1.5

Table 8.7: Comparison of keywords for 12-18 and 19-30 males in vulgar tweets

* These words are considered as keywords because of spam accounts, as revealed by checking the concordance lines.

In many aspects, the results obtained for the inter-generational comparison of male keywords are very similar to the results obtained for females. When comparing women's patterns to men's earlier, we noticed that abbreviations were salient among women, and that very few (even none in certain cases) appeared as keywords among males. Here, we realize that abbreviations are very present among males aged 12-18 (*fuq*, *y*, *aint*, *u*, *ma*, *gon*, *n*, *rn*, *wtf*, *m8*[141]) compared to those aged 19-30[142] (*nigga*). The fact that many abbreviations are now highlighted among 12-18-year-old males is an indication that abbreviations are overall significantly more present among women aged 12-18 than men of the same age, explaining why they could not be salient among men at this point. However, they are now striking among the 12-18-year-olds because men aged 19-30 do not use abbreviations as much. Again, this does not mean that 19-30-year-old men do not use abbreviations, as the detailed frequencies in Table 8.7 show that they do, but they simply use them significantly less frequently, hence they appear among those aged 12-18 only. Some of these abbreviations are variants of *fuck* (i.e. *fuq* and *wtf*), indicating that although females of the same age use these more than males aged 12-18 as we have seen earlier, younger generations of both women and men are the most likely to use abbreviations of the word *fuck*, as well as abbreviations as a whole. Although the presence of the base spelling *nigger* as a keyword for the 12-18-year-olds seems to indicate that this age group favors this word over the 19-30-year-olds, the fact that we find the spellings *nigguh* for males aged 12-18 and *nigga* for the 19-30-year-old males indicates that both generations use the word, but they have their own preferences regarding its spelling. In Table 8.1, we observed that overall, there was no gendered preference for the swear word *nigger* or its variants. Thus, we still cannot talk of it as being preferred by males, but here we once more notice that age plays a greater role than gender in determining how the word will be used.

This topic of politics is here again one which seems to be significantly more present among men aged 19-30 (*corbyn*, *labour*, *remain*, *trident*, *nigel*, *scraptrident*, *vote*, *farage*) than among the 12-18-year-olds (none). The same explanations given in the case of women can apply here as well, and the fact that among users aged 12-18, only those who are 18 are allowed to vote probably partially influences the presence of these words as keywords. Once more then, we realize that age is more relevant than gender in determining patterns of language use. In this

[141] These are abbreviations of *fuck*, *you*, *are not*, *my*, *going to*, *and*, *right now*, *what the fuck*, *mate*.
[142] Here again, abbreviations which are part of hashtags are not considered as such.

regard, we notice that the topic of school in a recurring one for 12-18-year-old males and females compared to those aged 19-30, as for both genders, *school* and *exams* are considered as keywords compared to the 19-30-year-olds. This seems logical[143], and this is likely to be related to what was said earlier about the fact that people tend to be influenced by the community they spend a lot of time with. However, beyond the mere influence of peers, in this case the fact that users from this age group are probably a majority to go to secondary school necessarily plays a role in the kind of words they will use, and as the 19-30-year-olds are unlikely to still be going to school, they will probably not mention it as often as users for whom this is the main occupation.

Still about generational tendencies, we observe that the TV show *Big Brother* is particularly salient among men aged 19-30 too (*cbb*, *bbuk*, *bbbots*[144]). When comparing women to men earlier, we noticed that the TV show *Big Brother* was salient among women keywords for those aged 19-30. With this additional data, we can assert that *Big Brother* is not particularly salient just among women aged 19-30, it is salient among *all* users aged 19-30. Of course, women from this generation mention it significantly more than men of the same age, otherwise the terms related to the show would not have appeared as keywords for women, but these results show that, once more, absence of evidence is not an evidence of absence, and the absence of keywords related to TV shows for men earlier cannot be said to be a sign that TV shows are a "female thing" based on these figures alone. The results of the inter-generational keyword comparison for males in all tweets (not just vulgar ones) shows the same tendencies, conforming to what has been observed several times by now, which is the fact that women and men mention the same topics when swearing as well as when they do not.

[143] Especially as the collection of tweets occurred at the end of the academic year.
[144] *bbbots* being the abbreviation of Big Brother's Bit On The Side.

Conclusion:

In this chapter, we have focused more closely on patterns highlighted in earlier chapters to try to better understand them and see whether the hypotheses made before hold when looking at the data more thoroughly and under various angles. Although I have been dealing with any kind of keyword and not just swear words only, what these inter-gender, intra-gender, and inter-generational analyses confirmed is that age is probably the most important factor determining the contexts in which users will swear. Indeed, although I have not been focusing on swear words only, remember that most of the results presented in this chapter were based on comparisons of patterns present in vulgar tweets. Thus, I would like to insist on two aspects mentioned several times already, but which are key in analyzing gender in this study, and when dealing with corpora more generally. 1) Interpretations should not be made too early in the analysis of the data, otherwise key aspects which can only be revealed by looking at the data under a different angle may be missed. This is how, in earlier studies on language and gender, women have been said to be deferential and powerless, and how men have been shown to be pragmatic and vulgar[145]. Most of these stereotypes have been contradicted thanks to more thorough investigations, but in order to prevent such erroneous associations with gender or any other social category in the future, we must be careful with the way we analyze the data. 2) One of the most important factors determining how Twitter users express themselves so far has been shown to be age instead of gender, so analyzing "gendered speech patterns" only is not enough to render the full spectrum of linguistic devices people resort to when speaking[146]. Even while taking into account parameters other than gender, this study probably misses other key components which may allow us to go deeper in our understanding of the mechanisms at play when swearing, and looking at race, occupation, or the socio-economic situation may add to this analysis. Thus, one should always take several parameters into account when trying to explain how a certain social group will use language, and not focus on the parameter in question only, as was done here when comparing intra-gender variation. In order to take more criteria into account, and to go even further in our analyses, the next chapter will focus on individual cases and examples to better account for the observations made so far.

[145] See Part 1 for a review of these stereotypes.
[146] Or, in this case, when tweeting.

Chapter 9: Collocational analyses

[C]ollocation networks […] can be used to operationalize the psychological notion of the 'aboutness' of a text
(Brezina et al., 2015: 142)

As explained earlier, studies have shown so far that what differs between women's and men's use of swear words 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 analyze 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. Following up on the examination of the patterns displayed among each gender and age group in the previous chapters, the present chapter will focus on specific cases highlighted through collocational analyses. This will allow us to build on the tendencies observed previously to study them in the context of the words frequently occurring in their vicinity to get a deeper insight into the gendered patterns. This will be a way to mainly focus on specific cases in order to carry out more qualitative analyses which will complete the ones seen so far.

Section 3.9.1 presents the tool used to carry out these analyses, as well as the parameters taken into account.

Sections 3.9.2 to 3.9.9 then each focus on various specific cases, namely the words *fuck*, *fucking*, *wtf*, *bitch*, *bloody*, *cunt*, *#euro2016*, and female names. These have all been spotted in our earlier observations as being linked to specific gendered (but not necessarily) patterns. Thus, studying them in the context of their collocates, and also looking at isolated examples of tweets will be a way to more thoroughly study how these words are used, and to better understand the implications these have on gendered patterns.

3.9.1: About the LancsBox tool:

In order to better understand the patterns which have been noticed earlier, and to confirm or refute these tendencies, I will, thanks to the LancsBox software[147], carry out collocation analyses on the (swear) words highlighted earlier. Collocations are a way of gaining deeper insight into what a text is about, and thus, in the case of swear words, collocation analyses can reveal what swear words are used to talk about. The means of analysis used earlier already provided valuable information regarding what swear words are used to talk about, and we have seen that Twitter users talk about sports, TV shows, other people etc. However, collocation analysis through the LancsBox tool allows to go beyond that by providing a visual representation of every one of the words co-occurring with the central node (i.e. the word of interest), while giving the opportunity to expand the context of any one of these collocates to see their own collocates, as can be seen in Figure 9.1 below, which will be taken as an example:

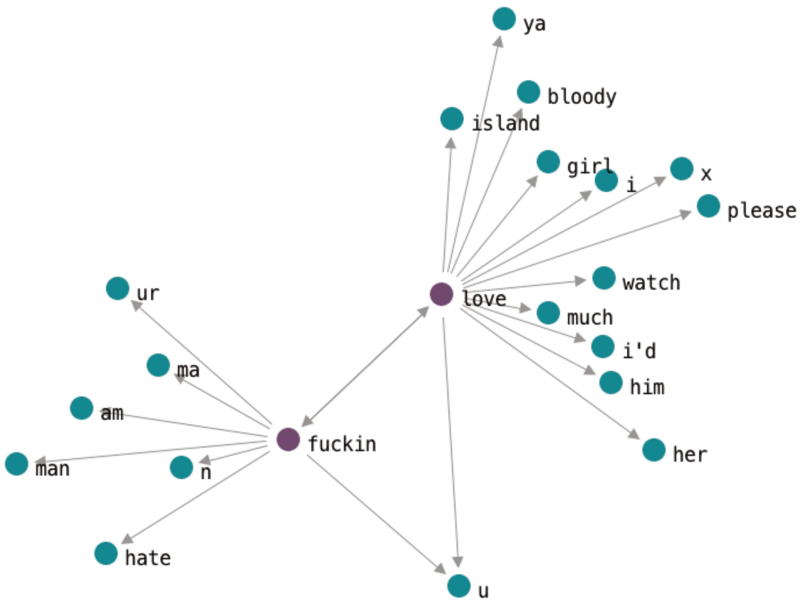


Figure 9.1: Collocates of *love* and *fuckin* in vulgar tweets for all women

In this graph, the central node is *love*, the words linked to it by an arrow are its collocates, the shorter the arrow, the stronger the collocational link between the two words. In this example, we can observe that *island* is a strong collocate, which echoes our earlier findings about the TV show *The Love Island*. As explained above, it is possible to expand any node in order to see its collocates. In Figure 9.1, this has been done for *fuckin*, which appeared as a salient collocate of

[147] See http://corpora.lancs.ac.uk/lancsbox/ for more details regarding the software.

love. The own collocates of *fuckin* then appear, as well as a collocate which is common to both *love* and *fuckin*, the abbreviation *u*. This illustrates the possibilities offered by the tool in terms of the flexibility one has to explore relations between words. This is crucial, as to "appreciate the complexity of the […] discourse we also need to look at how the immediate associations are connected with one another and, more importantly, how these are connected to other (more distant) associations" (Brezina et al., 2015: 155).

The tool also provides various parameters one can set up in order to decide the statistics applied to calculate the strength of the collocations (i.e. the association measure), the size of the left and right collocation window (i.e. the span), and the minimum frequency. These parameters are of primary importance in determining which words will appear as collocates or not. The same settings have been used for the majority of the cases which will be presented in this chapter. The reasons why I made these choices are detailed below.

Association measure

The association measure can be understood as the statistical test calculating the strength of the collocational link. Brezina et al (2015: 145) mention that association measures can be seen "as different ways of comparing the observed and expected values, putting different weight on different aspects of the collocational relationship". The association measure chosen was the MI Score, which was set at 4 in most cases (the default value being 3). So, candidate collocates displaying an MI score smaller than 4 were discarded, and thus do not appear on the graph. The MI score has been presented by Brezina et al. (2015: 151) as being "an association measure commonly used in corpus studies and implemented in a large number of corpus tools". This measure then represents a widely used one, which will allow for comparisons with other works to be made more easily. Setting the MI score at 4 instead of 3 (for the default cut-off value) allows to focus only on the words displaying the strongest collocational link. This has two main advantages: 1) It reduces the number of collocates appearing on the graph, making it more readable. Indeed, when several nodes are expanded, the number of collocates, nodes and arrows present on the graph can multiply so quickly that it becomes difficult to determine the relations between the nodes. 2) An MI score of at least 4 allows to only focus on words having a strong link with one another, and thus on the most salient patterns.

Collocation frequency

As the word implies, this parameter dictates the minimum number of times a candidate collocate has to appear around the word of interest for it to be taken into account. In most cases, the minimum collocation frequency was set at 10 (the default being 5). This serves as a way to only count the collocates which are frequent enough to be representative of an actual trend, in contrast to collocates which appear only once (hapaxes) or twice; these low-frequency collocates may be due to a spelling mistake, linguistic innovation or any other reason, making the collocation irrelevant in this case. Another key advantage of setting the minimum frequency at a higher threshold is the fact that it may compensate certain potential drawbacks of choosing the MI score as an association measure (see Poudat and Landragin, 2017: 203). Indeed, as Brezina (forthcoming) explains, "[s]ome collocation measures such as MI highlight the exclusivity of the collocational relationship favouring collocates which occur almost exclusively in the company of the node, even though this may be only once or twice in the entire corpus". Depending on one's needs, focusing on exclusivity only may be desirable, so in itself this is not necessarily a drawback. However, in my case, this tendency may result in hapaxes and improperly spelled words to be highlighted, especially because of the inconsistency of tweets in their grammar and spelling. Thus, increasing the minimum frequency threshold will ensure that only words which are both exclusive and not infrequent will be focused on.

Span

The span, or collocation window, chosen was 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. So, in the case of a tweet which would be "love eating cheese and crisps, but I don't like syrup fruits", *cheese* would be considered as a candidate collocate, but *I* would not if we were to select *love* as the central node here again. A span of five words to the left and five words to the right is a relatively standard one in studies analyzing collocations, and in the case of tweets, which are limited in length (and thus words), this means that the entire tweet will often be taken into account.

Depending on all those parameters then, not all collocates will appear. Again, this is not necessarily an issue, as one has to choose these parameters in order to filter the corpus and highlight the relevant collocates only. This is what Brezina et al. refer to when they say that "the graphs produced by the tool are exploratory in nature rather than providing a single answer

to the question of connectedness between words, as is collocation itself" (2015: 154). As mentioned several times before, replicability is key in any project, as it is the only way for others to test what is asserted in a study. For the sake of replicability, being as detailed as possible in the methodology used to carry out analyses is of major importance. Because of the number of parameters one has to take into account when analyzing collocates, and to provide a model allowing for more clarity about these parameters, Brezina et al. (2015) proposed what they called the Collocation Parameters Notation (CPN), and which they describe as follows:

> CPN has seven different parameters. Statistic ID refers to the number in the ID column of Table 3.3. 'a' after the statistic ID signifies an uncorrected and 'b' signifies corrected version of the same statistic […]. This is followed by the name of the statistic and the statistic cut-off value used (in brackets), the span of the left and the right context, the minimum frequency of the collocate in the whole corpus, and the minimum frequency of the collocation (i.e. the co-occurrence of the node and the collocate). The last parameter, the filter, specifies any further procedures in the collocation extraction process, for example a removal of certain words from the results (e.g. based on word class membership), or a minimum dispersion value.

Table 9.1[148] below summarizes everything this notation takes into account:

Stat ID	Stat name	Stat cut-off value	L and R span	Min. collocate freq. (C)	Min. collocation freq. (NC)	Filter
3b	MI	4	L5-R5	10	1	No filter
3b-MI(4), L5-R5, C5-NC1; no filter						

Table 9.1: Reproduction from Brezina (forthcoming) regarding the CPN

This notation will thus be used in the titles of the figures displaying the collocation networks under study, as a way to provide all the parameters easily and clearly, using one consistent format.

The collocational analyses which will follow will be based on the observations made in previous sections, and will aim at analyzing them in greater depth to provide more details regarding the organization of these patterns. Some cases will probably reveal to be more fruitful than others, but I will try to be as exhaustive as possible. However, reviewing everything is not possible, especially because of the possibilities offered by the LancsBox tool. Indeed, since any

[148] This table is directly inspired from Brezina (forthcoming).

node can be expanded to look deeper into the network, we virtually have infinite possibilities of exploration. Time and resources however, are not unlimited, so some choices will have to be made. For this reason, I will primarily focus on the tendencies which have been spotted so far.

Another important methodological issue which needs to be raised regarding the way I analyzed the data, is that in most of the following cases, I chose to use the vulgar sub-corpora only, and not the whole corpora of male or female tweets. So, when dealing with the way *fuck* is used by men and women, as it will be the first collocational analysis I will carry out, I focused on tweets containing swear words, and not on all tweets. This may at first seem like a limitation, as the first order collocates will not be the same in both cases. Taking the example given in Figure 9.1 above, the direct collocates of *fucking* may not be the same depending on whether I focus on vulgar tweets only or not. Indeed, when focusing on all tweets, what the LancsBox tool focuses on when asked to find the collocates of *fucking* is the frequencies of these collocates, and how many times other words appear in the proximity of *fucking*. So, these figures will not necessarily be the same from one corpus to the other, and thus the results will differ. However, and as we will see later, the tendencies are likely to remain stable, and the strongest collocates of the words will certainly remain whether I take into account vulgar tweets only or not. The reason why I chose to focus on vulgar tweets only is mainly because of the limited resources I have access to in terms of computing power. The LancsBox tool and my computer both having their limits, being able to efficiently process as much data as there is in the corpora containing all tweets is difficult (if not impossible in certain cases) and takes much longer. Actually, there is no ideal solution in this case, and focusing on vulgar tweets only simply means that the data will be filtered further, but does not limit the validity of the observations made. The only thing to keep in mind is that most of these observations apply to vulgar tweets only and may not necessarily hold in non-vulgar contexts or in bigger corpora. However, as I have explained earlier, context on social media, and especially on Twitter, is something which evolves very quickly because it is very dependent on what people react to. Thus, it is likely that the collocates observed in this study may differ from another collection of British tweets which would be carried out at a different moment in time anyway. However, this does not mean that these results are irrelevant, but simply that we should focus on what is the most salient, and thus on what is the most likely to remain stable in a different dataset[149]. By doing so, we are also likely to focus on what is the most stable in vulgar and in non-vulgar-only tweets, meaning that the decision to focus on

[149] In a different dataset collected using the same social, geographical and methodological parameters of course (i.e. focusing on the UK, on women and men from the same age groups etc…).

vulgar tweets only should probably not be perceived as a limitation[150]. I will, however, try to compare some of the observations made in vulgar contexts to the non-vulgar ones in order to see how comparable these can be, but the technical limitations prevent me from doing this systematically.

3.9.2 About *fuck*:

The word *fuck* is the one which has been shown to be the most popular swear word for both women and men from all age groups. Although these results took into account all the variants of *fuck*, I decided that looking at the collocates of the root form of the word (i.e. *fuck*) could be interesting. Without getting into too many details regarding the collocational networks, Table 9.2 provides an overview of the most salient collocates of the word for each age group and gender, the words in red being common for women and men of the same age group:

[150] Although it would of course be interesting to do a comparable study using all tweets, and not vulgar ones mainly, in order to see how similar or different the observations made here are.

All females	All males	F12-18	M12-18	F19-30	M19-30	F31-45	M31-45
sakes	sake	sake	sake	sake	sakes	off	sake
sake	sakes	hurry	thank	shut	sake	you	give
shut	thank	knows	shut	thank	thank		off
thank	shut	thank	off	gives	da		can
hurry	off	shut	savage	actual	off		as
off	actual	grow	bale	off	actual		all
knows	wit	off	knows	flying	outta		what
grow	flying	both	sheep	wales	flying		
flying	da	needs	bored	knows	shut		
gives	hungover	actual	actual	boys	tae		
outta	hurry	what	wales	what	11/06/2016		
actual	knows	did	hahaha	as	moving		
british	outta	nah	did	boy	knows		
jason	savage	as	wrong	jason	managed		
what	bale	yourself	kane	yes	bale		
andy	moving	doing		give	savage		
chill	robbie	wrong		happened	russia		
wales	11/06/2016			yourself	gives		
as	chill			wrong	em		
yourself	gives			voted	happening		
boy	russia			anyway	happened		
nah	tae				19		
wrong	ramsey				give		
needs	wales				means		
harry	19				yourself		
happened	em				wales		
give	give				as		
did	managed				hodgson		
	croatia				did		
	bored				what		
	did						
	what						
	happened						

Table 9.2: Most salient collocates of *fuck* for women and men according to age and in vulgar tweets[151]

[151] The CPN are: *3b-MI(4), L5-R5, C10-NC1; no filter* for all groups.

As can be observed in Table 9.2, up to and including age group 19-30 the majority of the collocates are common to males and females. This may not seem so surprising, *fuck* being the most popular swear word in the corpus, and the word itself being fairly common now, as we have seen earlier. Thus, it seems reasonable to imagine that at least part of the instances of *fuck* will occur in phrases or contexts which are shared by many users/speakers. And indeed, by looking at the top collocates we realize that the most frequent ones are shared across the various age groups taken into account here (*sake*(*s*), *off*, *what*…), which implies that a core set of expressions including these words and the word *fuck* are common to a lot of users, regardless of their age or gender, as the following examples show:

(#022) We've voted out, fuck sake. Odds on me getting deported 🤔😂

(#023) customers proper fuck me off. "Wheres the chicken" GO DOWN THE MEAT AISLE LOVE, that'd be a start X

(#024) Twitter has been so shit recently. It's been all politics. Please fuck off

(#025) The fucking exhaust's came off my car. What the fuck 😤

Fuck used in these cases has been referred to as being part of idiomatic set phrases (Thelwall, 2008), and generally denotes situations in which *fuck* is not used with a literal meaning (i.e. sexual intercourse). Thus, there seems to be an array of uses of *fuck* which are shared by women and men.

