Acronyms and Emoticons on a Popular Web Forum: Does Gender Makes a Difference? A Corpus-Based Study of Reddit

Marie Flesch

Université de Lorraine – ERUDI

Master Mention Langues et Cultures Etrangères

Spécialité Mondes anglophones,

Parcours Tourisme culturel

2ème session 2016

Acronyms and Emoticons on a Popular Web Forum: Does Gender Makes a Difference? A Corpus-Based Study of Reddit

Par Marie Flesch

Sous la direction de Alex Boulton

Table of contents

Abstract

Introduction……..…….....1

I. Background…….4

1. Computer-Mediated Communication……...4
   1.1. A History of Computer-Mediated Communication……5
   1.2. Computer-Mediated Communication and Gender: Access and Use……..6
   1.3 CMC: Defining the Linguistic Features of a New Medium……9
      A New Medium… .....……9
      Lexical Characteristics of Netspeak……...10

2. Language and Gender…….…13
   2.1. The Relation Between Language and Gender……..……13
      From Male Dominance to Gender Differences……...……13
      The Gender Similarity Approach……15
   2.2. Empirical Studies of Speech and Writing…….....17
   2.3. Language, Gender and CMC……....19
   2.4. Empirical Studies of Netspeak Features and Gender……20

3. Reddit: One of the Largest Online Communities……...…22
   3.1 Overview of Reddit…….…..22
   3.2. Describing Reddit with Herring's Faceted Classification……24
   3.3. Reddit : A Literature Review……27

II. Methods…….……...29

1. Criteria Used to Build the Corpus.……..29
2. Compiling the Corpus……30
3. A Corpus-Based Study of Reddit…….…….32

III. Results…….37

1. Emoticons……...……37
   1.1. Overall Emoticon Use…….....38
   1.2. Emoticon Use and Gender……39
   1.3. Individual Differences…….....……41
   1.4. Comparison of Two Smaller Corpora……..….41

2. Netspeak Lexical Items…….……..41
   2.1. Overall Use of Acronyms and Other Netspeak Features……..…41

2.2. Netspeak Lexical Items and Gender…….……42
2.3. Individual Differences……..44
2. 4. Contrasting Two Male and Two Female Corpora…….…..……45

IV. Discussion

1. Emoticons……...46
1. 1. Emoticons: Discussion of Overall Results……..46
1. 2. Emoticons and Gender……48

2. Acronyms and other Netspeak Features…….51
2.1. Netspeak Lexical Features: Discussion of Overall Results……..51
2.2. Netspeak Lexical Items and Gender……52

3. Limitations and Suggestions for Further Research…….54

Conclusion…….58

References……..…….……...59

Appendix……..…….…….72

Ouverture sur la thèse…….75

Abstract

This dissertation sets out to investigate if there are gender differences in the use of Netspeak features on Reddit, and if the possible differences reflect the traditional distinctions made between male and female communication styles in computer-mediated communication. A corpus-based study was conducted using a 2 million-word corpus built from the contributions of 100 female and 100 male Internet users on the news sharing site Reddit, which is one of the largest online communities. A number of items were examined with the Antconc concordancer, including abbreviations, acronyms, and emoticons. For Netspeak features, most results did not indicate significant gaps between male and female users, although there was a clear difference in the frequency of the acronym *SO*, or "significant other", which was used 89 times in the male corpus and 369 times in the female corpus. The most striking differences were found in the use of emoticons. Even though the female and the male sub-corpora contained about as many types of emoticons, female Redditors used more than twice as many emoticons as males. This finding is consistent with previous research (Witmer & Katzman, 1997; Herring, 1998; Wolf, 2000; Tossell et al., 2012; Baron 2004; Fox et al., 2007; Lyddy et al., 2014). Idiosyncratic use of Netspeak showed that a small number of Redditors were responsible for the production of most features especially in the case of emoticons. The results suggest other variables may interact with gender in the production of Netspeak features.

Introduction

With the Internet revolution of the 1990s and the mobile Internet revolution of the 2000s, computer-mediated communication or CMC has flourished and has become embedded in daily life. In 2015, an estimated three billion people were surfing the Internet worldwide (Internet Society global Internet report, 2015). At first used by the general public to write emails, browse the web and send files, the Internet now allows many different types of interaction, due to the invention of increasingly portable and polyvalent terminals.

Because of its unique constraints, computer-mediated communication has been described as a "third medium" (Crystal, 2001, p. 52), which is not simply written speech, and has different properties from speech and writing. This new medium of communication as been referred to as "electronic discourse" (Davis & Brewer, 1997), "electronic language" (Baron, 1998), "Computer-Mediated Communication" or "CMC" (Herring, 1996), and "Netspeak" (Crystal, 2001; Thurlow, 2001). It is not homogenous, and possesses features unique to the different types of outputs: email, chat, instant messaging, multi-user dungeons, blogs, message boards, social media and video chat. On the lexical level, computer-mediated communication or Netspeak is characterized by the use of non-standard spellings, neologisms, abbreviations, acronyms, emoticons, creative use of punctuation, and alpha-numeric substitutions (Crystal, 2001; Thurlow, 2001)

In the 1990s, computer-mediated communication brought in new hopes among activists who thought that it would act as a "great equalizer" and that the absence of face-to-face interaction would create a space where there would less discrimination, racism and sexism (Wolf, 1998, p. 15). Researchers in the field of language and gender have however abundantly commented on the persistence of gender differences in CMC. Studies have investigated various outputs in order to verify the existence of gendered communication styles. On many Internet platforms, men were found to adopt an adversarial style, while women tended to be more supportive. The male "adversarial" style has been characterized by the use of strong assertions, self-promotion, profanity, sarcasm, rhetorical questions and lengthy contributions. Females, on the other hand, have been reported to use more hedges, to apologize more, to express appreciation and to ask more questions (Herring, 1994; Cushing,

1996). It has also been found that they were more frequent users of emoticons, one of the lexical innovations brought about by computer-mediated communication (Witmer & Katzman, 1997; Wolf, 2000; Baron, 2004; Fox, Bukatko, Hallahan, & Crawford, 2007; Tossell et al., 2012). If emoticons have been studied extensively, a smaller number of studies have focused on the relation between the use of other Netspeak lexical features, such as acronyms, abbreviations and nonstandard spellings, and gender. They have found mixed results; some have revealed that males use more nonstandard forms online than females (Baron, 2004; Squires, 2007; Van Der Meij, 2007) while some have shown that females tend to produce more nonstandard typography and orthography than males (Herring & Zelenkauskaite, 2009). Other research has found no correlation between gender and usage of Netspeak features (Peersman, Daelemans, Vandekerckhove, Vandekerckhove, & Van Vaerenberg, 2016). Currently, the body of research about English Netspeak and gender is however too small to make broad generalizations.

This dissertation sets out to investigate the relationship between gender, Netspeak lexical features and emoticons by conducting a corpus study of a web forum. A two-million word corpus was compiled using content from the news sharing website Reddit. Launched in 2005, Reddit is one of the largest English-speaking online forums, and has recently attracted the attention of researchers (Danish & Talukdar, 2015; Noguti 2016). The Reddit corpus is made up of two gendered sub-corpora; each contains about one million words, and was created by compiling comments posted during the years 2015 and 2016 by a hundred female Reddit users and a hundred male Reddit users. A corpus-based analysis was conducted using, on the one hand, a list of emoticons adapted from Wikipedia, and, on the other hand, a list containing acronyms, initialisms, non-standard spellings, alpha-numerical substitutions and Netspeak terms, adapted from Wikipedia and other non-academic online sources. The overall results show that a small number of Netspeak features accounts for the bulk of the items found in the corpus. Females used more than twice as many emoticons tokens as men did, although they produced about the same number of types of emoticons. Analysis of other lexical items yielded less contrasted results. The most significant difference was found in the use of the acronym *SO*, which stands for "significant other". It is the second most frequent Netspeak item in the female corpus, and was used four times more often by female Reddit users than by their male counterparts. Despite the discrepancies in the

frequency of usage of emoticons and of some lexical items, the results reveal more similarities than differences between the two sub-corpora.

Part I takes a look at the key concepts and reviews previous research in the fields of language and gender, and computer-mediated mediated communication. It also provides an overview of Reddit. Part II describes the methods and the materials used to build the corpus. The results of the corpus-based study are presented in Part III. Part IV discusses the findings, addresses the limitations of the study, and touches on its implications.

I. Background

1. Computer-Mediated Communication

1.1. A History of Computer-Mediated Communication

The origins of computer-mediated communication can be traced back to the creation of the first digital computers, a little before World War II. Even though computers were invented to perform calculations, engineers and scientists soon understood that they could also be used to communicate. Vannevar Bush, a researcher at MIT, was probably the first to envision the possibilities that computers could offer. In the 1930s, he introduced the concept of the memex, a machine that would extend the powers of human memory, and allow scientists to keep up with the growing amount of studies (Press, 1993, p. 27). Several experimental interactive computers were created at MIT in the 1950s, including Jay Forrester's Whirlwind computer, which was able to process telemetry data in real time (Press, p. 29). The concept of time-sharing, or the sharing of a computer resource among many users, was described for the first time by British computer scientist Christopher Strachey, and would be instrumental in the development of the Internet.

The experimental time-sharing systems devised in the 1960s posed a problem; they were all printing terminals, and did not have a screen or display. Sutherland's Sketchpad led to the creation of the first WYSIWYG interfaces, and Engelbart's word processing programs allowed computers to process texts (Press, p. 28). While scientists such as Kememy, Kurtz and Albrecht were still trying to make computers less expensive and more user-friendly, the computer network was created by military-funded researchers. Because of the Cold War, the U.S. Defense Department's Advanced Research Projects Agency, or ARPA, was looking for a way to exchange data between geographically dispersed computers in the event of a nuclear war. It led to the creation of the ARPANET, the ancestor of the Internet (Dobrow, 2015). The first commercial online service, CompuServe, was launched in 1969. By the early 1980s, a number of networks had developed. Meanwhile, the personal computer was invented; it included music, drawing, animation, email and

word processing programs (Press, 1993, p. 31), and was becoming more and more portable and easier to use.

In 1990, the World Wide Web, a user-friendly way to access the information already on the Internet, was created by Tim Berners-Lee at the CERN physics laboratory in Geneva. It was based on the concept of hypertext, a way to store and retrieve information in a non-hierarchical structure, and was implemented the next year. This development, together with the passing of federal legislation which created a national information infrastructure promoted by then-Tennessee senator Al Gore led to the rapid growth of the Internet on an international scale. Non-experts quickly embraced the Internet, mostly because of email. By the mid-1990s, 20 to 30 million computers were connected to the Internet (Dubrow, 2015). At first, the Internet was mainly used for browsing, transferring files, reading and posting on bulletin boards and email; over the years, other types of technologies facilitating interaction were developed, such as newsgroups, bulletin boards, blogs, internet relay chat, instant messaging, metaworld, visual chat, and personal homepages (Thurlow, Lengel & Tomic, p. 31).

After the personal computer and the Internet, the invention of the smartphone and the mobile Web brought up a new revolution. Liberated from the constraints of wired connection, hundreds of millions of people gained access to the Internet. It was estimated that there were 3 billion Internet users in May 2015, a number expected to keep growing, as mobile Internet penetration is forecast to reach 71% in 2019. Not only did smartphones allow more people to use the Internet, but they also offered new functionalities, acting as telephones and web browsers, but also as GPS devices, cameras, dictaphones and storage devices. Even if mobile networks cover almost half the global population and 192 countries, there are still large disparities on a global level; 79% of Europeans and 84% of North-Americans are Internet users, but only 47% of population in Latin American and in the Caribbean, and 17% of the people living in Sub-Saharan Africa have access to the Internet (Internet Society global Internet report, 2015).

1.2. Computer-Mediated Communication and Gender: Access and Use

Even though past studies about the extent of gender difference in Internet and computer use have often been contradictory (Kimbrough, Guadagno, Muscanell & Dill, 2013), most researchers agree that early adopters of CMC were primarily men.

In 1994, a mere 5 % of Internet users were women (Weiser, 2000, p. 168). In the early 1990s, not only did men constitute the majority of users, but they also dominated the technology. The Internet was, as Weiser put it, a "boy toy", "an electronic pathway to riches of information and entertainment that was developed exclusively by men for men" (p. 168).

Several reasons have been given to explain the gender gap during this first era of Internet use. According to Weiser, the gender gap was due, in part, to the fact that women were socialized away from the science, technology and engineering fields, with girls being discouraged from studying math and science, and women thus less likely to choose technology-related jobs. Other arguments have been advanced, such as the "double day" of work of many women and the time constraints it involves, or the fact that women considered the Internet as a threatening place, rife with pornography and sexual harassment (Weiser, 2000, p. 169). Women mainly had access to computers at work, and used them primarily for word processing and spreadsheet work (Kaplan, 1994, cited by Wasserman & Richmond-Abbott, 2005, p. 253). As a result, males had greater technical competence than women, as reported by Durndell and Haag (2002) in their study of Romanian students. Another study of students' computer use noted that women considered themselves less familiar with computers than men, while men expressed more positive attitudes towards the technology (Sherman et al., 2000).