On the other hand, there are also collocates which are not shared by both genders. Most of the collocates of *fuck* which are specific to males are related to football, football teams, or football players (*savage*, *bale*, *robbie*, *wales*, *russia*, *rasmey*, *croatia*, *kane*). This is not surprising either, as we have seen before that one of the topics which distinguished men from women most of the time was that of football. The collocates of *fuck* which are specific to women are related to contestants present in TV shows (*jason*, *andy*, *harry*), or words used to refer to or address other people (*yourself*, *boys*), but these words are more heterogeneous than among men. Overall then, the observations made by studying the kinds of collocates of the word *fuck* are in line with the previous findings, no matter the age of the users. By studying collocational patterns, it is possible to get the general contexts in which certain words/names are mentioned, and to illustrate this I will take the example of *jason* for women, which is a first name present as a top collocate of *fuck* for women as a whole as well as for the 19-30 age group, and *robbie* for men, which is present as a strong collocate of *fuck* in every case (apart from the 31-45 age group).

Jason is the name of the contestant who won the 2016 edition of the TV show *The Big Brother*, and Robbie (Savage) is a football pundit and former player. Figure 9.5 shows the collocational network of the words *fuck* and *jason* for all women in vulgar tweets:

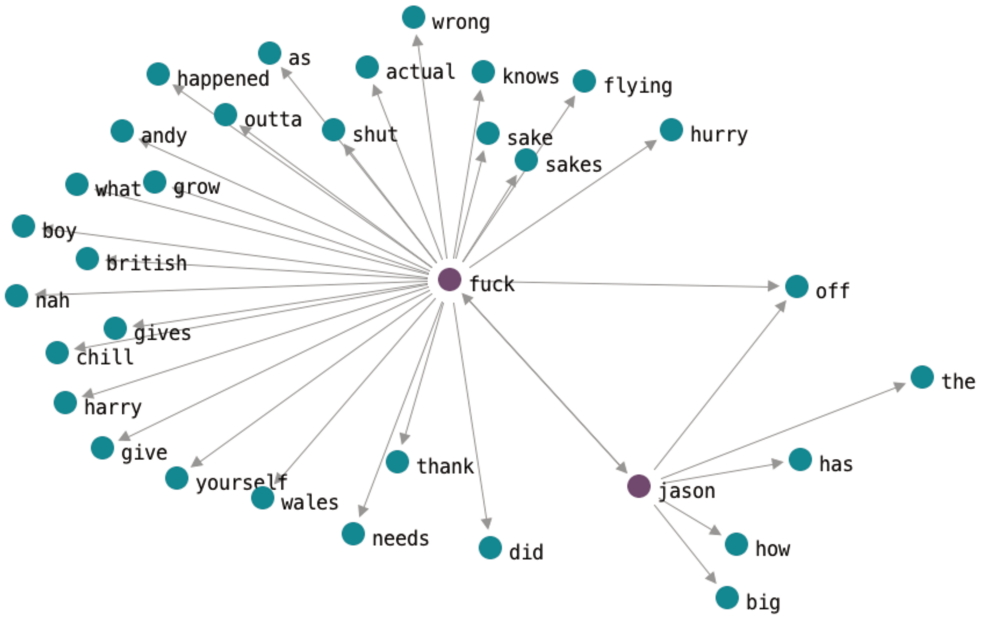


Figure 9.5: Second-order collocates: 3b-MI(4), L5-R5, C10-NC1; no filter

The fact that *big* is a collocate of *jason* is not a surprise, as this is linked with the mention of the TV show in which this contestant appears. What is interesting to observe is that *off* is the only common collocate of *fuck* and *jason*, which implies that this combination of words appears frequently enough for them to be considered as collocates. As observed in the examples above (and as may be known by the reader), *fuck off* is a phrasal verb which can be used as an equivalent of *go away*, or to refer to unimportant things which can waste time. Thus, the lack of other, more positive collocates implies that this contestant mainly seems to be referred to in a negative way, as the concordance lines confirm:

Occurrences 20/5 229 **Texts** 18/28 674 ▼ **Context** 25 ▼ **Corpus** f_ALL_V

Left	Node	Right
#bbcharlie FUCK OFF YOU OBSESSED FAKE ASS SLAG!!!! Leave Jason the	fuck	alone you fucking controlling whore! He don't want you now fuck off!
What the hell what has #bbjason been saved what the	fuck	is the public doing Jason should be □eavinf□□□□□□□
How the actual	fuck	is Jason still in there and Andy out?! #BBUK
Serious question though. How the	fuck	is Jason still in big brother?
How the	fuck	has Jason managed to be there □ #BigBrother
What a joke!!!!!	Fuck	off Jason
	FUCK	OFF JASON FUCK OFF FUCK OFF KYS NAH NAH FUCK THIS NAHHHHHH
FUCK OFF JASON	FUCK	OFF FUCK OFF KYS NAH NAH FUCK THIS NAHHHHHH
FUCK OFF JASON FUCK OFF	FUCK	OFF KYS NAH NAH FUCK THIS NAHHHHHH
WHY THE	FUCK	HAS JASON WON!!!!!
The show is fucking rigged how the	fuck	did Jason win #BBUK
How the	fuck	has Jason won
Who the	fuck	has been voting for Jason
	Fuck	off Jason you miserable twat
How the	fuck	has Jason won Big Brother, HOW!
@mention is my winner regardless of the result Jason	fuck	off!!!!!
Sorry but how the	fuck	has Jason won big brother?? #BBUK
How the	fuck	did jason win wtf! #bbuk
HOW THE	FUCK	HAS JASON YES JASON WON
	Fuck	jason todd and his fucking tank □ #arkhamknight #worstbossever

Figure 9.6: Concordance lines for the node *fuck* and *jason* as a collocate

Beyond the linguistic aspect of this analysis, it is interesting to note that in this case, the women in this sample mainly complain about this contestant winning the show, despite the fact that votes from the audience made him win. We can wonder then whether these are the males votes which mainly allowed him to win, or whether women who supported him were less likely to tweet about him[152].

Let us now look at the collocates of *fuck*, *robbie*, and *savage* for all men in vulgar tweets[153]:

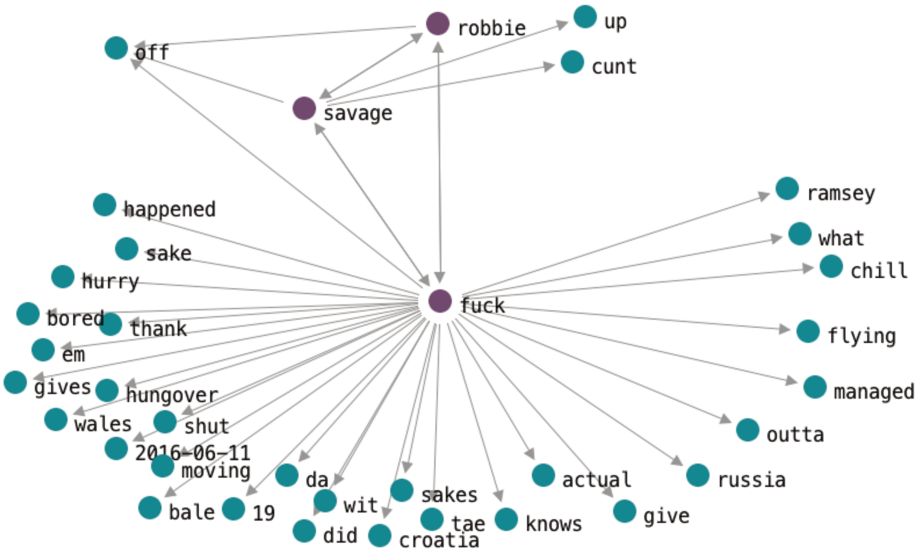


Figure 9.7: Third-order collocates: 3b-MI(4), L5-R5, C10-NC1; no filter

[152] A quick look at the collocates of *jason* with the same parameters for men revealed no relevant one, leaving this question unanswered…

[153] Note that the collocates of *robbie* and *savage* in all tweets are very similar, the only notable difference is that *cunt* is not present in this case.

The results are very similar to those obtained for *jason*, and we can see a clear pattern of both *robbie* and *savage* being used almost exclusively in the context of *fuck* and *off*, with *savage* being also frequently used with *cunt*. It seems that the situation is the same here with *robbie* as it was with *jason* then, and the concordance lines confirm it as well:

Occurrences 19/7 024 **Texts** 17/33 906 ▼ **Context** 25 ▼ **Corpus** m_ALL_V

Left	Node	Right
ROBBIE SAVAGE SHUT THE	FUCK	UP
	Fuck	Robbie savage fuck Wales fuck bale fuck sheep fuck you all
Fuck Robbie savage	fuck	Wales fuck bale fuck sheep fuck you all
Fuck Robbie savage fuck Wales	fuck	bale fuck sheep fuck you all
Robbie savage	fuck	off will ya
Wish Robbie Savage would shut the	fuck	up tbh.
Robbie Savage might hopefully	fuck	off when they've gone out.
What the	fuck	is Robbie Paul doing for the *@mention* tonight...talking random bollacks that's what #CCWigHul
Robbie Savage? Do	fuck	off
	Fuck	off Robbie savage
	Fuck	off Robbie Savage you stupid cunt
Robbie savage can	fuck	off honestly he's talking out his Arse
Shut the	fuck	up robbie savage
I fucking hate Robbie Savage	fuck	off you mug
Robbie savage can	fuck	off as well
@mention and	fuck	off robbie savage
	Fuck	Off. Robbie. Savage. You. Prick
I wish Robbie savage would shut the	fuck	up
Can Robbie Savage just	FUCK	OFF

Figure 9.8: Concordance lines for the node *fuck* and *robbie* as a collocate

The two examples chosen so far imply that when *fuck* is used in combination with a name, it may be used as a means of denigrating or criticizing the person. Despite the fact that swear words have traditionally been associated with taboo language, it has been attested that they can be used in a non-offensive or jocular way[154]. Thus, can we conclude that in the case of *fuck* on Twitter, the word is only used as an offensive word when in combination with a name? Another case study may shed more light on this, and I will give the example of two tweets given as a result of selecting *fuck* as the main node, and *bale*[155] as a collocate for all men (this can be seen on Figure 9.7 above):

(#026) FUCK YES GARETH BALE #ENGWAL

(#027) Bale is cool as fuck 😎🔥

These tweets clearly show that *fuck* can also be used to signal something positive, and to express joy, or satisfaction. *Fuck* is not only used as an expression of negative feelings then, but it must be acknowledged that in the case of *bale*, these two tweets were part of a total of 41 tweets mentioning *fuck* and *bale*, the 39 others being apparently denigrative. Although the findings of

[154] See for example McEnery (2005: 112).

[155] Gareth Bale being a Welsh football player.

this analysis cannot necessarily apply to contexts other than those in which these tweets were collected, it may seem that here, whether male or female, *fuck* mainly functions as an expression of negative feelings when directly addressing someone. However, this also raises another important point, which is that in written discourse, and especially in tweets, irony or sarcasm is very hard to detect. Indeed, what appears as offensive to someone who is exterior to the context of production of the tweet may very well be considered as friendly by a relative of the person who emitted the tweet. As mentioned earlier, context plays an important role in determining how we should interpret swear words, and in this case Twitter may not be the most adapted tool. A perfect example to illustrate this is another collocate of *fuck* for males 12-18 years old, *hahaha*, as shown in the concordance lines below:

Occurrences 14/2 089	Texts 14/9 725	Context 25	Corpus m_12-18_V
Left	No...		Right
Hahaha	fuck		off Russia
@mention hahaha	fuck		up, go to sleep
@mention hahaha	fuck		sake it's been a bad day □
@mention oh	fuck		hahaha I thought it was meant to be a fallow year
	FUCK		OFF KEEMSTAR BLOCKED ME HAHAHA
Senior doesn't give a	fuck		hahaha *@mention* *URL*
I try to text my friends back home when I wake up and then I'm like	fuck		they're all passed out hahaha
Yesssss hahaha	fuck		off Wales
Hahaha	fuck		off roofe, Leeds are having a class transfer window tbf and surely now Oxford will be favouries to come straight back
@mention hahaha does she	fuck		xx
@mention what the	fuck		hahaha
Hahaha literally I give up	fuck		you
hahaha am i	fuck		*URL*
@mention you inspired me to watch it and hahaha	fuck		sake what the hell

Figure 9.9: Concordance lines for the node *fuck* and *hahaha* as a collocate

For many of these examples, it can be hard to determine whether *fuck* and the constructions in which it may be included are meant to be offensive or not. At first it may seem that *hahaha*, which is used to express laughter, is necessarily used in a jocular (and thus non-offensive) context, but the lack of additional context prevents us from being assured of this. Indeed, in the first tweet, it could very well be that the user is simply teasing one of their Russian friends by writing this, or it could also be that the user is laughing at the fact that the Russian team lost a football match, in which case *fuck off* could be perceived as insulting. This observation also holds true for almost all these tweets, suggesting that the Twitter data cannot support any assertion of a correlation between certain swear words and specific functions. Thus, I will most of the time refrain from proposing such correspondences (unless the data enable me to), and will tend to focus on other aspects when possible.

Beyond this problematic question of the purpose with which swear words are used on Twitter, it could be hypothesized that the variant of *fuck* used has an influence on the connotation of the word, or on the context in which it will be used. We have seen earlier that *fucking*, and especially some of its abbreviations, were preferred by some users (it appeared as a keyword

for females aged 12-18). It then seems relevant to analyze this variant for two reasons. First, to see whether males and females aged 12-18 use the abbreviated form of *fuckin* as well as the regular one in similar ways, and especially to see why the variant *fucking* is preferred by females aged 12-18 as opposed to males of the same age. Secondly, it will be an opportunity to compare them to the way the root form *fuck* is used.

3.9.3 About *fucking* (and its variants):

This word is most of the time used as an adjective or adverb. Table 9.3 below presents the collocates of *fucking* according to gender and age. Here again, the collocates of *fucking* which are common to women and men of the same age group are highlighted in red:

All females	All males	F12-18	M12-18
idiot	rich	idiot	friday
joke	idiot	mess	fuming
idiots	idiots	hilarious	hilarious
christ	joke	joke	idiot
fuming	fuming	fuming	joke
brilliant	disgrace	annoying	tired
mess	insane	angry	goal
hilarious	hilarious	mad	hate
disgusting	disgusting	dead	buzzing
horrible	terrible	hate	hell
rude	buzzing	hot	amazing
annoying	scenes	fed	ball
ridiculous	mental	stupid	awful
hate	clue	absolutely	iceland
awful	christ	hard	hot
fed	nonce		love
stupid	legend		god
angry	payet		class
dead	shocking		sick
mad	awful		stupid
	wank		bale
	jesus		
	owl		
	hate		
	brilliant		
	11/06/2016		
	embarrassing		
	goal		
	friday		
	stupid		
	ridiculous		
	useless		
	unreal		
	kill		
	russians		
	15/06/2016		
	love		
	sick		
	hell		
	ball		
	annoying		
	amazing		
	bunch		
	class		
	god		

Table 9.3: Most salient collocates of *fucking* for women and men according to age in vulgar tweets[156]

[156] The CPN are: *3b-MI(4), L5-R5, C20-NC1; no filter* for both genders taken as a whole, and: *3b-MI(4), L5-R5, C10-NC1; no filter* for the 12-18. In this case, I decided to increase the frequency threshold for all users because

In this case, I chose to focus on women and men taken as a whole and on the 12-18-year-olds only, because as mentioned earlier, the 12-18 age group displayed specific preferences for this word, so I will compare the attitudes of this specific group to all users to see how similar or different younger users are. Contrary to *fuck*, the majority of the collocates of this word are not shared by women and men of the same age group. Indeed, men seem to use the word in more contexts than women, resulting in the majority of the collocates of *fucking* for men to be specific to them. Overall, the collocates of *fucking* can be grouped under certain categories, like the lack of intelligence (*idiot*(s), *stupid*), mental disorder (*mad*, *insane*, *mental*), unpleasant feelings or attitudes (*disgusting*, *horrible*, *awful*, *rude*, *terrible*, *shocking, embarrassing*), "negative" emotions (*fuming*, *annoying*, *hate*, *fed*, *angry*), "positive" emotions (*brilliant*, *love*, *amazing*), religion (*god*, *jesus*, *christ*, *hell*), humor (*joke*, *hilarious*), although I realize that some of these words can also be interpreted differently, or belong to other categories, implying that these are not fixed and can overlap[157]. Concerning the collocates of *fucking* which are specific to men, we notice here again the presence of words related to football, or used in this context, which I realized after looking at the concordance lines (*payet*, *goal*, *russians*, *ball*, *iceland*, *bale*).

About the collocates which are shared by both genders, here again, we notice that the tendency, as with *fuck* earlier, is for the collocates of *fucking* to carry negative implications, which seems to confirm the original impression of our analysis of the collocates of *fuck*, which was that the word tends to be associated with negative or undesirable characteristics. Despite the fact that the collocates of *fucking* which are specific to males outweigh those which are shared with women, most of the top collocates are the ones which are common to both genders. This indicates that although some contextual preferences may affect the use of *fucking*, the word is used in comparable ways by males and females for the most part. A look at the second-order collocates of *fucking* did not reveal as much depth or as many intricacies of the collocational networks as with *fuck*. Indeed, I looked at second-order collocates for *idiot*, *joke* and *hilarious*, which are part of the top collocates shared by both genders, and present in all sub-groups taken into account. However, the collocates of *joke* were very similar for both genders and age groups

of the great number of collocates present for these groups. By increasing the frequency threshold then, the graphs were made more readable, while still focusing on a relevant number of significant collocates.

[157] This also explains why *positive* and *negative* are inside quotation marks, as although the original connotations of these words may be one or the other, their use in context can vary from these associations and express opposite ideas, as we have seen many times now.

(*what*, *is*, *a*[158]). This is interesting in itself, as it reveals that women and men are similar in their use of the word, and that they often use *fucking* and *joke* together under the form of an exclamation (i.e. *what a fucking joke*), as the following concordance lines confirm:

(#028) what a fucking joke that I have to go back to college I wanna watch the football for fuck sake

(#029) I have two holiday to pay for and I've lost my job what a fucking JOKE

This is linked to what was said earlier about *fuck*, and the fact that there is a set of phrases and contexts of use of swear words which are implemented among users' linguistic patterns. Indeed, these patterns were true of both genders, when analyzed as a whole as well as among the 12-18 age group.

Concerning the collocates of *idiot*, which is also shared as a collocate of *fucking* for both genders and groups taken into account, the only collocate for women is *you*, and *a* for men. This is in line with previous observations[159] in which I hypothesized that women and men would use certain swear words differently when addressing other people. In this case, it seems that *idiot* is used more often by women than by men when directly addressing other people (with *you*). It does not mean that men do not use *idiot* to directly address other people, but that they do not do it frequently enough for *you* to be considered as a relevant collocate.

Concerning collocates of *fucking* which are specific to women and men, let us look at *mess*, which is the strongest female-specific collocate of *fucking* present among both groups of women considered. Its only collocate is *a* for both groups. For men, *buzzing*[160] is the strongest collocate of *fucking* which is common to both groups, and its only collocate is *for* for both age groups. This is linked to the construction of the expression *to buzz for*, which is used to express excitement, as can be seen in the examples below:

(#030) fucking buzzing for the game tonight, come on #ENG

[158] The collocational networks being very limited in these cases, I chose not to include figures of the networks themselves, as they would not add much to the discussion while potentially overwhelming the reader with many figures.

[159] See previous chapter.

[160] Note that in this case I did not take *rich* and *friday* into account, as these were present here because of the same tweets being published many times by the same author, and can thus be considered as spam, and not representative of more general trends.

(#031) If I don't get Melissa's cold by tomorrow I'll be fucking buzzing with my immune system considering I treat it like shite

What these results mainly show then, is that women and men most of the time use *fucking* in a way which is similar. Although some collocates are gender-specific, and despite the relative abundance of these for men, the top collocates are most of the time shared. The main difference with the observations made with *fuck* is that the collocational networks are much more unilateral with *fucking*, and present fewer connections between the nodes. Contrary to what it may seem[161], this is not less interesting, as it denotes that *fucking* will be used in as many different contexts as *fuck*, but that its collocates will not "interact" with each other. This seems to indicate that there is a greater exclusivity in the links between *fucking* and its collocates.

In the previous chapter, we saw that the abbreviation of *fucking* (i.e. *fckin*) was considered as a keyword for females, and particularly for the 12-18 age group, and this also corresponded to an overall preference for the 12-18-year-olds from both genders to use abbreviations. Here, a look at the collocates of *fckin* does not reveal much, as no collocate of the word is found in any of the age groups taken into account. Even lowering the thresholds[162] of the CPN does not reveal relevant collocates. This is a sign that Twitter users still seem to favor the non-abbreviated forms, although females in certain cases, and the 12-18 age group as a whole, display a preference for abbreviations compared to other (older) groups of users, which explains why these appeared as keywords earlier. About abbreviations, it was mentioned earlier (see Chapter 2) that some studies presented them used in online communication as being another form of self-censorship; this tendency is called obfuscation (Laboreiro and Oliveira, 2014). Showing that abbreviated swear words are used more often by one sub-group (i.e. women) or another would imply that they are used to mitigate the act of swearing. However, in the case of a study based on Twitter, is difficult to assert whether abbreviations are used to avoid swearing, or as a need for economy imposed by the 140 character limit. In order to address this, further study may consider analyzing the length of tweets in which abbreviations are used to see whether these are more likely to be used in longer or shorter tweets.

[161] Or at least contrary to what I may have felt at first.
[162] I went as low as *3b-MI(3), L5-R5, C10-NC1; no filter*.

We are now going to analyze another abbreviated form of *fuck* which also appeared as a keyword in certain cases: the abbreviation *wtf*.