However, despite the lack of confidence of female users, the gender gap narrowed quickly. In the United States, women outnumbered men online as early as 2000, and, between the years 2000 and 2001, new users of the Internet tended to be female (Shaw & Gant, 2002, p. 518; Cummings and Krout, 2002). By 2000, it was no longer possible to talk of an Internet gender gap in the United States, at least in terms of Internet access (Nie & Erbring, 2002). It was also true of a number of other countries. Gender parity was reached in 2000 in many European countries including France, the United Kingdom, Belgium, Denmark and Portugal (Norris, 2001, p. 83). In 2015, on a global scale, there were still large gender gaps in Internet use, especially in less economically developed countries, such as Senegal, where twice as many males have access to the technology as females, or Bangladesh, where 8% of men use the Internet while only 5% of women do (The Global Gender Gap Report, 2015, pp. 310 & 101). However, in many countries, such as Australia, Egypt, Canada, Brazil, Mexico, Thailand and Japan, the gap has disappeared or is very

narrow. The increasing presence of women on the Internet has been seen as a reflection of the changes occurring in society as a whole, with more women entering science and engineering professions and participating in office administrative activities (Wasserman & Richmond-Abbott, 2015, p. 253). The rapid decline of the gender gap, in regard to access, has also been attributed to technological progress, which made the Internet more affordable and easier to use (Weiser, 2000, p. 169).

Access to the Internet is however not the only area where gender discrepancies have been noted. There are other aspects to the gender gap, notably the frequency and the scope of use of the Internet (Wasserman & Richmond-Abbott, 2005, pp. 253-254). Even if gender equality has been reached in terms of access, men and women still seem to differ in the amount of time they spend online and in the ways they use the Internet (Fallows, 2005). However, some results are contradictory. In 2000, in a survey of 843 college students, Odell et al. found that males spent an average of 7.1 hours per week on the Internet, and females only 4.5 hours. Herring & Stoerger (p. 3) reported that, in 2009, men sent online more often than women, spent more time on the Internet, and visited more websites (2013, p. 3) while Kimbrough et al. (2013) reported that women were more frequent users of computer mediated communication than men.

Many studies seem to lend support to the existence of gender gap in the scope of use of the Internet. In several regards, research has shown that males and females still act, online, according to traditional gender roles. As Shaw & Gant have noted, there is a tendency "to compartmentalize different types of Internet use and to label them "male" or "female" (Shaw & Gant, 2002, p. 518). Herring & Stoerger have even gone so far as to claim that the web is becoming "increasingly specialized by gender" (2013, p. 3). In keeping with the gender differences paradigm, men have been found to dominate task-focused activities while activities geared toward interpersonal communication are favored by women (Fallows 2005). The results of Guadagno et al.'s (2011) study of the virtual world Second Life support the idea that, online, people are behaving according to traditional gender roles. Women tend to seek out communal activities, such as meeting people and shopping, whereas men focus more on agentic activities, like building things and owning property (Guadagno et al., 2011).

Email, an activity based on social interaction, has been said to be used more by females than males (Sherman et al., 1999, Odell, Korgen, Schumacher, & Delucchi

2000, Weiser, 2000). However, some studies and surveys have only found modest differences in use between the two genders. In Odell et al.'s study, for instance, 91% of the female participants reported to use the Internet for email, while 85.7% of males did. Some studies suggested that chatting is not gendered in the way email is (Odell et al., p. 857), and some studies have shown that men used it more than women (Sherman et al., 1999) or that women used it more than men (Kimbrough et al., 2013, p. 898). Other social types of CMC, such as video chat and text messaging have also been found to be more frequently used by women (Kimbrough et al. 2013, p. 898). Online forums, on the other hand, have been described as dominated by men (Herring, 1992, 1993). In their study of a small forum frequented by academics, Herring, Johnson & DiBenedetto (1992) found that women contributed to only 30% of the messages. When feminist topics were discussed, women contributed more, but disruptions ensued because some men felt that they were being silenced.

Task-focused activities, such as shopping, listening to music, building web pages, searching for romance, viewing pornography, playing games and reading the news have been found to be favored by males (Odell & al., 2000, Weiser, 2000). Consequently, male use of the Internet has been said to be "more flexible than female use" (Weiser, 2000, p. 175). More recent studies seem to confirm that some areas of the Internet are still dominated by men. Lam et al. (2011) found that there is a large gender gap in the population of Wikipedia contributors with only 16.1% of editors being females, and male editors making almost twice as many edits as female editors.

Social media, on the other hand, is an area where women outnumber men in many cases, as Kimbrough, Guadagno, Muscanell and Dill (2013) reported. Shepherd (2016), drawing on data from a 2011 survey of students' Facebook use, stated that women had had Facebook accounts longer than men, that they used it more often and spent more time on the social networking site (p. 22). Another study showed that women and men did not use social media for the same purposes. Women tended to use Facebook for maintaining existing relationships, whereas men tried to form new relationships and looked for jobs (Muscanell & Guadagno, 2012). However, surveys, such as those conducted by the Pew Research Center, reveal that the social media gap is disappearing, in the United States at least. A report published in August 2015 estimates that 73% of men and 80% of women who use the Internet have a social media account, while the gap was as large as 15 percentage points in 2010 (Anderson, 2015). There are still differences, however, on specific sites.

Female users dominate on Facebook, Pinterest and Instagram, while online forums such as Digg, Reddit and Slashdot attract mostly men. No significant gender differences were found on Twitter, LinkedIn and Tumblr.

1.3 CMC: Defining the Linguistic Features of a New Medium

A New Medium

Early researchers who tried to define the characteristics of CMC did not have access to a lot of data and often viewed the language of the Internet and of CMC as a whole, without making a distinction between the different genres. Ferrara, Brunner, & Whittemore (1991) were among the first to study the syntactic and stylistic features of what they called "Interactive Written Discourse" or IRW. Analyzing a corpus of an early form of instant messaging, they found IRW to be "a hybrid", "a novel writing style" (p. 30) that shares features with both speech and writing. Yates (1996) conducted a corpus study comparing CMC with corpora of written and spoken English. Using a corpus made of computer conference interactions, he found CMC was similar to speech in the use of pronouns and modals, and that it was more like writing in lexical density (Yates, 1996, pp. 41-42).

At first, researchers tried to analyze Internet language by comparing it to speech and writing. They used lists of classic distinctions between speech and writing, such as durable and ephemeral, planned and spontaneous, or complex syntax and simpler syntax, and used these binary features to characterize the different Internet outputs like the web, blogs, virtual worlds and chatgroups. This model proved to have limitations, however (Baron, 1998). The dichotomy it is based on has been viewed as too simplistic, since there are significant differences between various kinds of written and spoken language. Phone conversations between friends, for instance, do not share many features of lectures or news broadcast. Moreover, many CMC outputs blur the boundaries between speech and writing. In instant messaging, for instance, messages are typed, like in writing, but they are almost synchronous, which makes them closer to speech. To address this issue, a "spectral model" was proposed, which added to the traditional distinction made between speech and writing other media of communication, such as computer, telephones and teleconferencing (Baron, 1998), and tried to see where the various CMC platforms fall along this continuum. Baron also described a "cross-modality" mode, which takes into account the fact that linguistic messages are not always designed for a single modality.

Herring's (2007) classification of CMC features is probably a more appropriate tool for describing CMC. She adopts an approach inspired by the way information is organized in the fields of library and information science. Her model is based on two broad categories of "facets", or "concepts of the same inherent type". The first category includes technological features such as synchronicity, granularity, persistence, length, channels, identity, audience, adaptation and format. In the category of social facets, Herring identified participation structure, participant characteristics, purpose, activities, topic, tone, norms of organization, norms of social appropriateness, norms of language and code. Not all facets have to be used to describe a CMC output, and other features may be added. According to Herring, the model is useful because it cuts across the boundaries of different modes and "allow for the identification of a more nuanced set of computer-mediated discourse types, while avoiding the imprecision associated with the concept of genre" (para. 8).

Lexical Characteristics of Netspeak

Describing the lexical features of Netspeak is not an easy task since, as Crystal writes, language change seems to happen on the Internet at a much faster rate than at any previous time in linguistic history (2001, p. 98). Not only is Netspeak fluid, but it is not homogenous; its features vary depending on the different types of output, the devices used and the groups of people who use it. Thus, the language of the Internet is not a single dialect, and is "extremely diverse and segmented" (Driscoll, 2002, para. 2).

Thurlow (2001) and Crystal (2001) were among the first to attempt to list the characteristics of Netspeak. Thurlow describes Netspeak as "dynamic, unpredictable, spontaneous, and public", and claimed that it is characterized by strategies used "in pursuit of brevity, paralinguistic restitution, relationship, and play". He listed a number of these "strategies", which include lexical compounds and blends, abbreviations and acronyms, minimal use of capitalization, non-standard spellings, letter homophones, creative use of punctuation, emoticons, use of capitalization and other symbols for emphasis and stress, direct requests, interactional indicators and "emotes", or emojis (pp. 288-289). Crystal's description of Netspeak is more condensed, but contains the same key items: neologisms, abbreviations, distinctive graphology, random use of capital letters, new spelling conventions and minimalist punctuation. To this list, Crystal adds grammatical variation, which manifests itself

in phenomena such as verb reduplication, although he states that it may be rare, and that the most distinctive features of the variety are found in graphology and in the lexicon, "the levels of language where it is relatively easy to introduce innovation and deviation" (2001, p. 96-97).

In another early study of Netspeak, Driscoll (2002) investigates the language used by a group of gamers in a MUD chatroom and identifies a number of features: clips, acronyms, compounds, blends, suffixes, unique words, voice pauses, laughter, alpha-numeric substitutions, alternative spelling, and emoticons. When Baron (2004) researched the language of instant messaging, she found four lexical categories of items in her corpus: abbreviations, acronyms, contractions and emoticons. Other studies that have tried to quantify the use of Netspeak features have generally found abbreviations, acronyms and some non-standard spellings to be the most frequent items, followed by emoticons. Varnhagen et al. (2010), analyzing a corpus of instant messaging conversations produced by adolescents, reported that "short-cuts" were the most prevalent. Unconventional spellings such as the dropping of lowercase letters were also noted (Lyddy, Farina, Hanney, Farrell and O'Neill; 2014; Tagliamonte & Denis, 2008.)

The scarcity of Netspeak features in the language of instant messaging, chatroom, emails and online forums has also been reported by several studies. Tagliamonte and Denis noted the "sheer infrequency" of IM forms in their 1.5 million-word corpus, which represented less than 3% of their data. Referring to her own research of IM use among college students, Baron (2005) noted that students "eschew brevity" and use few abbreviations and acronyms, adding that "IM is unlikely to play a significant role in altering writing standards" (p. 31).

Corpus studies of computer-mediated communication reveal that acronyms and initialisms expressing laughter, such as *LOL* or *LMAO*, are among the most frequent Netspeak lexical items. *OMG*, which stands for "Oh my god", has also been reported to be one of the most frequent Netspeak acronyms (Baron, 2004; Tagliamonte & Denis, 2008). Abbreviations, acronyms and other types of "shortcuts" have been described as "shortcuts", and were reported to be prevalent in IM (Varnhagen et al., 2010). Non-standard spellings such as *u* for "you" have also been noted to be relatively frequent (Lyddy et al., 2014), although Tagliamonte & Denis (2008) reported that it was used by a minority of IM users (p. 14).

Perhaps because they are more noticeable and have become emblematic of the language of the Internet and text messaging, emoticons have been studied extensively. Emoticons are graphic representations of facial expressions created by combining ASCII characters. They are not to be confused with emojis, which are Unicode characters such as 😕 or 🙄. The majority of emoticons are meant to be read sideways. Most, like :-), are read left to right; other are meant to be read right to left, like d: or D:. Vertical emoticons also exist; they are mostly used in countries such as Japan or South Korea (Park, Barash, Fink, & Cha, 2013). Hundreds of emoticons exist, adding to an ever-growing list of combinations of characters. The paternity of the emoticon is often attributed to American computer scientist Scott Fahlman, who claims to have invented it in the early 1980s, when he was teaching at Carnegie Mellon University, on the university bulletin board system. The thread in question was unearthed in 2002 on the initiative of Mike Jones, a Microsoft research scientist. In a message, dated September 19th, 2002, Fahlman proposed "the following character sequence for joke markers :-)", adding that it is to was to be read sideways (Complete Bboard Post Sequence, n. d.). He also gives an example of a "sad face", to mark things "that are NOT jokes". The word emoticon itself dates back, according to the Merriem-Webster Dictionary, to 1987, and is a blend of "emotion" and "icon" (Emoticon, n. d.). Emoticons have gained tremendous popularity, especially since the advent of mobile communication (Parks & al., 2013 466), and has developed different forms and meanings (Skovholt, Gronning, & Kankaanranta, 2014, p. 781).