3.9.4 About *wtf*:

Among all the abbreviations of *fuck* taken into account in this thesis, *wtf* is the second most frequent one, as it appears 872 times in total[163]. The reason I chose to focus on *wtf* and not *ffs*, being more frequent, is that *wtf* appeared as a keyword for 12-18 males when compared to 19-30 males (see previous chapter), while *ffs* did not appear as being as salient. We saw earlier that although *fuck* (and its abbreviations) is the most popular swear word for both genders, males tend to use it more often. However, we also observed that when focusing on the abbreviated forms of *fuck* only, women were the ones favoring them, as well as the 12-18 age group overall. Thus, *wtf* seemed interesting as it represents a point of tension between what is favored by women, but also by men of different ages, as well as having characteristics shared by both genders.

Concerning the collocates of *wtf*, visually speaking, the first thing which can be noticed when comparing the two collocational networks is the fact that there are many more collocates of *wtf* for women, strengthening the idea that females use it (and abbreviations of *fuck* as a whole) more often than men, which we also observed earlier through other means of analysis.
Also, when looking at the overall patterns, we notice similarities in the way *wtf* is used by both genders, but we also observe major differences. Indeed, for both genders, a common collocate of *wtf* is *doing*, as can be seen in the examples below:

(#032) wtf are people doing out at night catching Pokemon like its 3am Pokemon Go the fuck to sleep
(#033) Wtf are these idiots doing its a restaurant not a fucking zoo

Doing is the strongest common collocate for both genders. It is also considered a salient collocate of *wtf* for women, as other inflections of the verb *to do* are present as strong collocates of *wtf* (*do*, *does*, *did*).
For these reasons, I decided to expand it to see how women and men use it, as can be seen in Figures 9.10 and 9.11 below:

[163] The most frequent abbreviation being *ffs*, which appears 1040 times in total.

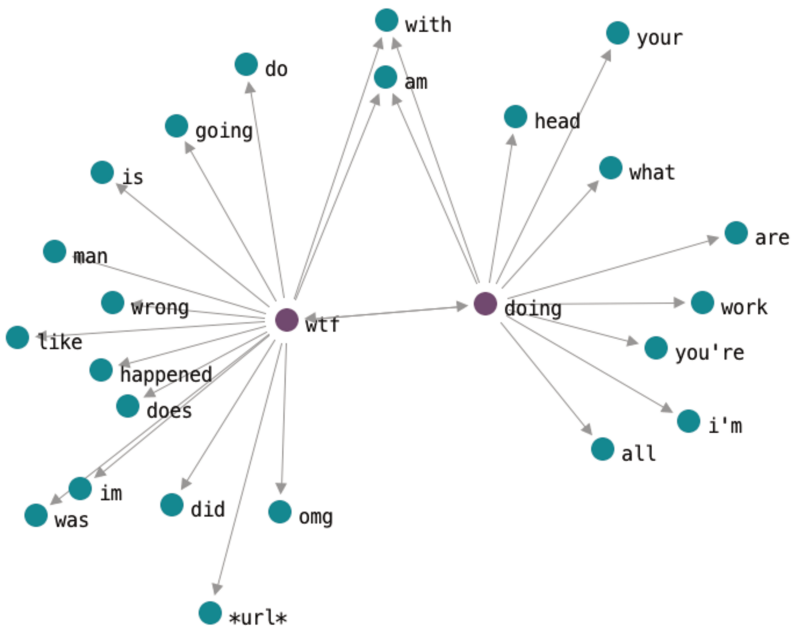
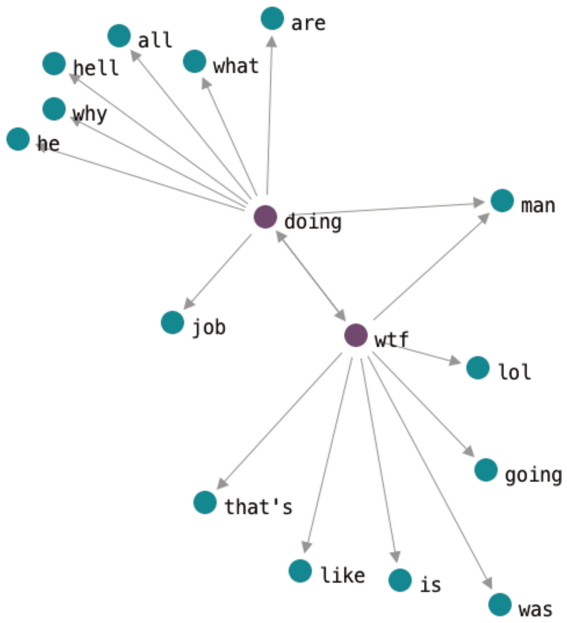


Figure 9.10: Second-order collocates for all women in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter



Figures 9.11: second-order collocates for all men in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

In doing so, we realize that a good deal of the collocates of the word are common to women and men (*wtf*, *what*, *are*, *all*, *work/job*[164]). On top of that, one word for men collocates both with *wtf* and *doing*: *man*, and there are two of these collocating words for women: *am*, *with*. The fact that these words collocate both with *wtf* and *doing* does not necessarily imply that the three will all co-occur frequently, although this is a likely possibility. What this indicates is at least that there is a significant link between *wtf* and *doing*, and between *wtf* and *man*.
However, this difference in the common collocates of *doing* and *wtf* between women and men is a noticeable difference, as *am* signals that the first person pronoun *I* is used by the speaker (or tweeter here), as in the cases below:

(#034) why am I doing jaegerbombs on a Sunday, especially after last night fuckin hell
(#035) Doing fucking well with my healthy eating then work piss me off so I eat cake wtf who am i

On the other hand, *man* is mainly used in this corpus as an interjection (when no one is targeted), as a term of address (when *man* is used to address or mention someone), or as a noun, as the examples below illustrate:

interjection ==> (#036) I just bought some Cheetos and got like 10 inside them, wtf man!
address term ==> (#037) *@mention* been fucking time man, what you doing now?
noun ==> (#038) Get u a man like Kanye wtf *URL*

So, apart from the cases where *man* is used as an interjection, for males it is used to refer to other people. In other words, *wtf* and *doing* are both frequently used by women and men, but for different purposes. Women use it to refer to themselves mainly, and men use it to mention other people, or as an interjection. This observation for females also explains why *im* is another collocate of *wtf*. This once more proves the need to go beyond basic observations, and especially beyond the mere focus on what is different in a corpus, to also delve into what is similar[165]. Furthermore, as can be seen in the expanded collocation network, females use *wtf* to talk about actions (*do*, *did*, *does*, *happened*), and particularly about ongoing actions or states (*doing*, *going*). As we have just seen, they also mention these actions or states to refer to themselves,

[164] I added these words in the list of common collocates as they can be considered as synonyms, although I realize that their meanings can vary in certain situations.
[165] Especially as the example I just gave shows that looking at what is similar can also point to differences!

which is made evident in Figure 9.12 by the links present between *wtf*, *am*, and *going*, as well as the links between *wtf*, *am* and *doing*:

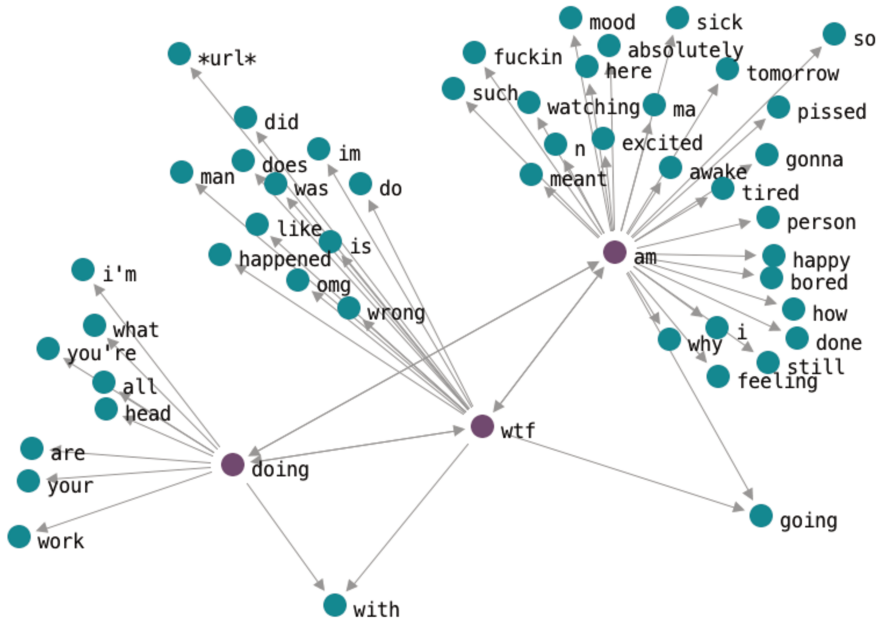


Figure 9.12: Second-order collocates for all women in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

I mentioned in the previous chapter that emotional states were a recurring topic for women compared to men when analyzing keywords. Here again, we can observe that this indeed seems to be a major center of interest for women, as a lot of the main collocates of *am* refer to emotions (*excited*, *mood*, *pissed*, *happy*, *bored*, *feeling*). The fact that these are strong collocates of *am* (used, as I mentioned earlier, to talk about oneself) then confirms the key importance of emotions in female discourse on Twitter. Another category of words present as key collocates of *am* is intensifiers (*such*, *fuckin*, *absolutely*, *so*). This indicates that female tweets often (although not necessarily mostly) refer to their emotional states in an emphatic way, as the examples below show:

(#039) I can't even explain how excited I am for tonight's Game of Thrones! #battleofthebastards

(#040) Literally in such a shit mood fuck I am never like this I'm so angry oh my CHRIST

For males, on the other hand, the only common collocate of *wtf* and *doing*, as mentioned earlier, is *man*. It is interesting to note that depending on what *man* collocates with, it seems to have

different functions. Indeed, as can be seen in the concordance lines in Figure 9.13 below, when used with *wtf*, *man* is mainly used as an interjection:

KWIC: wtf > man

Occurrences 15/330 Texts 15/33 906 Context 30 Corpus m_ALL_V

Left		Right
girl on this team is literally dressed as that blonde girl from Frozen,	wtf man	
Jon Jones man	wtf	are you playing at! That must be him done?!
I just bought some Cheetos and got like 10 inside them,	wtf man!	
Seriously Ezra man	wtf	you doing mate. You seen Aria?
Fucking found the cunt down the side of me bed man	wtf	
Chloe is the most vile thing to step foot onto Geordie Shore	wtf	is she man *URL*
Military coup in Turkey,	wtf man?!	
I love my late night tweets, people look at them and think wtf, that's cheesy af or	wtf	he talking about. Man I just want to enjoy life
□□□□	wtf man	*URL*
ll yea lizard planet! WINDOWS™: restart planet for important updates GOD: um ok *Dinos die, man appears* GOD:	wtf	
Decent goal from bale but fuck same Joe hart man	wtf	
@mention	wtf man	
	Wtf	is the world. A man with a bloody axe goes on a rampage on a train in Germany?! What the actual hell
I've got a reallllly annoying habit of answering questions before I've been asked □ "hello, I'm ok thanks how're you?"	Wtf man	
@mention	WTFman	told you not to do this

Figure 9.13: Concordance lines for the node *wtf* and *man* as a collocate

This seems logical, as the expression *what the fuck* itself is mainly used as an interjection; thus, it does not seem surprising to observe other words having the same function in the immediate proximity of the abbreviated form of this expression.

When used with *doing*, on the other hand, *man* has a different function, as can be observed in Figure 9.14 below:

Occurrences 9/317 Texts 8/33 906 Context 20 Corpus m_ALL_V

Left	Node	Right
Adam Lallana □□□□ wtf is he even	doing	in the 23 man squad, let alone the starting 11
@mention *@mention* trust me □an□□□ man thought he was	doing	bits but he fucked □p□□□□
You can't be fucking	doing	that man-,-
Seriously Ezra man wtf you	doing	mate. You seen Aria?
@mention I literally just went "Oh piss off, I'm gay you've got more chance of me	doing	it to a man love, don't even start"
@mention been fucking time man, what you	doing	now?
Osborne stop chatting shit man don't over exaggerate the ting to make it look like ur man are	doing	a good job
@mention I know that not much but I love you man, keep	doing	what your doing. Entertaining as fk man
@mention I know that not much but I love you man, keep doing what your	doing	Entertaining as fk man

Figure 9.14: Concordance lines for the node *doing* and *man* as a collocate

In this case, the word *man* rarely functions as an interjection, and is most of the time used either as a noun, or as an address term.

We saw that for women the collocates of *am* were mainly words referring to emotional states and intensifiers. Here, the collocates of *man* are mainly intensifiers/interjections (*fuckin*, *absolute*, *ffs*[166]), and one word related to football (*team*), as can be seen in Figure 9.15 below:

[166] Abbreviated form of *for fuck's sake*.

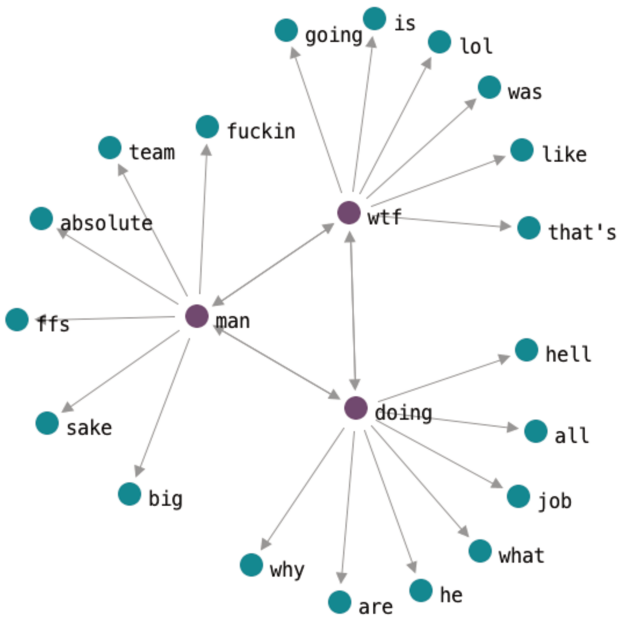


Figure 9.15: Second-order collocates for all men in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

Here then, it is shown that swearing is used in similar, yet different ways for both genders, but this additionally confirms some of the observations made earlier about the references to emotional states for women, or the differing ways of addressing other people when swearing.

3.9.5 About *bitch*:

Bitch has been shown earlier as the swear word being the most specific to females, and we will see if the collocational analysis confirms this trend, and try to understand why this is the case. A first look at the collocates of the word for various sub-groups seems to confirm this, as can be seen in Table 9.4 below, collocates highlighted in red being common to women and men from the same age group:

All females	All males	F12-18	M12-18	F19-30	M19-30
resting	little	face	a	resting	little
basic	#bbuk	stupid	you	fat	such
moody	such	such		face	your
natalie	her	little		little	a
lil	she	ass		such	
fit	being	a		stupid	
biggest	a	you're		please	
face	know	being		she's	
such	your			#loveisland	
little				always	
lucky				#bbuk	
stupid				being	
bye				a	
fat				hate	
birthday					
she's					
#bbuk					
please					
#loveisland					
a					
being					
ass					
you're					

Table 9.4: Most salient collocates of *bitch* for women and men according to age in vulgar tweets[167]

The first thing which can be noticed is the fact that there are a lot more collocates for women than for men in every age group. Although less marked, the same pattern was observed for *fuck* and *fucking* already. This is not surprising, as it seems reasonable to imagine that a word that is significantly more frequent among a sub-group will also be used in more varied contexts, thus producing more collocates. Consequently, this confirms the fact that on top of being quantitatively more used by women, *bitch* is also used in more linguistically rich situations.

Less surprisingly, *bitch* seems to mainly be addressed to other women (*natalie*, *she's*, *she*, *her*) more often than to men. It also appears that the word is often used by both women and men in reaction to some of the TV shows mentioned earlier (namely *Big Brother* and *The Love Island*).

[167] The CPN are: *3b-MI(4), L5-R5, C10-NC1; no filter* for all groups.

Concerning the categories of words collocating with *bitch*, there are many adjectives for women (*stupid*, *basic*, *little*, *lil*[168], *lucky*, *fat, moody*), and it seems addressed to other people (*you're*, *natalie*, *she's*) more often than among men (*she*). This is in line with what was argued earlier about the fact that women and men may differ in their use of swear words when addressing people. Also, the presence of *stupid* as a collocate of *bitch* for all groups of women taken into account echoes the findings we had for *fucking* earlier, which implied that the word was mainly used in such a way, and in this case to denigrate people. Here, the direct evidence of *bitch* being used like this is less obvious, as many of the collocates' potential to offend is not as noticeable (*resting*, *lucky*, *birthday*, *please*), especially among females. So, there seem to be a lot more ways for women to characterize *bitch* (with adjectives) than for males. This seems to be the key to explaining the gendered difference, as this discrepancy is steady across age groups: the array of adjectives used in combination with *bitch* is far greater for women than for men. I will now focus on some of these gender-specific collocates to try to better understand why they are specific to women. Let us consider the case of *resting*, which is the strongest collocate of *bitch* for women as a whole, as well as for women aged 19-30. Its only collocate is *face*, which in turn reveals interesting results when expanded, as showed in Figure 9.16 below:

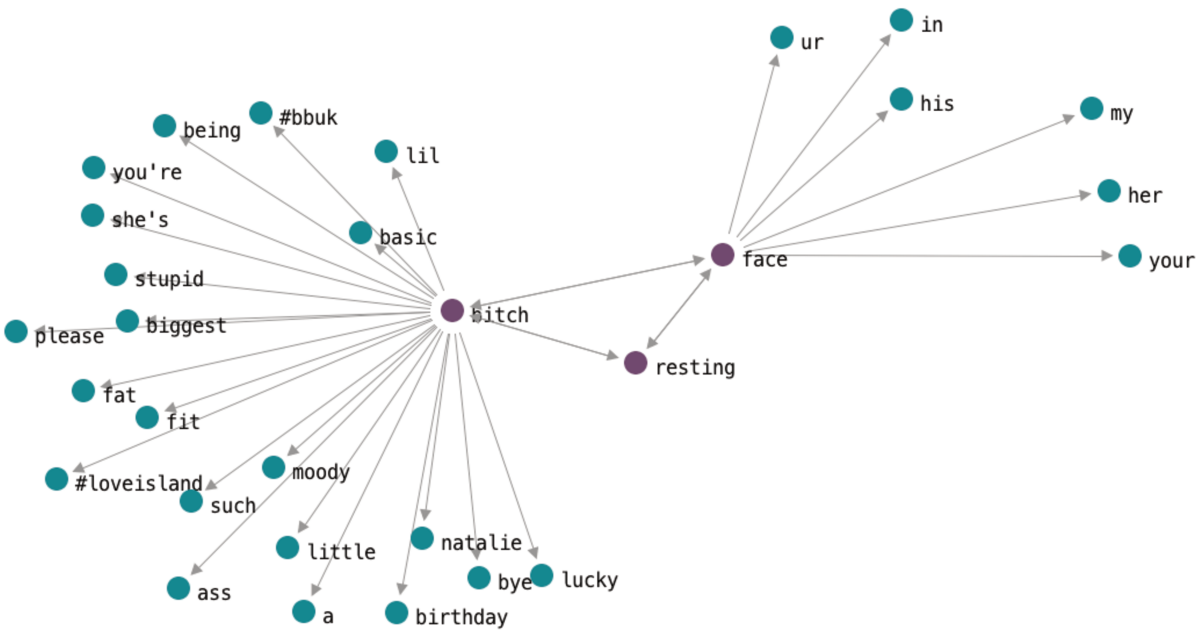


Figure 9.16: Third-order collocates for all women in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

[168] Abbreviation of *little*.

As can be observed, the collocational network seems to reveal a relatively exclusive pattern for resting as its only two collocates are *bitch* and *face*, which in turn also collocate with each other. This collocational triangle could be the sign of the existence of specific phrases mentioning these three words, as we saw earlier with *fuck* and collocates like *off*, *sake*, *what* etc… A look at the concordance lines of *bitch* having *resting* as a collocate revealed that indeed, *resting*, *bitch* and *face* do co-occur together every time in this case:

ch Term	Occurrences 25/1 016	Texts 25/28 674	Context 25	Corpus w_ALL_V

Left	No...	Right
lol everyone is saying #BBGeorgina doesn't actually like #BBJackson but i think she's just got extreme	resting bitch	face all the time
Bit late but v proud to be crowned as having the best "resting	bitch	face" in the rowing squad *URL*
Can't ppl see that #bbgeorgina just has a severe case of	resting bitch	face??? #bbuk
And just cause you've got severe	resting bitch	face it doesn't mean you're absolutely miserable!! #bbgeorgina #bbuk
her	resting bitch	face is still 10/10.... how *URL*
my	resting bitch	face is that bad that I looked at this guy and he said he genuinely thought I was gonna hit him lol a
My	resting bitch	face is seriously starting to ruin my □ife□I'm not a bitch, I don't hate you, nothing is wrong and I'm
I swear that cat has a	resting bitch	face! So funny. She clearly owns the our road! XD
	Resting bitch	face comes out so much in grays town centre, hate the □lace□□
So many eye rolls and	resting bitch	faces for today
	Resting bitch	face #niece #london #clerkenwell #italianprocession @ Clerkenwell London *URL*
i don't have a	resting bitch	face lmao i have a resting sad face, always look sad when im actually so happy
Does Kirsty have	resting bitch	face, or is just a bitch in general?
Love how people would rather stand up on the bus than sit next to me, perks of having a	resting bitch	□ace□□□
Have THE worst	resting bitch	face □
Had to post a smiley one to compensate for the	resting bitch	face *URL*
nave 'talking map' on my forehead. I always get asked for directions, even though I've got the biggest	resting bitch	face going rn □
So many strangers walk past me and tell me to smile.. Do I just have a permanent	resting bitch	face?
@mention	resting bitch	face, always #imhappyhonest
	Resting bitch	face for life
Forever a	resting bitch	face... #nofilterneeded @ Birmingham, United Kingdom *URL*
I have the worst	resting bitch	face in the universe
All I get is 'you're actually really nice' well yeah, there's a personality that comes with the	resting bitch	□ace□
Fs why do I have such a bad	resting bitch	face
Wish I didn't have	resting bitch	face... Always get told to smile

Figure 9.17: Concordance lines for the node *bitch* with *resting* as a collocate

This confirms that these words do collocate together, and consequently this also indicates that the expression *resting bitch face* is mainly used by women. The phrase seems to be a fairly recent one, and refers to a type of facial expression making the person look like they are angry when they are not[169]. Studies (ibid) showed that despite the female-oriented name given to this type of expression, it can be detected in both males and females, and the collocates of *face* in Figure 9.16 imply that women are aware of this, as they use the expression to mention other women (*her*), other men (*his*), themselves (*my*) and directly address people (*your*, *ur*). Although the expression does not appear frequently enough among males for resting or *face* to appear as collocates of *bitch*, some isolated tweets show that the expression is also used by males to talk about other males, as the example below shows:

(#041) There is an elderly man in my local tesco who has the most extreme case of resting bitch face

[169] For more details, see: http://www.noldusconsulting.com/blog/throwing-shade-science-resting-bitch-face

This shows two things. First, although women use the expression more often, and although it does not seem to appear in the collocational networks for males, they do use the expression. Secondly, and perhaps more interestingly, the gender-oriented aspect of the expression (*bitch* being female) does not influence women or men in using it to mention people from a specific gender.