The rapid and widespread diffusion of emoticons has been linked to the lack, in CMC, of nonverbal cues such as gestures and facial expression that are normally used in face-to-face interaction (Crystal, 2001, p. 39). They have been called the paralanguage of the Internet (Dery, 1993, p. 2, cited by Crystal, 2001, p. 37), and have been described as the primary tool used to express emotion in CMC (Riva, 2002), or as a "surrogate for nonverbal emotional expression (Derks, Bos, & Von Grumbkow 2008, p. 99). Rezabeck & Cochenour (1994) is generally considered to be the first academic study of emoticons. The authors studied listserv messages, and argued that emoticon use is a "highly personal decision" (p. 382). Witmer & Katzman (1997) is another early study of emoticons, which focused on gender.

Emoticons have often stigmatized by the media; they have been, for instance, described as "a paltry substitute for expressing oneself properly" and a danger to

punctuation marks (Truss 2004, p. 193). Some journalists have even called for their interdiction, arguing that they are childish and unnecessary (Andrews, 1994). They were also described as a "blunt" linguistic instrument, and as "an unnecessary and unwelcome intrusion into a well-crafted text" (Provine, Spencer, & Mandell, 2007, p. 305-306). Even a linguist like Crystal, who embraces the language of CMC, has pointed out the limitations of emoticons, "a potentially helpful but extremely crude way of capturing some of the basic features of facial expression", with a "limited" semantic role" (Crystal, 2001, p. 39).

More recent research, however, suggests that emoticons are not just graphic representations of emotions, and that their roles and functions of emoticons are far more complex. In their study of IM conversation, Garrison, Remley, Thomas, & Wierszewski (2011) came to the defense of the emoticon, writing that they "find the pervasiveness of the "emoticon as compensatory and intrusive "limiting" (p. 113). They showed that emoticons can function as an utterance on their own, and thus proposed that they must not been seen as having only a secondary role in CMC. Skovholt et al. (2014) emphasized the need for more in-depth research about the authentic use of emoticons. Looking at emails exchanged in the workplace, they found that emoticons do not primarily indicate emotions, and identified three communicative functions of emoticons: indicators of a positive attitude, joke and irony markers, and pragmatic markers that help contextualize or modify an utterance. Thomson and Filik have also argued that emoticons are markers of intention (2016, p. 105) and that they help strengthen messages (p. 106). Kaye, Wall & Malone point out that emoticons are used more consciously than nonverbal behavior, and that users have more control over the message they want to convey (2016, p. 380).

2. Language and Gender

2.1. The Relation Between Language and Gender

From Male Dominance to Gender Differences

Research about language and gender dates back from before the advent of feminism in the late 1960s and the 1970s. In 1922, Danish linguist Jespersen dedicated a chapter of his book *Language: Its Nature, Development and Origin* to "The woman", claiming that women have a smaller vocabulary than men (p. 247), or that they use incomplete sentences more often than men because they "start talking

without having thought out what they are going to say" (p. 249). Other linguists and ethnographers embarked on empirical fieldwork on gender differences in speech in the early twentieth century. Haas, for instance, studied the Native American language Koasati, and described the verb forms used by women as appearing to be "more archaic than that of men" (1944, p. 142).

After the emergence of the Women's Liberation Movement, language became a target of feminists, who claimed that it was inherently sexist and that it often adopted a masculine bias. Some features of the English language, such as the fact that women change their names when they get married, or the use of the generic "man", were seen to have for effect to degrade women and to make them invisible (Sunderland, 2006, p. 10). Soon, linguists joined activists, focusing not only "language sexism", but also on language use. In 1975, Lakoff published *Language and Woman's Place*, the first monograph about gender and language. In this pioneering book, Lakoff, adopting a feminist perspective, explores women's speech, the ways women are spoken of, the role of cultural institutions in creating and reproducing linguistic practices and ideologies, and the linguistically based cultural systems, such as politeness, which perpetuate gender power inequity. She presents women's language as deficient and powerless, because it encourages them to talk about trivial topics (p. 46) and it does not provide them with the resources needed to express themselves strongly and with confidence (p. 51). Even if it has been criticized since and seen as accepting sexist idea that women's language as deficient (Sunderland, 2005, p. 13), *Language and Woman's Place* remains one of the most influential works in language and gender (Bucholtz, 2004, p. 3). In the late seventies, Lakoff's work prompted other researchers, such as Fishman, to focus on the male-dominance approach. In 1978, she studied mixed sex-conversation in her article "What do couples talk about when they're alone?" and posited that women's role was to facilitate conversations by asking questions and using hedges (1978, p. 400). In the United Kingdom, Spender popularized the study of gender and language in her 1980 work *Man Made Language*, which used the male-dominance approach to show how men dominate conversations and silence women.

In 1990s, the focus of the discourse and research about language and gender shifted. The origins of this shift may be traced to Maltz and Borker's 1982 paper, "A cultural approach to male-female miscommunication", which interprets in a new light previous findings of works by West, Fishman and Hirschman. The authors

proposed a new framework for studying linguistic gender differences in the United States, claiming that "American men and women come from different linguistic subcultures" (1982, p. 200). The idea of gender differences, instead of male dominance, was more famously addressed by in 1990 by Tannen. In her bestselling book *You Just Don't Understand!: Women and Men in Conversation*, she argues that women and men have different world views, which are "distinct enough to make cross-gender conversation parallel to 'cross-cultural communication' " and posits that women see the world "as a network of connections", while men see it "as a hierarchical social order" (Tannen, 1990, pp. 18-25). Coates is another researcher who has contributed to change the focus of the field of language and gender. Instead of studying mixed-sex talk, she concentrated, in her 1996 book *Women Talk*, on women's speech, using empirical evidence gathered from conversations between friends. Her goal was to emphasize gender differences rather than male dominance, and not to show women as victims. The main way in which male and female communication styles differ, according to Coates, is the fact that men try to be informative and assertive, while women are supportive and want to achieve consensus and closeness.