Among the top collocates of *bitch* for all women, we can also observe two other relatively "unconventional" adjectives like *basic* and *moody*. The term unconventional is used here in comparison with most of the other adjectives present as collocates of the word like *little*, *biggest*, *stupid*, *fat*, for which the purpose of the word is relatively obvious: the goal is to characterize the person *bitch* is addressed to. In the latter examples given, the purpose seems to be to denigrate, whereas with the former examples, the meaning of these words in a context where *bitch* is present is not as clear, and may be the sign that these are other gender-specific phrases, as we saw with *resting bitch face*. Although expanding the node *basic* did not reveal any collocates, the concordance lines indicate that the presence of these two words collocating is indeed due to the expression *basic bitch* being used:

Term	Occurrences 11/1 016	Texts 11/28 674	Context 22	Corpus w_ALL_V

Left	No...	Right
	Basic Bitch	
people look at you for wearing Dr Marten heels on a night out. Like bitch these are worth ten times your	basic bitch	heels.
I feel my	basic bitch	level just increased to a new threshold *URL*
It'll be the most basic fitness	Bitch	Instagram posts but I just want to be able to record all my progress and thoughts
Basic goth	bitch	gets Halloween and summer mixed up *URL*
	basic bitch	flower crown □ *URL*
tooooo buzzed for tomorrow *@mention* snapchat story part fookin 2 oi oi can't wait to hear bout	basic bitch	ass zara
You look like a	basic bitch	now was I meant to stay in my scene kid phase forever? No
Is Amy Bolger a	basic bitch	
Love this	basic bitch	frannunn21 □□□□ #Sisters @ Gilgamesh London *URL*
OMG Zara is such a	basic bitch	□□□□ #LoveIsland

Figure 9.18: concordance lines for the node *bitch* and *basic* as a collocate

Here again, *basic bitch* seems to be a fairly new expression used to describe women seen as having plain and unoriginal tastes[170]. The fact that women use these recent expressions significantly more than men echoes previous claims that women would tend to be linguistic innovators (see Coates, 1986: 138).

The same is true of *moody*, whose only collocates are *such* and *a*, and which also appears as being part of a phrase used by women to talk about themselves or other women:

[170] For more details, see: http://time.com/77305/how-conformity-became-a-crime/

Occurrences 14/1 016	Texts 13/28 674	▼ Context 24
Left	**Node**	**Right**
I'm such a moody	bitch	if I haven't eaten □□□
Jackson is your mum a moody	bitch	too *@mention*
why is lucy always such a moody	bitch	on facetime *URL*
Need to stop being such a moody	bitch	24/7
Quitting smoking on Sunday this should be interesting, moody	bitch	Amber is on her way
Demmi actually has a chronic	bitch	face & is constantly moody. Smile bitch.
Demmi actually has a chronic bitch face & is constantly moody. Smile	bitch	
I'm such a moody	bitch	sometimes □
being a moody psycho	bitch	tonight □
1 hour sleep, right moody	bitch	x x
And off I go to sleep because I'll be a moody	bitch	in the morning if I don't sleep soon. Goodnight ❤□
Doesn't help that she was a right moody	bitch	
I've been such a moody	bitch	since coming home from Sheffield □
I've been such a moody	bitch	since coming back from □heffield....□

Figure 9.19: concordance lines for the node *bitch* and *moody* as a collocate

Thus, these examples imply that the word *bitch* may be significantly more used by women because they use the word in an array of set phrases that men do not seem to use as much. The gender of the people to whom users address *bitch* does seem to play a role in this, as men seem to address *bitch* to women (*she*, *her*) more than to men, the same pattern being observed among women as well (*she's*, *natalie*). It is also interesting to note that in the case of the expression *moody bitch*, there is no occurrence of it at all among men. Instead of *bitch*, they prefer using swear words which have been found earlier to be preferred by males, as the examples below show:

(#042) can't be doing with moody fuckers nowadays😬

(#043) When your mums being a proper moody cunt and you've done absolutely nothing wrong😡 *URL*

(#044) Love my moody little shit 💙 *URL*

(#045) Cant be dealing with moody bastards

(#046) I'm such a moody prick sometimes.

Thus, it seems that swear words can be used in set phrases by both women and men to refer to the same concepts (in the case of *resting bitch face*), but that they can still be preferred by one gender over the other. On the other hand, some set phrases can be adapted in accordance with the swear word preferences of the gender (in the case of *moody bitch*/*cunt*), but still preferred by one gender in particular (here, women). At this point, it seems that what may play a bigger role in accounting for these differences could be para, or extralinguistic features, like the cognitive representations of the words and expressions users have, for example. Further studies would be needed to better understand this phenomenon.

I am now going to focus on another word which has emerged as being favored by women, the word *bloody*.

3.9.6 About *bloody*:

Bloody has been shown in this study to be the third most frequent swear word overall, after *fuck* and *shit*. Despite its high frequency, MWU tests revealed that it was significantly more used by women as a whole, and by females aged 12-18. Table 9.5 presents the collocates of this word for both genders according to age, the red ones being the ones common to both women and men from the same group:

All females	All males	F12-18	M12-18	F19-30	M19-30
loved	brilliant	amazing	hell	brilliant	brilliant
brilliant	awful	hell	love	amazing	hell
amazing	hell	wait		hot	awful
awful	amazing	love		excited	hot
hot	hot	than		tired	hope
excited	ha	oh		love	love
hell	weather			annoying	tonight
cute	hope			hell	#euro2016
tired	keep			wait	better
hard	love				good
love	those				
wait	wait				
hope	long				
game	#euro2016				
won't	though				
long	tonight				
gone					
annoying					

Table 9.5: Most salient collocates of *bloody* for both genders according to age in vulgar tweets[171]

As we can see, there are a lot of similitudes between both genders inside all age groups. It may seem then, that as there is apparently no major difference in the collocates, the mere quantitative difference between the use of *bloody* between women and men observed before could explain

[171] The CPN are: *3b-MI(4), L5-R5, C10-NC1; no filter* for all groups.

the fact that it appears as a significant female word (at least for women as a whole, and the 12-18 age group). The fact that females aged 12-18 in particular display substantially more collocates for *bloody* than males of the same age seems to confirm this hypothesis, as this is the age group displaying the greater gendered imbalance in this regard. For the 19-30-year-olds, both genders present a similar number of collocates of the word, and this age group was shown to be gender neutral as far as the word *bloody* is concerned. The link between the gendered aspect of the word, and the difference in the number of collocates there is seems to be confirmed again then.

The words *love* and *hell* are the two strongest collocates of *bloody*, being present in every sub-group taken into account here. They are also the only two collocates present for males aged12-18, who seem to be the ones using the word the least frequently, implying that these two collocates are very steady ones, collocating (according to these parameters) 136 times for women, and 166 times for males (when taken as a whole). This pair is used as a set phrase (*bloody hell*), functioning most of the time as an exclamation, as the examples below demonstrate[172]:

(#047) I lost the cap for my camera while walking home bloody hell

(#048) Bloody hell, our country is going mad!

(#049) Mad nights on the snapchat! Bloody hell 😂

(#050) I do like David Haye but bloody hell does he chat alot of shit😂😂

Bloody and *hell* do not share any collocate in the male or the female corpus. This strengthens the idea that both genders use the two words in combination fairly frequently, but that apart from the set phrase "bloody hell", these two words do not share much. This is interesting as it shows that both genders are relatively similar in the exclusivity of these two words, and this also reveals that two words may not have much in common, but still exist as a defined entity if the collocational link between them is strong enough, which seems to be the case here for *bloody* and *hell*.

[172] Note the use of exclamation marks in some cases to insist on the exclamation. Such uses of punctuation do not seem to prevail however.

Concerning the word *love* now, its presence as one of the strongest collocates of *bloody* for all the sub-groups taken into account is surprising, especially as we saw earlier that other swear words were frequently used to denigrate others in this corpus. The mere presence of *love* however, does not necessarily mean that it is used to show appreciation, as it could very well be part of a tweet similar to the following example:

(#051) Used to love Sunday's and now I bloody hate them

In this case, *bloody* is not used in the same verbal group as *love*, but on the contrary, *bloody* is used to emphasize *hate*. Here then, the fact that *love* is present in the RL5R span specified in the LancsBox parameters may influence an erroneous association of *bloody* with the word *love*, whereas in this specific case the two are not directly associated with each other, as *bloody* is actually referring to a word with an opposite meaning. Another point which should be raised to temper the association of *bloody* with *love* as an expression of appreciation, is the frequent use of *love* as being part of the name of a TV show mentioned several times now; *The Love Island*. Indeed, the following tweets present examples of such cases:

(#052) 13 bloody episodes of love island to catch up on
(#053) I feel like I watch more adverts than bloody love island😒
(#054) Shoulda bloody watched love island instead of this dead final

However, in the case of the TV show examples, of the 146 tweets in which *love* and *bloody* collocate, these three examples are the only ambiguous ones which have been found. In this case, I am considering these occurrences of *bloody* as ambiguous in the context of *love*. Indeed, here *bloody* is not used as an emphatic marker of *love* as a verb, but as *love* being part of the name of the TV show. Thus, the collocational pattern *bloody* + *love* in this case could lead to the erroneous impression that *bloody* is used to express appreciation when it is not. The same principle applies for the case of *love* collocating with *bloody* but not used in a context of appreciation, as a very limited number of such counter-examples were found. Therefore, as I have explained all along in this thesis, we need to be careful when interpreting the contexts in which words are used, but in the case of *bloody* and *love* here, a manual examination of the data revealed that these findings do go against some of the previous observations linking swear words to the denigration of other people. This then represents one more example showing that

despite the taboo nature of swear words, they can in certain contexts (mainly) be used to express appreciation or support.

After reviewing two swear words which have been accounted for as being preferred by females, we are now going to focus on a swear word which has been shown to be one of the most frequently used by males, the word *cunt*.

3.9.7 About *cunt*:

As seen in Table 8.1, *cunt* is, according to the MWU tests, one of the swear words which is the most characteristic of males from all generations. We have also seen in Chapter 8 that one of the characteristics of male use of swear words was their more frequent use of plurals like *fuckers*, and these observations were particularly salient for *cunts*. So, in order to better understand this phenomenon, I provide an analysis of the collocates of the plural form, *cunts*. Table 9.6 first presents the collocates of *cunt* for women and men according to age, the words in red being, as usual, the ones which are common for women and men of the same group:

All females	All males	F12-18	M12-18	F19-30	M19-30
fat	daft	ya	fat	little	daft
ya	hello	such	stupid	an	fat
little	silly	you're	ya	being	stupid
absolute	owl	a	you're	a	useless
such	fat	being	a	you	little
an	massive	u	you		roy
being	savage		u		such
ur	stupid				old
you're	useless				you're
a	hodgson				you
u	welsh				ya
	ya				u
	you're				him
	such				haha
	little				absolute
	looking				hope
	old				a
	you				
	roy				
	u				
	absolute				
	a				
	hope				
	haha				
	look				
	kane				
	him				

Table 9.6: Most salient collocates of *cunt* for women and men according to age in vulgar tweets[173]

Cunt is very similarly used for women aged 12-18 and 19-30, with the use of some emphatic markers like *such*, or adjectives like *little*. There are relatively few collocates overall, which seems logical as we observed that the word was not used so much by women.

Men, on the other hand, use it more often, as there are a lot more collocates than for women. We also observe that men use a lot more adjectives, which could be perceived as a way to

[173] The CPN are: *3b-MI(4), L5-R5, C10-NC1; no filter* for all groups.

increase the strength of the word, or at least to characterize *cunt* (*stupid*, *fat*, *old*, *little*, *daft*, *absolute*, *useless*). Also, for men the word seems to be directed towards specific people (*roy*, *hodgson*, *savage*, *welsh*), when talking to other people (*ya*, *you're*, *you*, *u*), or about other men (*him*). This echoes some previous observations made when analyzing the collocates of *fuck*, as we noticed that men frequently used the word when directly mentioning someone. This observation is even more meaningful when looking at the way the plural form of *cunt* is used by men. Indeed, some of the collocates of *cunts* are *russia* and *wales*, which were both collocates of *fuck*, as well as part of the salient keywords in several cases for men. Figure 9.20 presents the collocational network of the words:

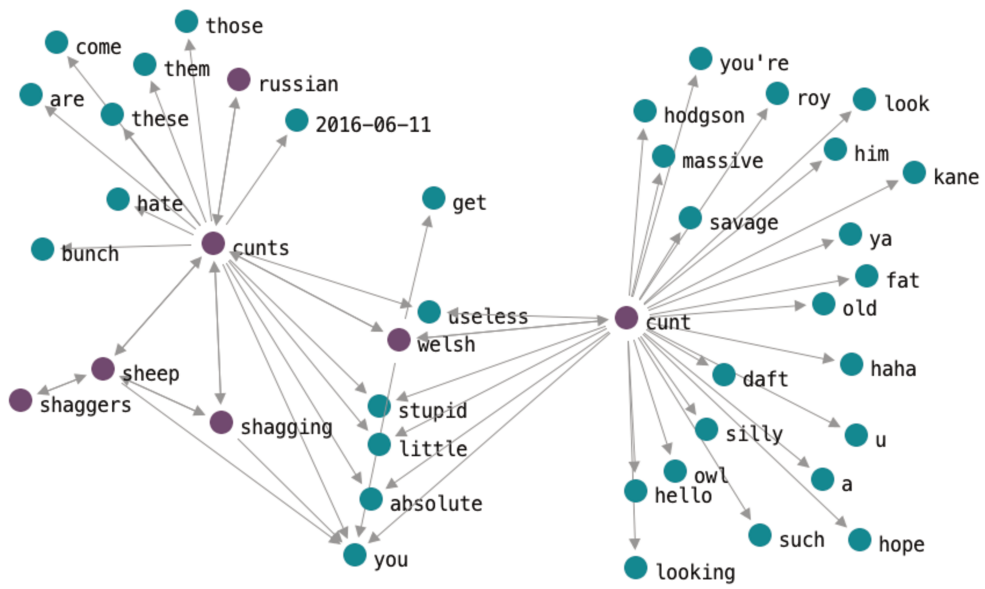


Figure 9.20: Fifth-order collocates for all men in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

As we can see, the singular and plural forms are used in relatively similar, yet different ways. Unsurprisingly, the plural form is used to target groups of people (*those*, *them*, *these*, *russian*, *bunch*), while the singular form is, as mentioned already, used to address specific people. What differs however is that for the singular form, men use more words characterizing the person *cunt* is addressed to (*massive*, *fat*, *old*, *daft*, *silly*). This does not mean that men never add adjectives when they use the plural form, as Figure 9.20 reveals that some of them are shared with the singular form (*useless*, *stupid*, *little*, *absolute*), but this may imply that characterizing groups of people is less important for men than characterizing individuals.

We also notice the presence of a "collocational square" between the words *cunts*, *welsh*, *you* and *sheep*, which in this case are indicative of a set phrase used to denigrate the Welsh. No such link is present for the word *russian*, as the word has no collocate other than *cunts*. This may indicate that there are, in the case of the Welsh, set phrases used to talk about them, whereas the way the Russians are talked about is more heterogeneous, explaining why no other collocate is present here. It could however, be the case that there are set phrases used to mention the Russians too, but that the words used are not frequent enough to be taken into account as additional collocates. In order to better understand this, I looked at the collocates of *russian* in the corpus composed of all male tweets, not just vulgar ones, to see if and how, Russians were talked about as a whole, and not just in vulgar contexts. Figure 9.21 presents the collocates of *cunts* and *russian* in all male tweets:

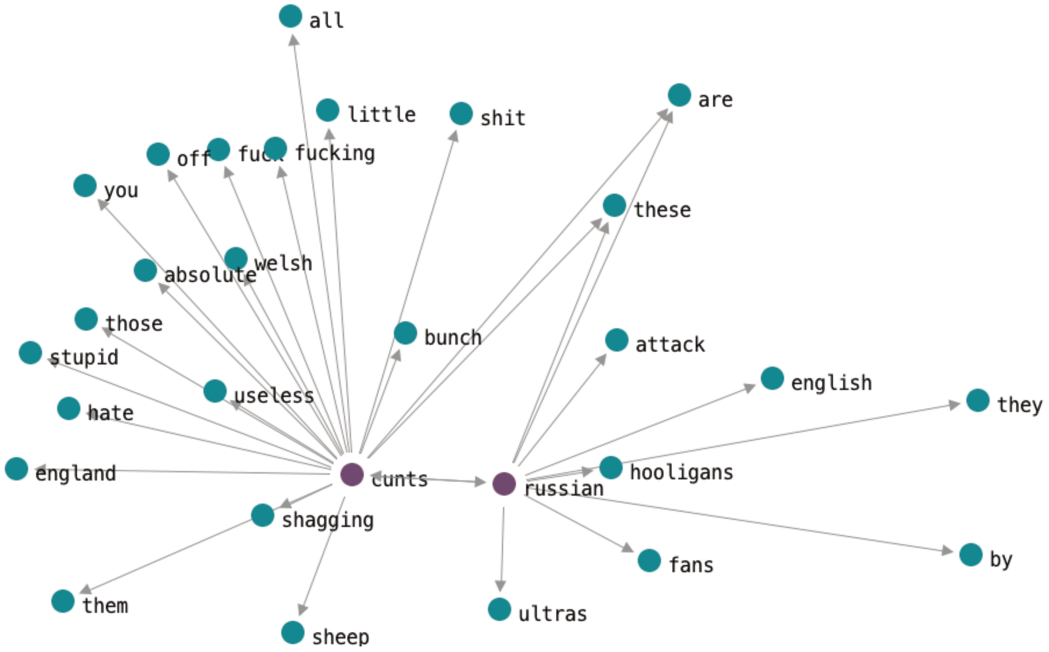


Figure 9.21: Second-order collocates for all men in all tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

Before even looking at the collocates of *russian* in this case, what is important to notice is that the collocates of *cunts* in all tweets are very similar to its collocates in vulgar tweets. In fact, of the 17 collocates of *cunts* in vulgar tweets, 15 are present here among the 21 collocates identified in all male tweets, the six new collocates being *shit*, *fuck*, *fucking*, *england*, *all*, *off*. This similarity indicates that the way people tweet does not differ so much when being vulgar. This once more implies that swearing is not restricted to a limited set of topics, and that Twitter users convey the same ideas when they swear as when they do not. On top of confirming the homogeneity of people's linguistic patterns in vulgar and non-vulgar-only contexts, this

reassures us that the methodology adopted for collocational analysis (i.e. focusing primarily on vulgar tweets) still allows for drawing conclusions which can be generalized to non-vulgar contexts, although we still have to keep in mind that certain swear words may present different, more specific patterns.

Concerning the collocates of *russian* in all tweets now, we observe that contrary to vulgar contexts only, there are here 10 collocates (including *cunts*). Through reading the concordance lines, and the strongest collocates of *russian* being words referring to the (violent) context of football (*hooligans*, *ultras*, *fans*), we also realize that the reason why *russian* is present as a collocate of *cunts* is because of an incident which happened between Russian and English fans[174]. This incident then triggered reactions from some British people leading to these two words collocating in the corpus, as can be seen in the following examples:

(#055) Stay safe out there #Eng fans- real men don't kick someone whilst they're down. Horrible Russian cunts.
(#056) Fuck the Russian cunts *URL*
(#057) Hope we batter the Russian cunts tonight! ENGLAND.

This last example is interesting as it highlights the fact that discussions and exchanges on Twitter are highly influenced by contextual and ephemeral events, which emphasizes the need to pay even more attention to context, and to dig deeper into the data in order to have a clearer idea of the intricacies of certain corpora.

Unsurprisingly, there are very few collocates of *cunts* for women[175]. This is probably related to the fact that women were shown to use the singular form much less frequently than men, so an infrequent use of the plural form may have been expected.

Having reviewed individual swear words, I am now going to take a look at certain topics which have been shown to be preferred by males and females, starting with what can be considered as the most male theme in this corpus, football.

[174] For more details, see https://www.theguardian.com/football/2016/jun/11/euro-2016-french-police-tactics-raise-fears-of-more-clashes-with-england-fans. Last accessed on May, 29th 2017.
[175] The collocates in vulgar tweets are *some*, *are*, *all* for all women, *are* for the 12-18 age group, and *are* and *on* for the 19-30 age group. The collocates are the same for all women in all tweets (*some*, *are* and *all*), with the addition of *fucking*.

3.9.8 About the hashtag *#euro2016*:

We are now going to take a look at what has been considered the topic most representative of male tweets in this corpus so far: football. When comparing keywords for both genders, we noticed the presence of many words related to the topic of football, and *euro2016*, which is part of the hashtag *#euro2016*, has been present as a top keyword every time. This keyword also displayed a much greater absolute frequency than many of the sports-related terms present as keywords in all tweets (see Table 8.5). Thus, it seems that this hashtag is frequently used as a landmark to signal that one is referring to the football competition. Consequently, it seems reasonable to consider this hashtag as the main node from where we will analyze male linguistic patterns when tweeting about football.

Table 9.7 presents all the collocates of *#euro2016* in various sub-groups among vulgar tweets[176], the one in red being common to all groups:

All males	M12-18	M19-30	M31-45
#engwal	#eng	#engwal	#eng
#por		#eng	
#eng		bloody	
#wal			
france			
tournament			
game			
bloody			

Table 9.7: Most salient collocates of *#euro2016* for men according to age in vulgar tweets[177]

Two things are salient in this table. Firstly, there are fewer collocates than in most of the other cases we have analyzed so far despite the greater frequency of the word we are focusing on. Secondly, the vast majority of the collocates of *#euro2016* are other hashtags. The most likely explanation for this phenomenon is that very frequently, hashtags are placed at the end of the tweet. In this case, they are used in combination with other hashtags, as a way to place the tweet

[176] Because the high frequency of *#euro2016* leads to a greater number of tweets taken into account, it was decided to add the 31-45 age group in the analysis, as in this case this sub-group displayed results which were as relevant as the ones from younger age groups of males.

[177] The CPN are: *3b-MI(4), L5-R5, C10-NC1; no filter* for all groups.

in a certain contextual frame, the hashtags serving as landmarks, or keywords, as in the examples below:

(#058) We have trouble leaving out Jamie Vardy, FRANCE LEFT OUT POGBA! #EURO2016 #FRAALB
(#059) Fair play to Iceland - played with pride and wrestled a key point #PORISL #EURO2016 #ICE #Iceland

Or, hashtags can be used to add information which was not explicitly mentioned in the tweet, as in the examples below:

(#060) Not one Spanish player singing there national anthem. #EURO2016 #SpainvsTurkey
(#061) Goaaaaaal #RUS 0-2 #WAL #EURO2016 #EURO

We also notice that Twitter users, on top of placing hashtags at the end of tweets, also tend to use several of them in a row. Thus, in the case of a hashtag placed at the end of a tweet, because the window used in the LancsBox parameters is of five words to the right and five words to the left, the likelihood of focusing on other hashtags is greater than usual, which probably explains why so many of them are considered as top collocates, and especially why there are almost exclusively other hashtags in this list. This is not necessarily a problem however, as the place where users choose to place hashtags is still relevant of a deliberate linguistic choice of the user. Also, although it has now apparently been accepted as standard practice to place hashtags at the end of one's tweet, this is still part of a sociolinguistic norm which has been adopted by the community, and thus is as relevant of a feature as any other linguistic practice.

We are now going to take a closer look at a collocational network having *#euro2016* as the central node for all males in vulgar tweets:

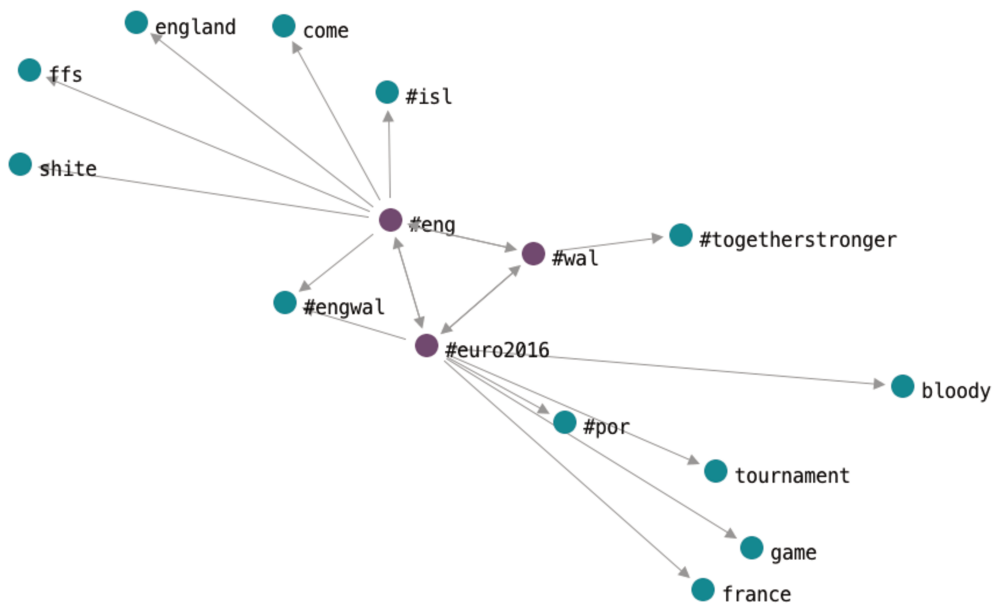


Figure 9.22: Second-order collocates for all men in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

Here, we notice that hashtags mentioning other countries tend to attract each other. This is especially true of the countries which are part of Great Britain[178]. This does not seem surprising considering that this corpus is based on British tweets, as it can be expected that users will frequently tweet about the football team representing their country, or about a football match in which their country will play[179]. What is interesting to notice is that the hashtag *#eng* (for England), has as many hashtags as "regular" words collocating with it, two of these words being swear words (*ffs* and *shite*). This is the sign that even if hashtags are mostly grouped together at the end of the tweets, this does not necessarily determine that other hashtags will always be collocating with them. This also indicates that the English team is the one which seems to be the most significantly associated with swear words.

The other age groups of males display similar patterns as when considered as a whole, and are displayed below except for the 12-18-year-olds for which there is only one collocate of *#euro2016*, which is *#eng*. Figures 9.23 and 9.24 present the collocational networks of *#euro2016* for males aged 19-30 and 31-45 respectively:

[178] Note that none of the other hashtags representing a country had additional collocates.
[179] As it is the case with *#engwal*, which is the hashtag used to tweet about the match between England and Wales.

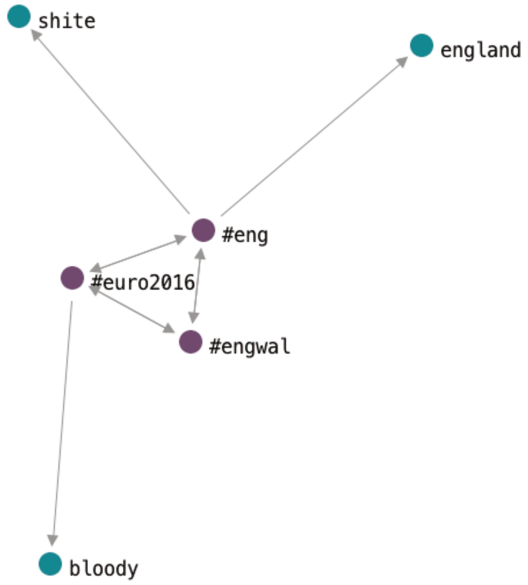


Figure 9.23: Second-order collocates for 19-30 men in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

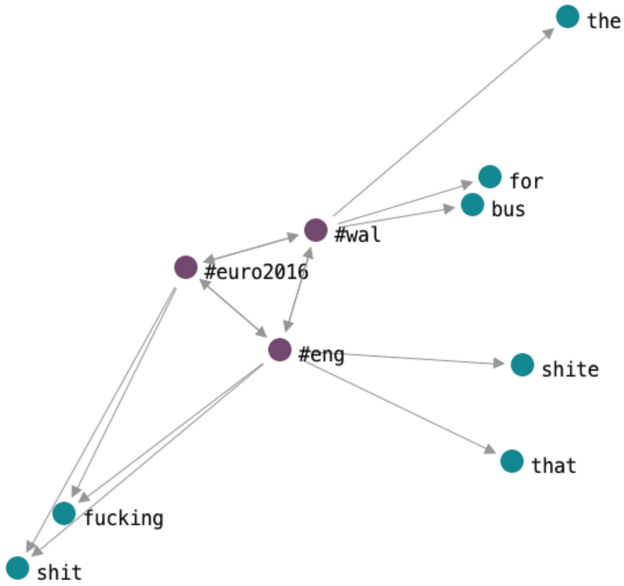


Figure 9.24: Second-order collocates for 31-45 men in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

First, it is interesting to note that in all three cases, the tendency of hashtags mentioning other countries to attract each other is present. Also, for men as a whole as well as for the 19-30 age group, *bloody* is present as a collocate of *#euro2016*. It is surprising that the pattern is considered a salient one in two cases, especially as this swear word has been shown earlier as

being significantly more frequent among women. Finally, all three groups of men use the variant *shite* in the same context; when also using the hashtag *#eng*.

Among older men, on the other hand, *fucking* and *shit* are present as collocates of both *#euro2016* and *#eng*, whereas in the two other groups, the only common links that these two hashtags had were with other hashtags. For men as a whole, and those aged 19-30, there is a partial agreement that *shite* is to be used with *#eng*, and that *bloody* is to be used with *#euro2016*, but men aged 31-45 do not seem to follow this pattern (at least not as markedly), and prefer to use different swear words (or variants, in the case of *shit*(*e*)) when mentioning these two #hashtags. This shows both that males of this age group deviate from this pattern, and that they swear in similar ways whether mentioning one or the other hashtag.

Although shown to mainly be a male feature, it would still be relevant to compare these observations to female patterns. However, no group or sub-group of females showed a single collocate of *#euro2016* with the same parameters. This hashtag thus emerges as more representative of male tweets.

After reviewing this male tendency, it will be interesting to take a look at a female one, so I will now focus on the way females tweet about other females.

3.9.9 About other females:

All along the keyword analysis, we noticed that females generally mentioned other women more than males in their tweets. Although we observed that this attitude may be more noticeable when comparing the results obtained when comparing male and female keywords in vulgar tweets among the 12-18-year-olds (see Table 8.3), this pattern was still present when comparing the 19-30-year-olds (see Table 8.4), as well as when looking at all tweets, and not just vulgar ones (see Table 8.5). Thus, it seems that mentioning other women is a particularly salient female characteristic in any context, whether it is a vulgar one or not. The most frequent word used to mention other women and present as a keyword across all these tables was the word *girls*. I am thus going to look at the collocational network of this word to try to better understand how and why females seem to display a gendered homophily in the way they tweet about other people, and especially how swear words are used in this context. First, Table 9.8

presents the collocates of *girls* according to age, the collocates in red being the ones which are shared by all sub-groups[180]:

All females	F_12-18	F_19-30
bitchy	other	some
other	some	why
hate	hate	are
some	are	do
are	about	with
why	how	so
do	do	
hate	with	
who	why	
how	when	
about	so	
with		
when		
can		
so		
they		

Table 9.8: Most salient collocates of *girls* for females according to age in vulgar tweets[181]

The first thing which is noticeable in this table is the vast amount of collocates which are common to all three groups taken into account. Apart from four words for all women (*bitchy*, *who*, *can*, *they*), every other collocate of *girls* considered here is shared across the sub-groups. Such a tendency has never been observed before in the other collocational analyses carried out so far, and this shows that on top of being a relevant keyword for women, the way females tweet about "girls" in vulgar contexts is very steady across age groups (at least across the ones taken into account here).

The only swear word present here is *bitchy*, which has repeatedly been present as a keyword for females in various contexts (see Chapter 8). *Bitchy* is mainly used to refer to a person's comments or behavior, and is defined as "malicious or spitefully critical[182]". Its presence here,

[180] Note that in this case, contrary to the analysis of *#euro2016* for males, there was no collocate of *girls* to display for the 31-45, hence the column is not included.

[181] The CPN are: *3b-MI(4), L5-R5, C10-NC1; no filter* for all groups.

[182] According to the New Oxford American Dictionary. See http://www.oxfordreference.com/search?q=bitchy&searchBtn=Search&isQuickSearch=true. Last accessed on May, 31st 2017.

as well as the fact that it is the most significant collocate of *girls* when all women are taken into account thus confirms that it is a salient swear word for women. Figure 9.25 presents the collocates of *bitchy* for all women, while having *girls* as the central node:

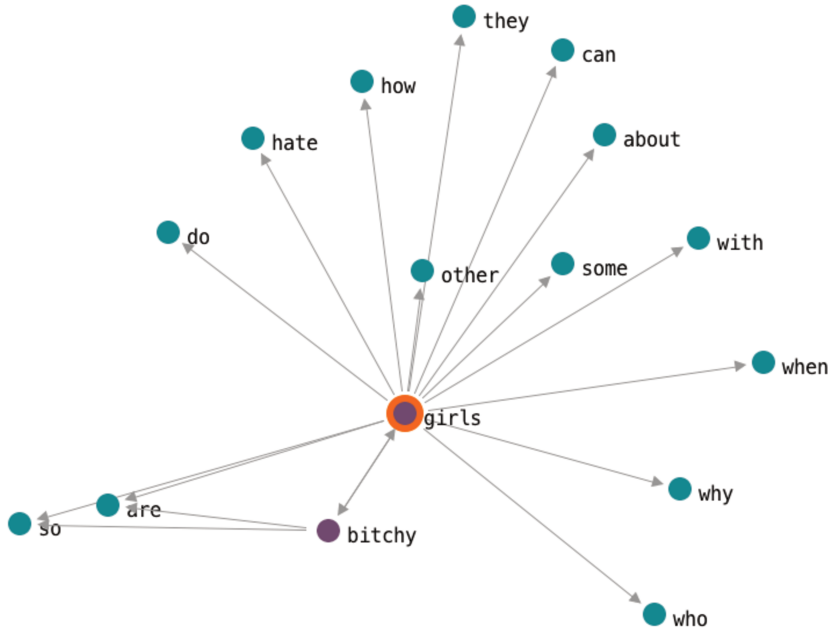


Figure 9.25: Second-order collocates for all women in vulgar tweets: 3b-MI(4), L5-R5, C10-NC1; no filter

As we can see, the most salient collocates of *bitchy* which are shared with girls as well are an emphatic marker (*so*) and the third-person plural form of *to be*, (*are*), used to talk about groups of people. The fact that these three words also all collocate with *girls* seems to indicate that "girls" may be the group of people talked about when mentioning *bitchy*, or at least that this group is mentioned significantly more than any other. Also, this implies that one of the most recurrent contexts in which the word *girls* is mentioned is with *bitchy*. The following examples present a few cases of tweets in which *girls* and *bitchy* collocate:

(#062) OMG I can't deal with the bitchy year 7 girls on my bus 😡😡😡

(#063) Girls aren't bitchy anymore, they're just downright nasty

(#064) Why are some girls so fucking bitchy?..... Absolutely no need

Although *girls* seems to be the most frequent group targeted when using *bitchy*, this does not necessarily imply that males are never qualified as such, as the following examples show:

(#065) He may be beautiful but nah. Can't be dealing with a bitchy boy.

(#066) Bitchy boys are the worst 😴

(#067) Boys are just as bitchy as girls, oh my god 😂😂

As we saw several times now, swear words, despite their somewhat taboo nature, are not always negative, or used to denigrate people, and I gave multiple examples in which swear words were used as a bonding factor, or to show proximity with someone. Here, on the other hand, *bitchy* does not seem to be used in such a positive way. Indeed, the fact that it is the most frequent collocate of *girls*, and the presence of *hate* as another of its collocates indicate that *bitchy girls* may regularly appear with the word *hate*. The examples given above also seem to confirm this, as two users mentioned that they "can't deal" with other people considered as being "bitchy". In another example, *bitchy* is given as a synonym of "nasty", and boys considered as such are referred to as being "the worst".

Note that for men in vulgar contexts, the only collocates of *girls* were *are*, *on*, *some* when taking all men into account, and no collocate of the word was found with the same parameters for men in specific age groups. Also, no collocate of *bitchy* could be found for males in any of the sub-groups usually taken into account, reinforcing the idea that this word is significant for females.

Thus, it seems that in this case, *bitchy* may be preferred by women as a way to denigrate other people, and that most women talk about other girls in the same way, as we saw that the majority of the collocates of *girls* were the same. The fact that, as seen earlier, *bitch* is one of the preferred swear words for females, may explain why *bitchy* is used a lot more by females too. However, the reverse may be true as well, and *bitch* may be used more by women because of their use of *bitchy*. Further analyses would be needed to understand this.

Conclusion:

In this chapter then, we have been able to dig deeper into and better understand most of the observations made in earlier analyses. Collocational networks allowed us to see that swear words were used in set phrases by both women and men, but that these phrases could be similar, gender-specific, or that they could be adapted by each gender according to the general preferences displayed when swearing. These analyses also allowed us to once more confirm that considering swear words as being either positive or negative was very often problematic and a relatively naïve position, as context plays a major role in determining this. Thus, the boundary between the two is often blurry, and it is difficult to categorically make such a distinction. We have also seen that some of the preferred gendered topics noticed earlier (e.g. football for men and emotions for women) were still present when looking at very specific contexts. This then implies that gendered patterns are present both at the macro and at the micro level of linguistic analysis. This means that these patterns, although not categorizing male and female speech as two distinct entities, are still representative of each gender, at least in this corpus.

The second key aspect that this chapter confirmed is that the topics that both genders mention are very similar when the use swear words as when they do not. Again, this is crucial in showing that, at least on Twitter, swearing is not set to be restricted to a certain type of context or discourse, and that users do not differentiate between what they talk about when they swear and when they do not.

The last main finding highlighted in this chapter is what could be considered as a relative predictability in the way swear words are used. It has been shown that depending on parameters such as the gender, the age, and the context, the use of one particular (swear) word seems to influence the presence of other words. These patterns have been spotted thanks to collocational analyses, and seem to be revealing of gendered and generational tendencies.

Conclusion

The original goals of this thesis were twofold: 1) it aimed at providing ways to determine whether younger generations of women swear more (or use "stronger" swear words) than men on social media and in the UK; and 2) it meant to offer new insights into the potential of social media (at least Twitter) in synchronic and diachronic sociolinguistic studies.

Concerning the first point, we can say that, at least according to this sample, younger generations of women do not swear more than men in the UK, nor do they use "stronger" words. We have seen on several occasions that both genders do not differ so much in their use of swear words, be it qualitatively or quantitatively[183], the general tendency being for younger age groups to display fewer gendered differences than the others. Instead of going towards an unbalanced situation where women would outweigh men in their use of swear words then, we seem to be converging towards a more balanced sociolinguistic situation in this regard. Some gendered words and patterns have been highlighted though, the main differences being in some regards limited to the content of what women and men tweet about: men are shown to tweet about sports more frequently than women, and the latter are found to tweet more about emotions and other people. However, here again these results are very often nuanced when looking at intra-generational differences, and the data reveal that age was overall much more significant in determining linguistic patterns than gender alone. From there, the need to pay as much attention to similarities as to differences has once more been made salient in order to fully account for the data present in this - or any - corpus.

Concerning the second point, these results show that, despite the absence of explicit information regarding the gender and the age of Twitter users in their profile, it is possible to overcome this with relatively high precision. Also, this shows that many Twitter users provide enough information in their profile to be able to determine these two parameters. Although these users remain a minority, the high volume of data transiting on this medium is such, that even a fraction of it allows for a corpus representing thousands of users. On top of Twitter being shown as a pool of informants then, this thesis also showed that this social medium (at least) may be an active center for studying the development of linguistic innovations. This is a crucial point

[183] Although it should be reminded that MWU tests confirmed a significant difference of use of swear words between women and men when taken as a whole.

to take into account as neologisms, as well as any emerging linguistic form, are hard to record in sufficient number for systematic analyses to be carried out. This thesis thus highlighted new ways of detecting and analyzing these forms, which may be one way of alleviating this (socio)linguistic issue to a certain extent.

However, throughout this thesis I also pointed to aspects of studying Twitter which may be seen as both advantages and drawbacks. Indeed, one of the most key results of this study is that the importance of context is paramount when analyzing gendered differences, and especially on Twitter. The limited number of characters imposed by the medium (at the moment of this study at least) incites a certain immediacy in the way it is going to be used. Thus, users will very often tweet (and swear) in reaction to something. This can be seen as a definite advantage for linguistic studies, as theoretically we could claim that such an immediacy prevents any form of self-censorship, but this could also be seen as a drawback if one wants to analyze linguistic data in a specific context. This sense of immediacy has been shown to be greatly influenced by the social, geographical, and personal contexts affecting the tweeters, and because of that, these contexts will most likely be ever-changing ones. This might be a problem for researchers hoping to analyze a specific register, or specific contexts of interaction. Also, it could be argued that because of this "instability", the tendencies observed in this thesis are likely to be as unstable. In order to verify this with certainty, we would need to replicate this study to see if we obtain the same results. However, because the results obtained here also match the results of other, fairly recent Twitter-based studies[184], it is likely that the topics present in corpora of tweets are unstable, but that the (gendered) tendencies will remain the same in similar contexts.