The Gender Similarity Approach

The "difference paradigm" spawned the publication of many non-academic books, such as the hugely popular *Men are from Mars and Women are from Venus* by Gray, which promoted the notion of fundamental gender differences and which, to this day, has sold more than 50 million copies (Men Are from Mars, Women Are from Venus, n.d.). However, in the academic community, some voices soon rose to put into question the notion of inherent psychological differences between men and women. In 1992, the same year Gray published his bestseller, sociolinguists Eckert and McConnell-Ginet gave a talk at the Berkeley Women and Language Conference, calling on their peers to explore a new direction. They insisted on the need to stop making generalizations about the topic of gender and language, pointing out that there is "tremendous variability" within gender categories (p. 93).

In 2005, psychologist Hyde, who had been working on gender for more than twenty years, advanced the idea that men and women are more alike than they are different. Her "gender similarities hypothesis" (2005, p. 590) was supported by

extensive evidence. Analyzing several hundred studies, she showed that the majority of them did not prove the existence of gender differences, except for some aspects of sexuality and motor behaviors. As for communication, she only found moderate differences in frequency of smiling and accuracy of spelling. She also emphasized the implications of a model that "reifies the stereotype of women as caring and nurturant and men as lacking in nurturance" (2005, p. 590) and the cost it has on relationships and work, with parents having lower expectations for their daughters in math, or heterosexual couples believing that they cannot communicate effectively. The gender differences paradigm was further deconstructed by Cameron with her 2008 book *The Myth of Mars and Venus*, in which she stated that it is "one of the great myths of our time" (2008, p. 14). Drawing on several decades of scientific research, she debunked the various claims made by what she called "the literature of Mars and Venus" (p. 14).

For academics such as Cameron or Eckert and McConnell Ginet, the fact that many studies found gender differences can be explained in part by the role of social norms. Gendered behavior, the latter state in their 2003 book *Language and Gender*, is learned, and biological differences are exaggerated. For instance, they point out that five-year-old boys start to lower their pitches while girls raise theirs at the same age, imitating men and women, even though they have the same vocal apparatus (p. 18). The same phenomenon exists at the lexical level, with women being more likely to know words relating to sewing, and men words from the domain of mechanics (p. 70). The prevalence of the differences paradigm may also be explained, according to Cameron, by the fact that researchers are sometimes tempted to interpret ambiguous evidence by using what they know about gendered communication styles, and that any claim they make thus becomes "a self-fulfilling prophecy (2003, p. 46). Most of the time, this process is not conscious, and is due to what Nickerson called the "confirmation bias", that is, the tendency to search for or to interpret information that confirms pre-existing beliefs or hypotheses. Baker (2014, p. 42) pointed out, for instance, that, because of the widespread belief that men swear more often than women, people tend to notice more when males use offensive words.

Reviewing the literature about gender differences, Baker advanced other explanations to the claims made by many studies. First, he emphasized the fact that gender difference research often based its findings on inappropriate or unconvincing sources, and that a lot of studies focused on small samples. He thus argued that many

studies overemphasized their findings (p. 20). Baker also noted that differences are observed whenever two groups of people are compared, and that contrasting the speech of a group of women with another group of women, will also produce differences, as will comparing two groups of men. He added that gender roles can also influence the outcome of studies. The problem often lies in the way research is conducted, with men's speech being often recorded in the workplace, and women's in a domestic setting. In studies examining outputs recorded in similar settings, "the amount of difference reduces and is only slightly larger than comparisons of single-sex groups". Finally, Baker insisted on the existence of idiosyncrasies and "atypical speakers" (p. 41), which are sometimes overlooked by researchers. Conducting a corpus study of the use of the words "lovely" and "fucking" in the BNC, he nuances its finding that men produced the swear word more often than women by pointing out the fact that a minority of speakers used it a lot more than the majority of men (p. 33).

2.2. Empirical Studies of Speech and Writing

Most linguistic research of gender has focused on speech rather than writing. Sociologists began to investigate the relationships between language and gender in the 1960s. At the time, they did not make the distinction between sex and gender. Sex was only one of the many variables of their studies, together with age, social class, or network, although research would show, as Trudgill wrote in 1983, that the correlation between gender and certain features of language "is the single most consistent finding to emerge from sociolinguistic work" (p. 96). Many studies have investigated the use of standard versus nonstandard features by males and females. Labov's (1966) study of speech in New York examined the use of easily quantifiable linguistic features, such as the postvocalic /r/, and showed that men are more likely to use nonstandard variants than women, and that women had a tendency to produce hypercorrect forms, especially in the lower middle class. Trudgill (1974), who studied variation in Norwich, England, also found that women tended to use standard or prestigious linguistic features. Women's tendency to reject nonstandard forms has been explained by their awareness of their lower social status; it has been argued that using standard features is for them a means to gain more power in society, and to avoid the stigma of nonstandard speech. Studies have also suggested that men may

eschew some standard features because of their historical association to a perception of "feminity" that they do not want to convey (Romaine, 2003, pp. 104-100).

Other sociolinguistic studies of speech and gender have focused on the differences between a "social" and an "informative" style. Many support the argument made by proponents of the gender differences paradigm that women tend to use language to facilitate social interaction, while men are more information- and task- focused. For instance, in 1987, Van Alphen, studying Dutch children, reported that girls tended to avoid disagreement, and boys to challenge each other. Similarly, Goodwin (1990), in her study of African American children in West Philadelphia, found that boys tended to be more aggressive when they disagreed with someone than girls. Pilkington (1992), recording all-male and all-female groups gossiping, found that "cooperative talk" dominated amongst women, and that men used less verbal feedback and expressed criticism in more direct way.

Even though writing has not been studied as extensively as speech by sociolinguists focusing on gender, studies of writing have also commented on the existence of an "involved" female style, and of an "informational" male style. Newman, Groom, Handleman, and Pennebaker (2008), analyzing a database of more than 14,000 text files including 93% of written texts, found that women used more words relating to psychological and social processes, whereas men referred more to impersonal topics and object properties. According to Olsen (2005), who conducted a diachronic corpus study of 300 literature texts written by male and female authors between the 16th and the 20th century, there is a "distinct female literary voice" (p. 147), characterized by a higher frequency of words expressing emotions and relationships. In his corpus, males, on the other hand, used more abstract words and concepts.