In order to go beyond the results presented here, further research may consider comparing the tendencies observed in the UK to other regions. Although one of the main goals of this study was to confirm a tendency which seemed to be salient in the UK, thus justifying the isolation of one geographical region only, looking at other parts of the English-speaking world may yield results allowing for these tendencies to be clarified, or even contrasted. It could, for example, be relevant to see if age has as big an influence elsewhere as it has in the UK in determining the use of swear words. This could be one way to show that age is more relevant than gender alone in this case, thus allowing to have more empirical proof that women's and men's languages are just myths in the UK and/or in other parts of the world.

[184] See for example Wang et al (2014).

Other aspects which may be advised for future studies of the kind are directly related to the methodology adopted to analyze the data. The first of these aspects concerns the type of sub-corpora taken into account for the investigation. I explained earlier that due to technical limitations, I chose to mainly focus on sub-corpora of vulgar tweets only. Although I frequently resorted to the analysis of all tweets, not just vulgar ones, a systematic examination of the former has not been possible. I showed that many times when comparing vulgar and non-vulgar-only tweets, the results of collocational analyses were very similar, thus comforting us in the idea that this choice has not been detrimental to the reliability of the findings. However, I sometimes highlighted differences between the two, which still indicate that systematically comparing the two would be ideal for optimally accounting for all aspects of the corpus.

The second of the methodological choices which may be refined is related to the collocational analysis. As mentioned earlier, Gries (2013) insisted on the importance of taking directionality into account when analyzing collocates, as in any observation, collocate A may collocate with B, but the reverse may not necessarily be true. This is an important aspect to take into account, as this may point to specific uses of the words which may allow for a better representation of certain (gendered) tendencies. In the case of this study, choosing the MI score as an association measure prevented me from carrying out analyses based on the directionality, as, unlike Delta P for example, the MI score does not allow for direction to be examined. Thus, further research may want to include the use of various association measures in order to once more, fully account for every aspect of a corpus.

Bibliography

Andersen, Gisle. 2010. "How to Use Corpus Linguistics in Sociolinguistics." In *The Routledge Handbook of Corpus Linguistics*. Routledge.

Bailey, Lee Ann, and Lenora A. Timm. 1976. "More on Women's—and Men's—expletives." *Anthropological Linguistics* 18 (9): 438–49.

Bailey, Richard. 1985. "South African English Slang: Form, Function and Origins." *South African Journal of Linguistics* 3 (1): 1–42.

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press.

Baker, Paul. 2011. "Times May Change, but We Will Always Have Money: Diachronic Variation in Recent British English." *Journal of English Linguistics* 39 (1): 65–88.

Baker, Paul. 2014. *Using Corpora to Analyze Gender*. A&C Black.

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. "Gender Identity and Lexical Variation in Social Media." *Journal of Sociolinguistics* 18 (2): 135–60.

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

Baudhuin, E. Scott. 1973. "Obscene Language and Evaluative Response: An Empirical Study." *Psychological Reports* 32 (2): 399–402.

Beers Fägersten, Kristy. 2007. "A Sociolinguistic Analysis of Swear Word Offensiveness."

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

Beers Fägersten, Kristy. 2014. "The Use of English Swear Words in Swedish Media."

Biber, Douglas. 1993. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8 (4): 243–57.

Blommaert, Jan, and Chris Bulcaen. 2000. "Critical Discourse Analysis." *Annual Review of Anthropology* 29 (1): 447–66.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2 (1): 1–8.

Bourdieu, Pierre. 1977. *Outline of a Theory of Practice*. Vol. 16. Cambridge university press.

Brezina, Vaclav, Tony McEnery, and Stephen Wattam. 2015. "Collocations in Context: A New Perspective on Collocation Networks." *International Journal of Corpus Linguistics* 20 (2): 139–73.

Brezina, Vaclav, and Miriam Meyerhoff. 2014. "Significant or Random?: A Critical Review of Sociolinguistic Generalisations Based on Large Corpora." *International Journal of Corpus Linguistics* 19 (1).

Burgoon, Michael, Stephen B. Jones, and Diane Stewart. 1975. "Toward a Message-Centered Theory of Persuasion: Three Empirical Investigations of Language Intensity." *Human Communication Research* 1 (3): 240–56.

Chambers, Jack K., and Peter Trudgill. 1980. "Dialectology. Cambridge Textbooks in Linguistics." *Cambridge: Cambridge University Press. E-Kirja, EBSCOhost. Luettu* 1: 2013.

Cheng, Na, Rajarathnam Chandramouli, and K. P. Subbalakshmi. 2011. "Author Gender Identification from Text." *Digital Investigation* 8 (1): 78–88.

Coates, Jennifer. 1986. “Women, Men and Languages: Studies in Language and Linguistics.” *Longmen. London*.

Coates, Jennifer. 1991. *Women Talk: Conversation between Women Friends*. Wiley-Blackwell.

De Klerk, Vivian. 1991. “Expletives: Men Only?” *Communications Monographs* 58 (2): 156–69.

De Klerk, Vivian. 1992. “How Taboo Are Taboo Words for Girls?” *Language in Society* 21 (02): 277–89.

Diakopoulos, Nicholas A., and David A. Shamma. 2010. “Characterizing Debate Performance via Aggregated Twitter Sentiment.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1195–98. ACM.

Dibbell, Julian. 1993. *A Rape in Cyberspace–How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database Into a Society, The Village Voice*.

Dubois, Betty Lou, and Isabel Crouch. 1975. “The Question of Tag Questions in Women’s Speech: They Don’t Really Use More of Them, Do They?” *Language in Society* 4 (03): 289–94.

Dunbar, Robin Ian McDonald. 1989. “Grooming, Gossip and the Evolution of Language (Cambridge, MA, 1996).” *Deborah Cameron, Fiona McAlinden and Kathy O’Leary, “Gossip Revisited: Language in All-Female Groups,” in Women in Their Speech Communities: New Perspectives on Language and Sex*.

Eckert, Penelope. 2008. “Variation and the Indexical Field.” *Journal of Sociolinguistics* 12 (4): 453–76.

Eckert, Penelope, and Sally McConnell-Ginet. 2003. *Language and Gender*. Cambridge University Press.

Edelsky, Carole. 1976. “The Acquisition of Communicative Competence: Recognition of Linguistic Correlates of Sex Roles.” *Merrill-Palmer Quarterly of Behavior and Development* 22 (1): 47–59.

Ervin-Tripp, Susan. 2001. “The Place of Gender in Developmental Pragmatics: Cultural Factors.” *Research on Language and Social Interaction* 34 (1): 131–47.

Freifeld, Clark C., John S. Brownstein, Christopher M. Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. 2014. “Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter.” *Drug Safety* 37 (5): 343–50.

Garside, Roger, Geoffrey N. Leech, and Tony McEnery. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Taylor & Francis.

Gauchat, Louis. 1905. *L’unité Phonétique Dans Le Patois D’une Commune*. Niemeyer.

Goffman, Erving. 1978. “Response Cries.” *Language*, 787–815.

Grimm, M. n.d. “When the Sh*t Hits the Fan.” *American Demographics* 25 (10): 74–75.

Guille, Adrien, and Cécile Favre. 2015. “Event Detection, Tracking, and Visualization in Twitter: A Mention-Anomaly-Based Approach.” *Social Network Analysis and Mining* 5 (1): 18.

Harris, Roy. 1990. “Lars Porsena Revisited.” *The State of the Language*, 411–21.

Haustein, Stefanie, Timothy D. Bowman, Kim Holmberg, Andrew Tsou, Cassidy R. Sugimoto, and Vincent Larivière. 2016. “Tweets as Impact Indicators: Examining the Implications of Automated ‘bot’ Accounts on Twitter.” *Journal of the Association for Information Science and Technology* 67 (1): 232–38.

Herring, Susan. 1999. “Interactional Coherence in CMC.” *Journal of Computer-Mediated Communication* 4 (4): 0–0.

Herring, Susan C. 1992. “Gender and Participation in Computer-Mediated Linguistic Discourse.”

Herring, Susan C., J. Holmes, and M. Meyerhoff. 2003. “The Handbook of Language and Gender.”

Herring, Susan C., and John C. Paolillo. 2006. “Gender and Genre Variation in Weblogs.” *Journal of Sociolinguistics* 10 (4): 439–59.

Herring, Susan C., and Sharon Stoerger. 2014. “Gender and (a) Nonymity in Computer-Mediated Communication.” *The Handbook of Language, Gender, and Sexuality* 2: 567–86.

Hjort, Minna. 2015. “Vittu and Fuck–tales from a Literary Coexistence.” *Miscommunication and Verbal Violence Du Malentendu À La Violence Verbale Misskommunikation Und Verbale Gewalt*, 319.

Holmes, Janet. 1984. “‘Women’s Language’: A Functional Approach.” *General Linguistics* 24 (3): 149.

Huffaker, David A., and Sandra L. Calvert. 2005. “Gender, Identity, and Language Use in Teenage Blogs.” *Journal of Computer-Mediated Communication* 10 (2): 00–00.

Hughes, David John, Moss Rowe, Mark Batey, and Andrew Lee. 2012. “A Tale of Two Sites: Twitter vs. Facebook and the Personality Predictors of Social Media Usage.” *Computers in Human Behavior* 28 (2): 561–69.

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

Hyde, Janet Shibley. 2005. “The Gender Similarities Hypothesis.” *American Psychologist* 60 (6): 581.

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*. John Benjamins Publishing.

Junco, Reynol, Greg Heiberger, and Eric Loken. 2011. “The Effect of Twitter on College Student Engagement and Grades.” *Journal of Computer Assisted Learning* 27 (2): 119–32.

Kergl, Dennis, Robert Roedler, and Sebastian Seeber. 2014. “On the Endogenesis of Twitter’s Spritzer and Gardenhose Sample Streams.” In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, 357–64. IEEE.

Kiesler, Sara, Jane Siegel, and Timothy W. McGuire. 1984. “Social Psychological Aspects of Computer-Mediated Communication.” *American Psychologist* 39 (10): 1123.

Kilgarriff, Adam. 2012. “Getting to Know Your Corpus.” In *Text, Speech and Dialogue*, edited by Petr Sojka, A. Horák, Ivan Kopeček, and Karel Pala, 7499:3–15. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kramarae, Cheris, and H. Jeanie Taylor. 1993. “Women and Men on Electronic Networks: A Conversation or a Monologue.” *Women, Information Technology, and Scholarship*, 52–61.

Kramer, Cheris. 1974. “Stereotypes of Women’s Speech: The Word from Cartoons.” *The Journal of Popular Culture* 8 (3): 624–30.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. “What Is Twitter, a Social Network or a News Media?” In *Proceedings of the 19th International Conference on World Wide Web*, 591–600. ACM.

Laboreiro, Gustavo, and Eugénio Oliveira. 2014. "What We Can Learn from Looking at Profanity." In *International Conference on Computational Processing of the Portuguese Language*, 108–13. Springer.

Labov, William. 1966. "The Effect of Social Mobility on Linguistic Behavior." *Sociological Inquiry* 36 (2): 186–203.

Labov, William. 1972. *The Social Motivation of a Sound Change. Sociolinguistic Patterns, Ed. by William Labov, 1-42*. Philadelphia: University of Pennsylvania Press.

Ladegaard, Hans J. 2004. "Politeness in Young Children's Speech: Context, Peer Group Influence and Pragmatic Competence." *Journal of Pragmatics* 36 (11): 2003–22.

Lakoff, Robin. 1973. "Language and Woman's Place." *Language in Society* 2 (01): 45–79.

Lakoff, Robin Tolmach, and Mary Bucholtz. 2004. *Language and Woman's Place: Text and Commentaries*. Vol. 3. Oxford University Press, USA.

Magué, Jean-Philippe. 2006. "Semantic Changes in Apparent Time." In *32nd Annual Meeting of the Berkeley Linguistics Society*.

McConnell-Ginet, Sally, and G. G. Corbett. 2013. "Gender and Its Relation to Sex: The Myth of 'natural'gender." *The Expression of Gender*, 3–38.

McElhinny, Bonnie. 2003. "Theorizing Gender in Sociolinguistics and Linguistic Anthropology." *The Handbook of Language and Gender*, 21–42.

McEnery, T., and Zhonghua X. 2004. "Swearing in Modern British English: The Case of Fuck in the BNC." *Language and Literature* 13 (3): 235–68.

McEnery, T., and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and practice/Tony McEnery, Andrew Hardie*. Cambridge: Cambridge University Press.

McEnery, T. 2005. *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Routledge.

Mehl, Matthias R., and James W. Pennebaker. 2003. "The Sounds of Social Life: A Psychometric Analysis of Students' Daily Social Environments and Natural Conversations." *Journal of Personality and Social Psychology* 84 (4): 857.

Mercury, Robin-Eliece. 1995. "Swearing: A' Bad' Part of Language; A Good Part of Language Learning." *TESL Canada Journal* 13 (1): 28–36.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. "Understanding the Demographics of Twitter Users." *ICWSM* 11: 5th.

Montagu, Ashley. 1967. *The Anatomy of Swearing*. University of Pennsylvania press.

Murphy, Bróna. 2009. "'She's a Fucking Ticket': The Pragmatics of Fuck in Irish English–an Age and Gender Perspective." *Corpora* 4 (1): 85–106.

Murray, Thomas E. 2012. "Swearing as a Function of Gender in the Language of Midwestern American College Students." *A Cultural Approach to Interpersonal Communication: Essential Readings*, 233–41.

O'Barr, William M., and Bowman K. Atkins. 1980. "' Women's Language' or' Powerless Language'?"

O'Keeffe, Anne, and Michael McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge.

Ochs, Elinor. 1992. "Indexing Gender." *Rethinking Context: Language as an Interactive Phenomenon* 11 (11): 335.

Oliver, Marion M., and Joan Rubin. 1975. "The Use of Expletives by Some American Women." *Anthropological Linguistics*, 191–97.

Otto, Jespersen. 1922. *Language, Its Nature, Development and Origin*. London: Allen and Unwin.

Paquot, Magali, and Yves Bestgen. 2009. "Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction." *Language and Computers Studies in Practical Linguistics* 68: 247.

Poudat, Céline, and Frédéric Landragin. 2017. *Explorer Un Corpus Textuel: Méthodes-Pratiques-Outils*. De Boeck Superieur.

Rathje, Marianne. 2014. "Attitudes to Danish Swearwords and Abusive Terms in Two Generations." *Swearing in the Nordic Countries*, 37–61.

Reppen, Randi. 2010. "Building a Corpus: What Are the Key Considerations?" *The Routledge Handbook of Corpus Linguistics*, 31–37.

Ridge, Enda, and Daniel Kudenko. 2010. "Tuning an Algorithm Using Design of Experiments." *Experimental Methods for the Analysis of Optimization Algorithms*, 265–86.

Risch, Barbara. 1987. "Women's Derogatory Terms for Men: That's Right, 'dirty' Words." *Language in Society* 16 (03): 353–58.

Rosenfeld, Barry, and Steven D. Penrod. 2011. *Research Methods in Forensic Psychology*. John Wiley & Sons.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. 2010. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors." In *Proceedings of the 19th International Conference on World Wide Web*, 851–60. ACM.

Sapir, Edward. 1929. "The Status of Linguistics as a Science." *Language*, 207–14.

Schiffrin, Deborah. 1996. "Narrative as Self-Portrait: Sociolinguistic Constructions of Identity." *Language in Society* 25 (02): 167–203.

Schmid, Hans-Jörg. 2003. *Do Women and Men Really Live in Different Cultures? Evidence from the BNC*.

Scott, Mike. 1997. "PC Analysis of Key Words—and Key Key Words." *System* 25 (2): 233–45.

Selfe, Cynthia L., and Paul R. Meyer. 1991. "Testing Claims for on-Line Conferences." *Written Communication* 8 (2): 163–92.

Selnow, Gary W. 1985. "Sex Differences in Uses and Perceptions of Profanity." *Sex Roles* 12 (3): 303–12.

Sewell Jr, William H. 1992. "A Theory of Structure: Duality, Agency, and Transformation." *American Journal of Sociology* 98 (1): 1–29.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

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.

Sood, Sara, Judd Antin, and Elizabeth Churchill. 2012. "Profanity Use in Online Communities." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1481–90. ACM.

Staley, Constance M. 1978. "Male-Female Use of Expletives: A Heck of a Difference in Expectations." *Anthropological Linguistics* 20 (8): 367–80.

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

Steadman, J. M. 1935. "A Study of Verbal Taboos." *American Speech* 10 (2): 93–103.

Stroh-Wollin, Ulla. 2010. *Fula Ord-Eller?: En Enkät Om Attityder till Svordomar Och Andra Fula Ord*. Uppsala universitet.

Subrahmanyam, Kaveri, David Smahel, and Patricia Greenfield. 2006. "Connecting Developmental Constructions to the Internet: Identity Presentation and Sexual Exploration in Online Teen Chat Rooms." *Developmental Psychology* 42 (3): 395.

Sutton, Laurel A. 1995. *Bitches and Skanky Hobags: The Place of Women in Contemporary Slang. Gender Articulated: Language and the Socially Constructed Self. Ed. Kira Hall and Mary Bucholtz*. New York: Routledge.

Svensson, Annika. 2004. "Gender Differences in Swearing; Who The**** Cares?" *A Study of Men and Women's Use of Swearwords in Informal and Formal Situation*.

Tagliamonte, Sali A., and Derek Denis. 2008. "Linguistic Ruin? LOL! Instant Messaging and Teen Language." *American Speech* 83 (1): 3–34.

Thelwall, Mike. 2008. "Fk Yea I Swear: Cursing and Gender in MySpace." *Corpora* 3 (1): 83–107.

Thomson, Rob, and Tamar Murachver. 2001. "Predicting Gender from Electronic Discourse." *British Journal of Social Psychology* 40 (2): 193–208.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Vol. 6. John Benjamins Publishing.

Trudgill, Peter, and Lars-Gunnar Andersson. 1990. *Bad Language*. Cambridge, USA: Penguin books.

Valenzuela, Sebastián, Daniel Halpern, and James E. Katz. 2014. "Social Network Sites, Marriage Well-Being and Divorce: Survey and State-Level Evidence from the United States." *Computers in Human Behavior* 36: 94–101.

Valkanas, George, and Dimitrios Gunopulos. 2012. "Location Extraction from Social Networks with Commodity Software and Online Data." In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, 827–34. IEEE.

Villessèche, Julie. 2016. "The Board and the Commission (1909-Present): Study of a Language Criterion through Film Classification". Université de Lyon.

Vingerhoets, Ad JJM, Lauren M. Bylsma, and Cornelis De Vlam. 2013. "Swearing: A Biopsychosocial Perspective." *Psihologijske Teme* 22 (2): 287–304.

Vives, Juan Luis. 2007. *The Education of a Christian Woman: A Sixteenth-Century Manual*. University of Chicago Press.

Wang, Wenbo, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. "Cursing in English on Twitter." In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 415–25. ACM.

Weiss, Gilbert, and Ruth Wodak, eds. 2003. *Critical Discourse Analysis*. London: Palgrave Macmillan UK.

Wentworth, Harold, and Stuart Berg Flexner. n.d. *Dictionary of American Slang. 1975. 2d Supp. Ed*. New York: Crowell.

Williamsson, Joy. 2009. *How Brits Swear: The Use of Swearwords in Modern British English*.

Yates, Andrew, Alek Kolcz, Nazli Goharian, and Ophir Frieder. 2016. "Effects of Sampling on Twitter Trend Detection." In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'16)*.

Zahn, Christopher J. 1989. "The Bases for Differing Evaluations of Male and Female Speech: Evidence from Ratings of Transcribed Conversation." *Communications Monographs* 56 (1): 59–74.

Zappavigna, Michele. 2011. “Ambient Affiliation: A Linguistic Perspective on Twitter.” *New Media & Society* 13 (5): 788–806.

Appendices

Appendix 1: Comparison of gendered keywords for users aged 12-18 in all tweets

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
12-18 females						12-18 males					
mistress	571	297.1	3	2	3.9	euro2016	766	522.3	114	59.3	3.9
woof	399	207.6	3	2	3	league	438	298.7	41	21.3	3.3
xxx	1049	545.9	172	117.3	3	team	953	649.8	266	138.4	3.1
n	2108	1097	452	308.2	2.9	game	1459	994.8	481	250.3	3.1
pet	415	216	15	10.2	2.9	player	445	303.4	56	29.1	3.1
darling	445	231.6	26	17.7	2.8	players	414	282.3	57	29.7	2.9
puppy	392	204	21	14.3	2.7	eng	564	384.6	127	66.1	2.9
makeup	457	237.8	40	27.3	2.7	wales	711	484.8	215	111.9	2.8
xx	1776	924.2	429	292.5	2.6	england	1547	1054.8	613	319	2.8
cute	902	469.4	175	119.3	2.6	goal	570	388.7	150	78.1	2.7
excited	1494	777.5	365	248.9	2.5	nonces	265	180.7	5	2.6	2.7
loveisland	678	352.8	131	89.3	2.4	tournament	319	217.5	33	17.2	2.7
u	5081	2644.1	1562	1065.1	2.4	class	548	373.7	148	77	2.7
mum	1401	729.1	370	252.3	2.4	fans	572	390	166	86.4	2.6
prom	732	380.9	157	107.1	2.3	white	635	433	201	104.6	2.6
crying	618	321.6	122	83.2	2.3	play	954	650.5	368	191.5	2.6
hair	900	468.4	216	147.3	2.3	mate	1341	914.4	570	296.6	2.6
ur	1411	734.3	386	263.2	2.3	iceland	360	245.5	68	35.4	2.6
omg	1500	780.6	418	285	2.3	payet	263	179.3	20	10.4	2.5
tidevip	241	125.4	0	0	2.3	alura	208	141.8	0	0	2.4
lovely	953	495.9	250	170.5	2.2	france	421	287.1	125	65	2.3
love	7423	3862.9	2522	1719.7	2.2	kane	256	174.6	35	18.2	2.3
girls	973	506.3	262	178.6	2.2	vardy	308	210	65	33.8	2.3
beautiful	1054	548.5	297	202.5	2.1	roy	241	164.3	34	17.7	2.2
miss	1558	810.8	477	325.2	2.1	score	287	195.7	63	32.8	2.2
fab	395	205.6	64	43.6	2.1	bro	287	195.7	64	33.3	2.2
boyfriend	308	160.3	35	23.9	2.1	ronaldo	319	217.5	84	43.7	2.2
oitnb	343	178.5	51	34.8	2.1	logo	186	126.8	6	3.1	2.2
xxxx	353	183.7	58	39.5	2	against	382	260.5	123	64	2.2
literally	1518	790	501	341.6	2	imo	200	136.4	16	8.3	2.2

Appendix 2: Comparison of gendered keywords for users aged 19-30 in all tweets

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
19-30 females						19-30 males					
loveisland	1599	480.8	409	104.1	2.8	euro2016	3439	875	650	195.4	3.3
excited	2449	736.3	856	217.8	2.6	eng	1975	502.5	300	90.2	3.2
hair	1442	433.5	464	118.1	2.4	mate	3670	933.8	815	245	3
omg	1735	521.6	618	157.2	2.4	players	1228	312.5	142	42.7	2.9
lbloggers	435	130.8	0	0	2.3	player	1259	320.3	165	49.6	2.8
cute	1260	378.8	423	107.6	2.3	tournament	1004	255.5	102	30.7	2.7
makeup	497	149.4	46	11.7	2.2	labour	1055	268.4	143	43	2.6
girls	1332	400.5	491	124.9	2.2	team	2381	605.8	579	174.1	2.6
mum	1602	481.7	635	161.6	2.2	league	874	222.4	93	28	2.5
boyfriend	615	184.9	118	30	2.2	goal	1441	366.7	294	88.4	2.5
xxx	932	280.2	295	75.1	2.2	corbyn	910	231.5	123	37	2.4
fab	610	183.4	144	36.6	2.1	game	4168	1060.5	1270	381.8	2.4
my	38426	11553.1	21888	5569.3	2.1	england	4051	1030.8	1235	371.3	2.4
literally	2416	726.4	1206	306.9	2	pogba	540	137.4	17	5.1	2.3
oitnb	564	169.6	130	33.1	2	iceland	1062	270.2	214	64.3	2.3
holiday	1838	552.6	878	223.4	2	fans	1506	383.2	382	114.9	2.2
xx	1591	478.3	739	188	2	against	1132	288	258	77.6	2.2
girl	1493	448.9	691	175.8	2	wal	945	240.5	209	62.8	2.1
lovely	1563	469.9	742	188.8	2	ronaldo	955	243	214	64.3	2.1
sleep	2460	739.6	1308	332.8	1.9	score	759	193.1	135	40.6	2.1
bed	2269	682.2	1197	304.6	1.9	sterling	620	157.8	80	24.1	2.1
nails	388	116.7	49	12.5	1.9	wales	2099	534.1	685	206	2.1
thank	2966	891.8	1637	416.5	1.9	win	2769	704.6	966	290.4	2.1
baby	870	261.6	349	88.8	1.9	payet	520	132.3	43	12.9	2.1
dress	476	143.1	108	27.5	1.9	cheers	938	238.7	216	64.9	2.1
love	9283	2791	5579	1419.6	1.9	play	2227	566.7	748	224.9	2.1
so	25906	7788.8	16202	4122.5	1.9	signing	604	153.7	79	23.8	2
am	4743	1426	2818	717	1.9	games	915	232.8	212	63.7	2
miss	1891	568.5	1016	258.5	1.9	bro	687	174.8	121	36.4	2
cry	711	213.8	270	68.7	1.9	vardy	662	168.4	120	36.1	2

Appendix 3: Comparison of gendered keywords for users aged 31-45 in all tweets

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
31-45 females						31-45 males					
ppl	385	1128.4	16	33.4	9.2	followed	372	775.7	13	38.1	6.3
abt	280	820.7	1	2.1	9	n	376	784.1	24	70.3	5.2
thrilled	263	770.8	3	6.3	8.2	eng	304	633.9	19	55.7	4.7
xxx	569	1667.7	96	200.2	5.9	game	370	771.6	41	120.2	4
count	401	1175.3	61	127.2	5.6	euro2016	378	788.3	45	131.9	3.8
islam	153	448.4	2	4.2	5.3	fucking	310	646.5	33	96.7	3.8
women	300	879.3	49	102.2	4.8	players	169	352.4	7	20.5	3.8
chance	569	1667.7	139	289.9	4.5	ha	422	880	58	170	3.6
muslims	133	389.8	7	14.6	4.3	wal	146	304.5	6	17.6	3.4
bhlove	111	325.3	0	0	4.3	mate	572	1192.8	98	287.2	3.3
religion	134	392.7	9	18.8	4.1	brexit	227	473.4	27	79.1	3.2
fingerscrossed	107	313.6	2	4.2	4	manchester	176	367	17	49.8	3.1
shld	93	272.6	0	0	3.7	team	232	483.8	33	96.7	3
coatbridge	91	266.7	0	0	3.7	tae	102	212.7	2	5.9	3
giveaway	221	647.7	51	106.4	3.6	player	141	294	13	38.1	2.9
religious	92	269.6	1	2.1	3.6	vinyl	87	181.4	0	0	2.8
lanarkshire	93	272.6	5	10.4	3.4	f	118	246.1	8	23.4	2.8
trans	91	266.7	5	10.4	3.3	m1	80	166.8	1	2.9	2.6
thanks	1117	3273.9	444	925.9	3.3	play	218	454.6	40	117.2	2.6
rights	115	337.1	16	33.4	3.3	xo	78	162.7	1	2.9	2.6
men	194	568.6	50	104.3	3.3	league	88	183.5	4	11.7	2.5
retweeting	87	255	5	10.4	3.2	wales	234	488	45	131.9	2.5
gender	80	234.5	2	4.2	3.2	fans	166	346.2	26	76.2	2.5
plz	82	240.3	3	6.3	3.2	france	115	239.8	12	35.2	2.5
thank	472	1383.4	175	364.9	3.2	england	378	788.3	87	255	2.5
criticise	76	222.8	2	4.2	3.1	fuck	239	498.4	48	140.7	2.5
happily	87	255	8	16.7	3	tournament	88	183.5	5	14.7	2.5
dundee	74	216.9	2	4.2	3	voteremain	83	173.1	4	11.7	2.4
human	112	328.3	20	41.7	3	shite	79	164.7	3	8.8	2.4
liberal	68	199.3	1	2.1	2.9	shit	229	477.5	47	137.8	2.4

Appendix 4: Comparison of gendered keywords for users aged 46-60 in all tweets

Keyword	AF	NF	AF_rc	NF_rc	Score	Keyword	AF	NF	AF_rc	NF_rc	Score
46-60 females						46-60 males					
fab	240	2587.7	11	40.3	19.2	mm	1090	3994.5	0	0	40.9
joseph	180	1940.7	4	14.7	17.8	temperature	1089	3990.8	0	0	40.9
xxxx	141	1520.2	2	7.3	15.1	barometer	1089	3990.8	0	0	40.9
xxx	154	1660.4	12	44	12.2	mph	1088	3987.2	0	0	40.9
joe	229	2469.1	33	120.9	11.6	mb	1088	3987.2	0	0	40.9
completed	102	1099.8	1	3.7	11.6	humidity	1088	3987.2	0	0	40.9
puzzle	96	1035.1	0	0	11.4	wind	1094	4009.1	2	21.6	33.8
jigsawepic	96	1035.1	0	0	11.4	c	1129	4137.4	12	129.4	18.5
jigsaw	96	1035.1	0	0	11.4	steady	466	1707.7	1	10.8	16.3
puzzles	97	1045.8	1	3.7	11.1	slowly	509	1865.3	2	21.6	16.2
xxxxx	87	938	0	0	10.4	rain	1136	4163.1	17	183.3	15
epic	100	1078.2	5	18.3	10	rising	311	1139.7	0	0	12.4
ya	145	1563.4	28	102.6	8.2	falling	331	1213	3	32.3	9.9
xx	273	2943.5	87	318.8	7.3	wsw	192	703.6	0	0	8
yes	397	4280.4	149	546	6.8	sw	145	531.4	0	0	6.3
weymouth	42	452.8	0	0	5.5	w	221	809.9	6	64.7	5.5
prize	52	560.7	6	22	5.4	wnw	121	443.4	0	0	5.4
brighton	46	496	6	22	4.9	mate	182	667	6	64.7	4.7
goodnight	43	463.6	5	18.3	4.8	follow	284	1040.8	15	161.7	4.4
sweet	54	582.2	15	55	4.4	arlesey	81	296.8	0	0	4
yum	36	388.1	4	14.7	4.3	station	116	425.1	3	32.3	4
sleep	70	754.7	28	102.6	4.2	game	285	1044.4	18	194.1	3.9
hayley	31	334.2	2	7.3	4	rt	111	406.8	3	32.3	3.8
omg	35	377.4	5	18.3	4	today	1507	5522.7	134	1444.8	3.6
ss	26	280.3	0	0	3.8	season	146	535	7	75.5	3.6
ch	26	280.3	1	3.7	3.7	railway	71	260.2	0	0	3.6
cu	24	258.8	0	0	3.6	ssw	70	256.5	0	0	3.6
awesome	48	517.5	21	77	3.5	greater	88	322.5	2	21.6	3.5
silverstone	26	280.3	3	11	3.4	nw	66	241.9	0	0	3.4
xxxxxx	22	237.2	0	0	3.4	eng	106	388.5	4	43.1	3.4

Résumé en français

Introduction

Le genre (naturel) est présent à tous les niveaux de notre société et définit un ensemble de règles déterminant ce que sont la masculinité et la féminité. Parmi les traits traditionnellement associés au genre d'une personne, la manière dont les individus utilisent le langage est un facteur tout aussi important que d'autres aspects, potentiellement plus évidents, tels que la manière de s'habiller par exemple. Ces idées préconçues influencent de manière plus ou moins forte les attentes que nous pouvons avoir des autres. La littérature recense de nombreux traits linguistiques faisant partie de ces idées associées à un genre ou l'autre tels que la tournure interrogative (tag question) ou la déférence par exemple. Comme je le montre en détails plus loin (voir Chapitres 1, 2 et 3), la plupart de ces idées préconçues ont été réfutées, et certaines sont encore discutées. De toutes les caractéristiques linguistiques genrées, celle qui a probablement été le plus débattue est celle concernant l'utilisation des jurons. En effet, la vulgarité est un sujet qui est généralement une source de tensions et de débats, que la question du genre soit abordée ou non. Ceci est principalement dû à des idées qui ont longtemps été associées à l'utilisation de la vulgarité : celles d'un langage « impur », blasphématoire ou encore d'un langage tabou. En anglais, l'expression « bad language » est souvent employée, ce qui sous-entend que cette façon de parler n'est pas « la bonne », et n'est donc pas désirable. Cette expression, au-delà de faire référence à l'emploi d'insultes, englobe aussi parfois tout un ensemble de traits linguistiques considérés comme devant être évités (« bad »), comme l'argot, le jargon, l'utilisation d'une grammaire non-standard, les formes dialectales ou même les néologismes. A cause d'une interaction complexe entre pression et pouvoir social, la vulgarité a traditionnellement été associée à l'idée de masculinité avant tout (voir Chapitre 1). Coates (1986 : 97) explique que la croyance populaire selon laquelle les hommes utilisent des jurons et termes tabous plus souvent que les femmes est très répandue[185], et il avance que ceci mène à la création de stéréotypes stigmatisants pour les femmes utilisant ce registre linguistique. De plus, utiliser des jurons est souvent considéré comme étant l'affirmation linguistique d'une forme de pouvoir social (Lakoff, 1973 ; G. Hughes, 2006 ; Beers Fägersten, 2012 ; Murray, 2012). Par conséquent, l'association intrinsèque de la vulgarité comme forme de pouvoir à un

[185] Citation originale: "the folklinguistic belief that men swear more than women and use more taboo words is widespread".

genre ou l'autre pourrait conduire à l'association d'autres caractéristiques sociales aux questions de masculinité ou de féminité, qu'elles soient fondées ou non. Certaines études ont démontré que, contrairement aux idées longtemps répandues, les femmes n'utilisent pas la vulgarité moins souvent que les hommes, pas plus qu'elles n'utilisent un registre linguistique fondamentalement différent (voir Chapitre 1).

En effet, ces enquêtes ont montré que ce qui diffère généralement chez les hommes et les femmes dans leur manière d'utiliser la vulgarité n'est pas la fréquence à laquelle ils/elles jurent, mais le contexte dans lesquels les jurons sont utilisés, ainsi que le type de juron utilisés. Certaines études ont prédit que l'utilisation de jurons dits « forts » (i.e. « strong swear words ») chez les femmes augmenterait dans certains contextes (Murray, 2012), et en particulier sur les réseaux sociaux (Thelwall, 2008) ; ceci s'appliquerait particulièrement aux jeunes générations de femmes (les 16-19 ans dans le cas the Thelwall) au Royaume Uni. Il a même été anticipé qu'une égalité linguistique, voire un inversement des tendances genrées pour l'utilisation des jurons « forts », deviendraient de plus en plus courant, au moins sur les réseaux sociaux[186]. En d'autres termes, l'utilisation de certains jurons chez ces jeunes générations de femmes deviendrait à terme plus fréquente que celle des hommes du même âge. Etant donné que les adolescents sont souvent présentés comme étant à l'avant-garde de l'innovation linguistique, l'hypothèse de Thelwall suggère que cette préférence linguistique pourrait s'appliquer à d'autres générations de femmes, et pas seulement les plus jeunes. Cette hypothèse correspond à une des conclusions de Herring (2003), qui suggère que la communication assistée par ordinateur participe à l'autonomisation des femmes (voir Chapitre 2).

Il serait donc intéressant de vérifier ces observations et prédictions près de dix ans après qu'elles aient été formulées afin de les confirmer ou de les réfuter. A ma connaissance, aucun travail de ce type n'a été effectué en prenant en compte suffisamment de paramètres sociaux et démographiques permettant de vérifier l'existence des phénomènes susnommés. Par conséquent, la question suivante se pose : les prévisions faites par Thelwall en 2008 se sont-elles réalisées près de dix ans plus tard, dans une société où les médias sociaux n'ont jamais eu autant d'importance dans notre vie quotidienne ? Le but de cette thèse est donc double : tout d'abord elle vise à offrir une meilleure compréhension de la manière dont les femmes et les hommes utilisent la vulgarité sur les réseaux sociaux. Le second objectif de ce travail est de démontrer le potentiel de ce type de médias en tant qu'objet d'étude sociolinguistique synchronique (et potentiellement diachronique) de grande échelle.

[186] Citation originale de Thelwall (2008 : 102): "gender equality in swearing or a reversal in gender patterns for strong swearing, will slowly become more widespread, at least in social network sites."

Cependant, répliquer les études précédentes (e.g. Thelwall, 2008) ne représente pas une solution idéale car MySpace, le réseau social sur lequel Thelwall avait basé son étude, a souffert d'une baisse considérable de sa fréquentation depuis 2008. De ce fait, et pour d'autres raisons méthodologiques décrites plus loin (voir Chapitres 2, 4 et 6), j'ai choisi Twitter comme mode de collecte de données. Avec près d'un demi milliard de tweets émis chaque jour (au moment de cette étude) à travers le monde, Twitter est l'un des réseaux sociaux les plus populaires. Cette étude est basée sur un corpus composé à l'origine[187] d'un peu plus de dix-huit millions de tweets représentatifs de près de 739 000 utilisateurs (voir Chapitre 6). Le corpus a été constitué à partir de tweets provenant du Royaume Unis, émis par des utilisateurs masculins et féminins, et qu'appartenant à différentes tranches d'âge. La focalisation géographique permet de comparer ces données avec celles d'études antérieures se concentrant également sur la même région. Une méthodologie et des outils d'analyse issus de la linguistique dite de corpus ont été utilisés pour mener à bien ce projet et tenter de répondre aux problématiques soulevées précédemment (voir Chapitres 7, 8 et 9). Aussi, en raison du manque d'informations démographiques directement associées aux profils Twitter (e.g. le genre et l'âge des utilisateurs), il fût nécessaire de recourir à des techniques issues de la programmation informatique afin d'inférer le genre et l'âge de ces personnes (voir Chapitre 6).

Analyser des changements linguistique grâce aux réseaux sociaux est une approche relativement nouvelle, surtout si l'on compare l'impact qu'ont les réseaux sociaux actuellement par rapport à la place beaucoup plus limitée qu'ils avaient lorsque ces prévisions furent effectuées. Selon une étude d'Ofcom (voir leur rapport de 2013[188]), le temps consacré aux réseaux sociaux augmente chaque année chez des utilisateurs de toute catégorie socio-économique et de toute tranche d'âge (voir aussi Smith et Brewer, 2012). Cette thèse entend donc améliorer les connaissances que nous avons du lien qu'il existe entre le genre, l'âge, l'utilisation de la vulgarité et les réseaux sociaux. En ce faisant, le but est également d'améliorer le socle méthodologique actuel et nécessaire à l'analyse linguistique des réseaux sociaux.

Pour accomplir ces objectifs, cette thèse se divise en trois parties, chacune se focalisant sur un des piliers de cette étude, à savoir un état de la littérature, la description de la méthodologie utilisée, et la présentation des résultats. Ces trois parties sont, elles aussi, composées de trois chapitres. Le plan de cette thèse est le suivant :

[187] C'est à dire avant que l'âge et le genre des utilisateurs n'aient été détectés, comme nous le verrons plus tard.

[188] Dernière consultation le 27 juin 2017. URL: https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/children-parents-oct-2013

Première partie

Le but du chapitre 1 est de présenter les notions principales sur lesquelles cette étude repose : les questions de genre, de vulgarité et de réseaux sociaux. Afin d'insister particulièrement sur le côté interdisciplinaire de cette thèse, je m'attarderai ici principalement sur les relations qui existent entre ces notions.

Le chapitre 2 a pour but principal de déconstruire les idées préconçues existant à propos de la manière dont les hommes et les femmes utilisent le langage.

Le chapitre 3 est une opportunité d'effectuer un passage en revue approfondi des fondations sur lesquelles repose la linguistique dite « de corpus ». Cette méthodologie sera centrale lors de l'analyse des données, il paraît donc capital d'en aborder les concepts clés afin de mieux comprendre les choix méthodologiques qui ont été faits.

Deuxième partie

Le chapitre 4 présente les avantages des données Twitter par rapport à d'autres types de réseaux sociaux. Ce chapitre sera aussi une opportunité de fournir plus de détails quant à la manière dont l'interface de Twitter fonctionne, car cet aspect est fondamental pour appréhender correctement la manière dont j'ai eu accès à ce corpus.

Dans le chapitre 5 je fournirai des explications approfondies concernant l'outil en ligne (i.e. CATS) que j'ai utilisé pour collecter les données. Cette interface étant un élément central de cette étude, il est nécessaire de comprendre la manière dont je l'ai utilisée afin de mettre en lumière les avantages, mais aussi les potentiels inconvénients de la méthodologie employée.

Le chapitre 6 donne plus de détails sur la manière dont j'ai inféré l'âge et le genre des utilisateurs dont j'ai collecté les tweets. Ces deux informations n'étant pas ouvertement accessibles, j'ai dû recourir à des méthodes informatiques et statistiques que je présenterai dans ce chapitre.

Troisième partie

Le chapitre 7 fournit une vue d'ensemble des données collectées. Ces informations varient d'éléments basiques tels que le nombre d'utilisateurs à l'intérieur de chaque tranche d'âge, à des données plus détaillées telles que le nombre d'occurrences de chaque juron au sein des différentes tranches d'âge prises en compte, ou la manière dont le corpus est organisé en différents sous-corpus.

Le chapitre 8 va plus en détails dans la présentation des données, et fournit des informations telles que les jurons qui se sont révélés comme étant significativement plus utilisés par les

hommes ou par les femmes, ainsi que par certaines tranches d'âge. Pour ce faire, des tests statistiques tels que le Mann-Whitney U ou le simple maths parameter ont été utilisés. Ce chapitre est une opportunité d'analyser les différences qui existent entre chaque sous-groupe d'utilisateur, mais ne manquera pas de présenter les éléments se révélant similaires également. Le chapitre 9 combine des analyses à la fois quantitatives et qualitatives afin de concentrer l'analyse sur des cas spécifiques qui sont représentatifs des tendances mises en lumière lors des analyses précédentes. Ce sera une manière de confirmer ou de réfuter les tendances observées jusqu'alors, et de mieux en comprendre l'organisation. Dans ce chapitre, l'exploration des données sera principalement articulée autour de l'analyse des cooccurrences rendue possible par le logiciel LancsBox.

PREMIERE PARTIE : CADRE THEORIQUE

Chapitre 1 : Concepts généraux

Très souvent, une insulte est considérée comme étant un mot offensant, utilisé la plupart du temps pour exprimer un sentiment de colère ou de douleur. Si l'on garde à l'esprit qu'un juron est un mot à caractère offensant, il n'est pas difficile d'imaginer des foules de contextes où des mots ou expressions usuels pourraient offenser un interlocuteur. Inversement, il peut également exister des contextes où un juron ne sera pas considéré comme tel, voire sera considéré comme neutre dans certaines communautés de pratique. Beers Fägersten (2007 : 32) insiste particulièrement sur cette notion de contexte afin de déterminer si le mot en question sera perçu comme étant offensant.