Identifying the gender of the author of a text was shown to be possible by Koppel, Argamon, & Shimoni (2002), who used stylometric techniques. Comparing more than a thousand lexical and syntactic features across 566 documents from the British National Corpus, they were able to correctly identify the gender of writers 80% of the time, which suggests that there are quantifiable differences between a male and a female writing style. Using a corpus study of English letters written from the 15th to the 17th century, Nevalainen (2000) examined supralocalization, the replacement of local or regional linguistic features by others having a wider

currency. The findings were contrasted; men and women were equally associated the changes of the suffix –s and the disappearance of multiple negation, while women used "you" more frequently than men, which suggests that they led the change in this instance. Even if research reveals that there are gender differences in different types of writing, it has been pointed out that there is a relative dearth of corpus-informed research, despite its vast potential (Baker, 2014, p. 20).

2.3. Language, Gender and CMC

Computer-mediated communication provides linguists with new opportunities to investigate the relationship between language and gender. At first, it was thought that the relative anonymity of the Internet would reduce the differences between gender roles, but, early on, studies showed that there was a parallel between gender distinctions offline and online (Baron, 2004, p. 405). Cushing was one of the first to research how gender interactions were translated into CMC. Grounding her research in gender theory and in Tannen's sociolinguistic research, she described the Internet as a "frontier environment" favorable to the "aggressive" linguistic style of males, which "closely mirrors the male syndrome and mutes women's voices" (1996, p. 65-69). Herring adopted the same approach; she has claimed that men and women have two recognizably different styles, and that gender distinctions are not neutralized by CMC. In her empirical study of six discussion lists, she characterized male style as adversarial; she found that men used more strong assertions, self-promotion, putdowns, sarcasm then females, and wrote lengthier messages. Female style, she argued, combines supportiveness and attenuation. Women thanked more other users, made them feel welcome, expressed doubts, used more hedges, apologized more, asked more questions and made more suggestions than males. Herring claimed that the existence of these gendered styles disproves the fact that CMC is "gender blind", and shows that, online, females and males constitute different discourse communities, with differing cultures and communicative norms and practices (Herring, 1994, p. 198).

On the whole, many empirical studies have supported Herring's and Cushing's claims. Females have been found to be more supportive and to show more personal sensitivity than males in a study of email sent to friends (Colley & Todd, 2002). Another study of emails showed that it was possible to find out the gender of senders with accuracy, since females used more modal verbs, intensifying adverbs, asked

more questions and apologized more than males, who were more likely to give opinions and use insults (Thomson & Murachver, 2001). Cheng, Chandramouli, and Subbalakshmi analyzed Reuters newsgroup messages and emails from the company Enron and found that women made more frequent use of emotionally intensive adverbs and affective adjectives, while men's conversational style used more first-person singular pronouns and directive sentences.

Some studies have also uncovered some nuances. A study of the nature of offensive conduct on five Usenet newsgroups revealed that women were more likely to thank and to apologize than men, but that women who did criticize other users were no more friendly and helpful than males (Smith, McLaughlin & Osborne, 1997). In their study of IM, Fox et al. (2007) found that, even if women sent more expressive messages than men, men did not talk more than women, and that in many ways, messages were similar, regardless of the gender of the sender (2007, p. 395). Thelwall, investigating the use of cursing in a corpus of 9,000 British and American MySpace pages, found that males swore more often than females, but that the difference was less great for younger females in the United Kingdom. His study also revealed that younger users swore more than older users, with "a disproportionate increase for females" (2008, p. 17). He linked this finding to deeper changes in gender roles in society and to the rise of the "ladette culture".

2.4. Empirical Studies of Netspeak Features and Gender

Several studies have investigated the relation between emoticon usage and gender. Differences were not always found; Thompson & Filik (2016) conducted two experiments at the University of Glasgow; they presented students with written conversations and asked them to produce the final message of the conversation by making their intention clear. The aim of the researchers was to investigate to what extent emoticons acted as markers of sarcastic intent. Their results revealed that gender had no effect on the likelihood of producing emoticons (p. 111). Most studies have however suggested a possible relation between gender and use of emoticons. A study of teenage blogs revealed that males used more emoticons than females, and had a greater tendency to produce "sad" and"playful" emoticons, such as :p, referred to by the authors as "flirty" (Huffaker & Calvert, 2005). Tossell et al. have also observed that, if men do use fewer emoticons in terms of frequency, they use a more diverse range of emoticons (2012, p. 659). The majority of findings, on the other

hand, have reported women to be more frequent emoticon users than men (Witmer & Katzman, 1997; Wolf, 2000; Baron, 2004; Fox et al., 2007; Tossell et al., 2012). For instance, in a study of text messages sent by students, Lyddy et al. (2014) found that 81% of emoticons were sent by women.

In mixed-gender settings, gender differences have been found to be less significant. In Lee (2003), males rarely used emoticons when sending messages to other males, and did use them when interacting with females. Females, on the other hand, used an equal amount with males and with other females. This finding is congruent with Wolf's study of newsgroups; in her corpus, males used more emoticons in mixed-gender newsgroups, while female users produced fewer emoticons than in same gender newsgroups. (2000, p. 831).

If previous research has shown a clear tendency for females to use more emoticons, the situation is more contrasted in the case of other Netspeak lexical items. Some studies support Labov's and other sociolinguists' findings that women are more likely to produced standard forms in speech and writing. This is the case, for instance, of a study of email communication in Dutch elementary schools (Van Der Meij, 2007). Girls' emails were described to be more elaborate than boys' and to contain more words and more clauses, which the author suggested "may reflect a difference in language proficiency" (p. 351). Squires, in her empirical study of the use of apostrophes in instant messaging, observed that females included apostrophes more frequently than males (2007, p. 11). Similarly, a study has found that males used a higher percentage of contracted forms than females, and also suggested that females had "a greater tendency toward treating IM as a written medium" (Baron, 2004, 418). By contrast, another study revealed that women use more nonstandard typography and orthography than men (Herring & Zelenkauskaite, 2009). The outputs investigated by these two studies were however very different; the first examined IM conversations among college students in the United States, and the other text messages posted to a television program in Italy. On the other hand, Peersman et al. (2016), who analyzed an 18-million word corpus of Flemish Dutch chatspeak, did not find any effect of gender on the likelihood of producing Netspeak features such as abbreviations, acronyms, omission of letters or words, and character flooding (p. 19).

3. Reddit: One of the Largest Online Communities.

One of the most popular web forums, especially in the United States, Reddit provides a large amount of data, which is very easy to access and navigate, and thus seemed a good source of content for a study of Netspeak.