Le contexte est donc un élément primordial à prendre en compte afin d'établir les mots et expressions considérés comme étant des insultes. C'est pour cette raison que des institutions telles que la BBC ont mis au point des listes de mots qui, selon eux, doivent être censurés dans leurs programmes à certaines heures de diffusion. Ceci entraine donc une standardisation des mots considérés comme étant vulgaires dans un contexte audiovisuel au Royaume Uni. Différentes situations nécessitent donc différentes approches, et ce qui s'applique dans un contexte ne s'applique pas forcément dans un autre. C'est pour cette raison qu'il est difficile d'établir une liste clairement définie de termes étant considérés comme des insultes dans le cadre d'une étude sociolinguistique. Ainsi, chaque groupe ou institution constitue des listes qui leurs sont propres et qui s'appliquent dans leurs cadres respectifs.

Dans le cadre d'une étude sociolinguistique portant principalement sur les différences genrées donc, il est nécessaire de prendre en compte ce qui fut rapporté concernant ces différences dans des études antérieures. Ervin-Tripp (2001 : 135) parle de « mondes séparés » (« Separate Worlds ») pour illustrer la manière dont les hommes et les femmes utilisent le langage. Lakoff quant à elle (1973) parle de différentes langues. Selon ces théories donc, il existerait une « langue masculine » et une « langue féminine » qui se développeraient chez les individus dès l'enfance et ce, à cause des différences sociales établies entre les filles et les garçons dès le plus jeune âge. La vulgarité fait partie de ces éléments de la langue qui ont longtemps été considérés comme étant des traits masculins uniquement, et des manuels décrivant la manière « correcte » de parler pour les femmes existaient déjà plusieurs siècles auparavant. Toutefois, la vulgarité

étant une affirmation linguistique de pouvoir social avant tout (Lakoff, 1973 ; Hall, 2004), le statut généralement attribué aux hommes et aux femmes sera un facteur plus déterminant de la légitimité accordée à un individu étant vulgaire, que le contexte d'élocution à proprement parler.

Il apparaît donc que s'intéresser au contexte social plus qu'à la biologie ou au genre permet de mettre en lumière des résultats plus nuancés quant aux raisons poussant les individus à utiliser la langue et les jurons de certaines manières. De plus, l'emploi d'une méthodologie plus rigoureuse pour étudier les attitudes linguistiques genrées a permis de faire apparaître les défauts existants dans beaucoup d'études préliminaires, basées plutôt sur de supposées prédispositions génétiques que sur un fondement social valide. Ce renforcement des méthodes d'enquêtes, de collecte de données et d'analyses, ont permis de nuancer les visions très axiomatiques souvent proposées pour décrire la manière dont la vulgarité est utilisée.

De récentes études portant sur la manière dont est utilisée la vulgarité sur Internet semblent indiquer que le Web, et les réseaux sociaux en particulier, pourraient être un contexte d'expression relativement neutre faisant en sorte que les femmes et les hommes utilisent la vulgarité de manière beaucoup plus similaire que dans d'autres contextes. Il semblerait donc que les différences genrées dans l'utilisation des jurons soient bien plus limitées que les stéréotypes le laissent entendre, et cet écart semble se réduire d'autant plus sur Internet.

Chapitre 2 : Jurons, individus et réseaux sociaux : évolution et liens

Il n'est pas rare d'entendre ou de lire des articles de presse mentionnant une certaine banalisation de la vulgarité. Des études de perception ont révélé que les jurons semblaient apparaître comme étant moins offensant au fil du temps, que ce soit chez les hommes ou chez les femmes. Il semblerait également que l'âge des locuteurs/utilisateurs joue un rôle déterminant dans la manière dont les jurons sont perçus : les jeunes générations sont en général les plus tolérantes au regard de la vulgarité et de leur propension à être offensées. La littérature sur le sujet est relativement abondante et indique qu'en plus d'être plus tolérants, les plus jeunes sont également ceux qui utilisent les jurons le plus souvent (Thelwall, 2008 ; Stroh-Wollin, 2010 ; Oliver and Rubin, 1975).

Parmi tous les jurons étudiés de manière approfondie, le mot « fuck » semble être celui qui a connu la plus grande augmentation de sa fréquence d'utilisation, que ce soit dans le cadre privé ou dans les médias télévisés. Ce mot, qui créa un scandale et des excuses publiques lorsque Kenneth Tynan le prononça à la télévision britannique en 1965, peut maintenant être entendu plusieurs centaines de fois dans des programmes accessibles au grand public. Selon certains résultats (McEnery et Xiao, 2004), il semblerait que le développement de nouvelles inflexions de « fuck » serait ce qui a conduit ce mot à pouvoir être utilisé dans de nouveaux contextes, et donc à démocratiser son utilisation.

Certains contextes linguistiques peuvent donc être à l'origine d'une utilisation plus grande de la vulgarité, mais d'autres facteurs sont également à prendre en compte pour attester de ces changements. En effet, Internet semble être un contexte particulièrement intéressant pour des études de ce genre, car les jurons sont beaucoup plus fréquemment utilisés dans certaines de ces sphères. Selon une étude de Wang et al. (2014) basée sur Twitter, environ 8% de ces mini messages (i.e. tweets) contiendraient un juron. Cela semble être bien plus fréquent que dans d'autres types de forums en ligne, où seulement 3% des phrases seraient vulgaires (Subrahmanyam et al., 2006). Pour cette raison, Twitter semble être un contexte idéal pour l'étude des jurons.

Chapitre 3 : La linguistique de corpus comme ensemble de d'outils polyvalents

La linguistique de corpus est le nom donné à l'ensemble des techniques et outils pouvant être utilisés pour l'analyse de corpus textuels. Ces méthodes sont à la base de l'approche utilisée dans cette thèse, et un passage en revue des grandes tendances de cette discipline est nécessaire afin de l'adapter au mieux à mes propres besoins. L'utilisation de corpus à des fins linguistiques n'est pas nouvelle, et de telles pratiques ont été recensées il y a plusieurs siècles de cela. Cependant, ce n'est qu'à partir des années 1950 que cette pratique s'est développée, en grande partie grâce à l'avènement de l'informatique et des corpus numérisés.

Les approches dites « quantitatives » et « qualitatives » sont souvent mises en opposition, telles deux méthodes incompatibles parmi lesquelles il faudrait impérativement choisir. La linguistique dite de corpus se base au départ principalement sur une approche quantitative, mais ceci n'exclut pas une approche qualitative détaillée pour autant. Comme je le montre en détails lors de l'analyse des données, il est capital de toujours prendre le contexte dans lequel un mot/juron est utilisé afin de correctement interpréter son utilisation. Il est donc possible de combiner une première approche quantitative à une approche qualitative basée sur des analyses de concordances par exemple.

J'ai expliqué précédemment que pour une analyse sociolinguistique de la manière dont les jurons sont utilisés, il est nécessaire de se baser sur le contexte précis dans lequel l'étude va être menée. Afin de compiler une liste de jurons considérés comme tels dans cette étude donc, je me suis appuyé sur différentes sources. La première est l'étude de Wang et al. (2014) qui fut basée sur un vaste corpus de tweets en langue anglaise émis dans le monde entier. Ce corpus fut annoté grâce au travail de deux locuteurs anglophones natifs qui ont déterminé, à partir d'une liste de plusieurs centaines de jurons, lesquels étaient, selon eux, utilisés le plus souvent dans le but d'être vulgaire. Ensuite, à partir de cette liste de jurons, Wang et al. ont pu calculer lesquels étaient les plus fréquents. Ils ont ainsi pu conclure que les sept jurons les plus fréquemment utilisés représentent à eux seuls plus de 90% de l'ensemble des occurrences de ce type de mots. Il semble donc que le panel des jurons anglais sur Twitter est restreint à une poignée d'entre eux utilisés massivement. En se basant sur ces jurons-ci donc, ainsi que sur ceux présentés par la BBC (une institution britannique) comme devant être censurés, j'ai pu

établir une liste de jurons dans cette étude, et étant représentatifs du contexte en question, à savoir des tweets britanniques. La liste finale de jurons pris en compte pour cette étude est la suivante : *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*.

DEUXIEME PARTIE : METHODOLOGIE

Chapitre 4 : De « la vraie vie » à l'API

Le titre de ce chapitre fait référence aux liens et différences qui existent entre les interactions qui ont lieu lors d'échanges verbaux entre deux personnes se trouvant face à face, et les interactions ayant lieu sur les réseaux sociaux. L'acronyme API (Application Programming Interface) fait référence aux protocoles qui permettent d'interagir entre des sites Internet et des bases de données, et ce procédé est celui que j'ai utilisé pour collecter des données depuis Twitter.

Twitter fut dans mon cas une source de données extrêmement intéressante car aucun corpus préétabli ne correspondait à mes attentes. De plus, comme expliqué précédemment, la forte présence de jurons sur ce réseau social en fait un atout majeur.

Le hasard des rencontres a fait que j'ai pu créer des liens académiques avec des chercheurs en informatique, liens qui ont mené à la création d'une interface Web, CATS (Collection and Analysis of Tweets made Simple), servant à la collecte et analyse facilitées de données depuis Twitter. Manipuler les APIs est en effet une tâche ardue pour beaucoup de chercheurs en sciences sociales qui, comme moi, n'ont pas les connaissances nécessaires en programmation pour écrire eux-mêmes leurs propres programmes. CATS a donc pour but de combler ce manque en proposant un outil facile à utiliser pour la collecte de tweets.

Chapitre 5 : Comment CATS m'a permis de collecter et d'ordonner mes données

La présentation détaillée de la méthodologie utilisée est centrale à toute étude. C'est grâce au fait de pouvoir répliquer une étude qu'il devient possible d'en comprendre les avantages et limites, et donc de pouvoir faire avancer les connaissances que l'on a de ce domaine. Le but de ce chapitre est de donner plus de détails quant à la manière dont j'ai utilisé CATS, afin de comprendre l'organisation du corpus et le type de données le composant.

La durée de collecte des tweets fut fixée à deux mois. Il est capital d'avoir un flux ininterrompu de données pendant une période prolongée afin de limiter une possible surreprésentation d'évènements ponctuels ayant incité les utilisateurs à réagir pendant une durée prolongée. C'est le cas par exemple du Brexit, qui a eu lieu durant la phase de collecte des données. Un événement aussi marquant a nécessairement suscité de nombreuses réactions des utilisateurs britanniques et pourrait donc être considéré comme un biais potentiel influençant la manière dont les individus se sont exprimés sur Twitter pendant la collecte. C'est pour cette raison que la durée de collecte a été étendue à deux mois.

Grâce à la géolocalisation des tweets, il fut possible de collecter uniquement des messages ayant été détectés comme étant émis depuis le Royaume Uni. Ceci inclut toutefois également les tweets de personnes n'étant pas britanniques, mais tout de même situées sur le territoire lors de la collecte.

A cause du nombre important de métadonnées associées à chaque tweet[189], il est préférable, pour des raisons logistiques et de stockage, de ne conserver que les plus importantes. Dans mon cas, je n'ai gardé que les informations concernant l'heure d'émission du tweet, la description du profil de l'utilisateur, le contenu du tweet en lui-même, la date, et le nom de l'utilisateur.

[189] Ces métadonnées incluent des informations sur le tweet et l'utilisateur, telles que le nom de l'utilisateur, son nombre de « followers », la date d'émission du tweet, la description du profil de l'utilisateur, la langue du tweet etc…

Chapitre 6 : quel âge a-t-il/elle ?

L'un des problèmes qui se pose dans le cadre d'une analyse de tweets prenant en compte le genre et l'âge des utilisateurs est qu'en dépit du grand nombre de métadonnées fournissant des informations supplémentaires sur le tweet et son émetteur, ces deux informations n'en font pas partie. Il est donc nécessaire de recourir à des techniques permettant d'inférer le genre et l'âge des utilisateurs avant de pouvoir analyser le corpus.

Afin de détecter si les utilisateurs se déclarent comme étant des hommes ou des femmes, je me suis appuyé sur le prénom renseigné. Les utilisateurs ayant la possibilité de fournir un nom dans leur profil, se baser sur ce nom et procéder à une analyse automatique du genre de chacun d'entre eux est un moyen d'avoir accès au genre déclaré de l'utilisateur. Bien sûr, tous ne fournissent pas de noms, ou de noms valides tout du moins, qui m'auraient permis de prendre en compte 100% des utilisateurs. Pour les autres en revanche, en me basant sur des listes de recensement, et sur le sexe des enfants à qui ces prénoms ont été donnés, il m'a été possible de composer une liste de prénoms masculins, et une liste de prénoms féminins. Au total, ce sont donc près de 53 000 prénoms genrés qui sont à la base de ce processus. Toutefois, j'ai dû recourir à certaines techniques afin de contourner le problème posé par les prénoms ambigus (qui peuvent être à la fois masculins et féminins), ainsi que pour résoudre le problème de caractères spéciaux parfois utilisés dans les noms des utilisateurs.

Pour ce qui est de détecter l'âge, j'ai pour cela utilisé les informations fournies dans la partie « description » du profil des utilisateurs, qui correspond en général à une brève biographie dans laquelle l'âge est souvent mentionné. En analysant manuellement ces données, je me suis rendu compte que le plus souvent, les utilisateurs ont recourt à des méthodes très spécifiques pour mentionner leur âge, qui sont :

- L'âge est le premier élément de la description.
- L'âge est situé entre deux caractères non-alphanumériques.
- L'âge est explicitement mentionné et est suivi d'un élément tel que « … years old ».
- L'âge est explicitement mentionné et est précédé d'un élément tel que « I am… ».

Une fois le programme optimisé, et afin de palier à certains cas exceptionnels, le corpus a dû être découpé en différentes sous-parties prenant en compte l'âge et le genre des utilisateurs.

TROISIEME PARTIE : RESULTATS

Chapitre 7 : Quelques chiffres

Selon les fréquences d'utilisation des différents jurons pris en compte dans le cadre de cette étude, nous pouvons en déduire que la prédiction de Thelwall (2008) impliquant que les femmes utiliseraient des jurons « forts » plus fréquemment que les hommes n'est pas avérée. En effet, les résultats indiquent même l'inverse, car les deux jurons principalement considérés comme tels par Thelwall (i.e. *fuck* et *cunt*) se révèlent ici être les deux jurons les plus utilisés par les hommes, toutes générations confondues. Cependant, nous pouvons observer que lorsque nous nous focalisons uniquement sur les formes abrégées de *fuck* par exemple, celles-ci sont en fait utilisées bien plus fréquemment par les femmes. Il apparaît donc que l'analyse quantitative globale réfute plutôt l'hypothèse de Thelwall, mais une analyse plus spécifique semble nuancer ces observations, car il se pourrait que l'utilisation des abréviations soit un moyen pour les femmes de se réapproprier certains de ces mots qui sont parfois considérés comme étant typiquement masculins. Nous pouvons donc en déduire que bien qu'une analyse quantitative assez générale semble indiquer, au moins partiellement, que la vulgarité est plutôt utilisée par les hommes, des analyses bien plus poussées sont nécessaires avant de clairement confirmer cette dichotomie. De plus, il doit être signalé que malgré ces différences dans la proportion de tweets vulgaires, la grande majorité des tweets masculins et féminins, toutes tranches d'âge confondues, ne comporte aucun juron. Aussi, et au-delà de l'opposition hommes/femmes qui peut parfois être faite, ces résultats nous permettent de nous rendre compte du fait que l'âge a ici beaucoup plus d'influence sur la variation linguistique que le genre, et donc que l'influence mutuelle de ces deux paramètres est un élément crucial afin de pleinement rendre compte des contrastes existants.

Chapitre 8 : Qui préfère quoi ?

Ce chapitre a pour but d'utiliser des techniques statistiques afin de déterminer les jurons, ainsi que les mots, les plus significatifs de chaque genre et chaque tranche d'âge. Les principaux tests statistiques utilisés sont le Mann-Whitney U et le simple maths parameter dont je me suis servi pour effectuer une analyse des mots-clés les plus marqués au sein de chaque sous-groupe. Le but est ici d'aller plus loin que les tendances dégagées dans le chapitre précédent en utilisant des tests plus spécifiques, et en analysant les données sous des angles différents, notamment en mettant en lumière les tendances inter et intra générationnelles, ainsi que les tendances genrées. Ces observations ont permis de confirmer que l'âge est bien un facteur plus important que le genre lorsqu'il s'agit de déterminer quels jurons sont préférés au sein de chaque groupe d'utilisateurs. Ce chapitre a donc pour objectifs de montrer que :

1) L'interprétation des données ne doit pas se faire trop tôt dans le processus d'analyse, sans quoi le risque de négliger certaines approches pouvant nuancer les observations préliminaires s'accroît.
2) Le genre ne semble pas être un facteur aussi déterminant que les stéréotypes le présentent. Il semblerait donc qu'analyser les attitudes linguistiques « genrées » sans prendre d'autres paramètres en compte n'est pas suffisant pour faire état du spectre complet des procédés linguistiques auxquels les individus ont recourt lorsqu'ils s'expriment. C'est aussi le signe que parler de « langage féminin » et de « langage masculin » n'est pas fondé sur des bases empiriques, et que ces idées sont effectivement plus fondées sur du folklore que sur des réalités linguistiques avérées.

Chapitre 9 : Analyse des cooccurrences

En linguistique, la cooccurrence fait référence à l'apparition répétée de certains mots à proximité les uns des autres. Il est possible, en analysant les cooccurrents les plus marqués d'un mot, d'étudier le contexte dans lequel ce mot est utilisé. Ceci peut être une indication forte des contextes dans lesquels certains sous-groupes vont préférer utiliser certains jurons par exemple. Le but de ce chapitre est d'analyser les cooccurrents des jurons ou des mots ayant été révélés comme particulièrement représentatifs de certains sous-groupes dans les chapitres précédents. Ces analyses permettent de mettre en lumière le fait que les jurons sont très souvent utilisés par les hommes comme par les femmes dans des expressions figées. Ceci montre une fois de plus des similarités marquées entre les deux genres, mais nous pouvons toutefois également observer quelques différences propres à ces groupes. En effet, bien que ces expressions soient utilisées par les femmes comme les hommes, elles sont aussi parfois adaptées en fonction des préférences linguistiques propres de chacun de ces groupes.

Nous remarquons aussi que les thématiques observées précédemment comme étant propres aux hommes et aux femmes (e.g. le football pour les hommes et les émotions pour les femmes), sont toujours présentes, et ce même lorsque l'on se focalise sur des événements ou des cas très spécifiques. Cela implique que certaines tendances genrées se retrouvent à tous les niveaux d'analyse, du plus global au plus spécifique.

Nous nous apercevons également que les thématiques caractéristiques de chaque sous-groupe se retrouvent aussi bien dans des contextes vulgaires que non-vulgaires, ce qui indique que l'utilisation de la vulgarité n'est pas limitée à un contexte particulier, et que les utilisateurs ne font pas de différenciation entre les sujets abordés en étant vulgaires.

Le dernier point important soulevé dans ce chapitre est le fait que pour chaque sous-groupe, nous observons des cas dans lesquels l'utilisation d'un juron semble fortement influencer l'apparition d'autres mots. Il semblerait donc qu'il existe une directionalité forte entre certaines paires de mots, et cela semble être le signe de préférences propres à certaines tranches d'âge ou à un genre donné.

Conclusion

Les objectifs principaux de cette thèse étaient doubles : 1) elle visait à déterminer si les jeunes générations de femmes utilisent la vulgarité (ou les jurons dits « forts ») plus fréquemment que les hommes au Royaume Uni et sur Twitter ; et 2) elle avait pour but de fournir de nouvelles pistes méthodologiques afin d'explorer les possibilités offertes par Twitter dans le cadre d'analyses sociolinguistiques.

Concernant le premier point, et comme nous l'avons déjà vu plus haut, nous pouvons conclure que, au moins en ce qui concerne cet échantillon, les jeunes générations de femmes ne sont pas plus vulgaires que les hommes au Royaume Uni, pas plus qu'elles n'utilisent des jurons « forts » à une plus grande fréquence. En effet, la tendance la plus marquée est celle incitant les jeunes générations à se démarquer des autres en étant plus vulgaires. Au lieu d'annoncer une situation de déséquilibre dans laquelle les femmes utiliseraient la vulgarité démesurément plus que les hommes donc, il semblerait que nous nous dirigions plutôt vers une sorte de stabilité sociolinguistique. Certains jurons ou thèmes se sont toutefois révélés comme étant plutôt masculins ou féminins, mais là encore ces tendances tendent à être nuancées en observant de plus près les différences inter et intra-générationnelles existantes. De là, le besoin de prêter autant attention aux différences qu'aux similarités dans un corpus a été mis en lumière afin de pouvoir rendre compte des différents aspects que peuvent présenter les données.

Concernant le second point, ces résultats ont montré qu'en dépit de l'absence d'informations explicites concernant le genre et l'âge des utilisateurs sur Twitter, il est possible de surmonter cet obstacle avec une précision relativement grande. Aussi, cela démontre que de nombreux utilisateurs fournissent suffisamment d'informations pour pouvoir inférer ces deux informations. Malgré le fait que ces utilisateurs ne soient pas une majorité, le très grand volume de données en transit sur ce réseau social est tel que même une fraction seulement d'utilisateurs représente tout de même des milliers de personnes. En plus de démontrer que le panel d'informateurs potentiellement utilisable pour une étude de ce genre est vaste, cette thèse montre que Twitter peut être un outil extrêmement intéressant pour étudier le développement de nouvelles formes linguistiques. C'est un aspect crucial à prendre en compte car les néologismes, ainsi que les formes linguistiques émergentes, ne sont pas aisés à analyser en nombre suffisamment grand pour pouvoir tirer des conclusions généralisables.