3.1 Overview of Reddit

Reddit was founded in 2005 by Alexis Ohanian and Steve Huffman, who met at the University of Virginia. They developed the site to make it easier for Internet users to find interesting information online. According to the site's Frequently Asked Questions page, its name derives from the play on words "I read it on reddit" (Reddit FAQ, n. d.). The site quickly gained popularity and attracted investors. In January 2006, Reddit merged with the company Infogami before being acquired by Condé Nast Publications in October 2006. By this time, five hundred thousand users were visiting the site every day (Caffrey, 2016). Reddit's user base rapidly grew; in 2010, the site averaged 12 million monthly active users and, in 2015, it boasted 170 million monthly regular users (Isaac, 2015).

Self-proclaimed "front page of the internet", Reddit is a community-driven platform. Designed to be a news sharing site, it has evolved to include other topics, such as entertainment and advice giving. On Reddit, users, who are called "Redditors", a blend of "Reddit" and "editor", can share links to articles, videos and pictures, or submit texts, called "self-posts". They can also comment on the items submitted, and are able to "upvote" or "downvote" them. The posts that obtain the most votes are featured higher up on the front page of the site, which is constantly updated to feature the most popular content. On any given day, Reddit's front page might feature links to videos of sloths or of people shooting guns, detail recipes, discuss news articles, or try and answer questions such as "Why are there 'early birds' and 'night owls"? or "How is orange juice economically viable?" The voting system is also used in the comment sections of each post, the most popular comments appearing directly beneath a post.

Reddit is organized into smaller communities or sub-forums, called "subreddits", which are always titled using the prefix /r/. According to the website Reddit Metrics, there were, as of May 7, 2016, 857,444 subreddits (Reddit Metrics, n.d.). Hundreds of subreddits are created everyday (History, n.d.). Subreddits deal with all sorts of topics, ranging from the serious, like /r/science, /r/worldnews or /r/AskAnthropology,

to the bizarre, with subreddits about the right of toasters or about photographs of a famous politician eating sandwiches. With more than 11 million subscribers, /r/IamA is one of the most popular subreddits (Reddit Metrics, n. d.). Celebrities or people with an unusual story or skill use it to answer Redditors' questions. Apple co-founder Steve Wozniak, astronaut Scott Kelly, actor Patrick Stewart, and, most famously, President Obama in August 2012, have all used Reddit as a platform to interact with the Internet crowd.

On Reddit, users have a great degree of freedom**.** Not only can they post comment and contribute to decide which items are the most popular, but they can also create subreddits. Setting up an account is an easy and entirely anonymous process, since no email address is required. This feature may have contributed to the success of the site, which is, in 2016, one of the largest online communities according to Alexa, a company specializing in web traffic data. Even if the accuracy of Alexa measurements has been disputed, since it only takes into account a small portion of Internet users (Gibbs, 2006), they are still useful to estimate the success of a website. In 2016, according to its global Alexa rank, Reddit is the ninth most visited website in the United States, behind Google.com, Facebook.com, YouTube.com, Amazon.com, Yahoo.com, Wikipedia.com, Ebay.com and Twitter.com. On a global scale, Reddit was ranked #28 in May 2016. Alexa also provides statistics on the countries of origins of visitors, which indicate that more than half of Reddit's visitors are located in the United States. The United Kingdom and India come next, each representing a little more than 5% of Reddit's visitors, while Canada and Australia account respectively for 4.3% and 2.4% of the visitors to the site (Site overview, n.d.).

A study published in February 2016 by the Pew Research Center, a think tank based in Washington, reveals more information about Reddit users (Barthel, Stocking, Holcomb, & Mitchell, 2016). The researchers, who wanted to understand the news dynamic on the site by focusing on the 2016 presidential campaign, surveyed users in order to estimate the age and gender makeup of Reddit. They found that about 7% of US adults use Reddit, and that Redditors are "more likely to be male, young and digital in their news preferences". Fifty-eight percent of Redditors, the authors reported, are aged 18 to 29, 33% are aged 30 to 49, and 97% of them go online every day (p. 7). In 2013, there were, in the United States, twice as many men Redditors as there were women (Duggan & Smith, 2013). In 2016, men still

dominate the site, even though the gender gap has narrowed a little bit, with 69% of Redditors being male, and 31% female (Barthel et al., 2016, p. 7). This type of division by gender is not unique to Reddit; it has been shown to exist in other online discussion forums such as Digg or Slashdot (Anderson, 2015, para. 3), which are especially popular among men.

Reddit has been known for allowing misogynistic and racist content. For years, it was based on a model which supported complete freedom of speech, and which only relied on users and moderators of the various subreddits for its regulation. The fact that users are not required to indicate their email address to create an account, which makes for a very low barrier of entry, has also been seen a factor which negatively impacts discourse (Ovadia, 2015, p. 37). Reddit has thus been surrounded by a number of controversies over the years. It is home to "The Chimpire", a network of several dozens of hyper-racist subreddits, several of which have been banned, and to the subreddit /r/MensRights, which has been described as a "hate group" by the American non-profit advocacy organization Southern Poverty Law Center (Hankes, 2015). A subreddit played a key role in the leaking of photographs of nude celebrities, known as the Fappening, while another subreddit wrongly identified a number of people as suspects following the Boston Marathon bombings. The controversies surrounding Reddit contributed to bring about a crisis in management (Dewey, 2014; Valdes, 2013). Reddit's policy of allowing all speech on its site (Dewey, 2015) changed in June 2015 when Steve Hoffman, the co-founder of the site, came back as the CEO of the company after a six-year absence (Isaac, 2015). He quickly announced a new content policy, which banned the posting of personal information, child pornography, spam, illegal activity, and harassment (Weinberger, 2015). This new set of rules provoked a backlash from many Redditors (Caffrey, 2016).

3.2. Describing Reddit with Herring's Faceted Classification

Herring's "faceted classification" (2007) was applied to the site. Her instructions were taken into account; she explained that the categories of her model are open-ended and that additional factors can be added, and recommended choosing only the most important features of an output, in order for the model not to seem too "verbose" (2007, para. 69). However, all facets proved to be useful to give a clearer

picture of Reddit, and to better understand how it differs from other forums and social networking websites.

Technological features

- Synchronicity of participation: communication is asynchronous, and users do not have to be logged on at the same time to send and read messages, by contrast with live chat, for instance.
- Granularity of the unit transmitted: messages are posted in their entirety, and not line by line or character by character.
- Persistence of transcript: comments remain on the system indefinitely after they have been posted, although they can be removed by the moderators of the various subreddits.
- Size of message buffer, or length: by default, the comment limit is 10,000 words, although some subreddits allow longer comments (FAQ Writing Prompts, n. d.). However, most comments are a lot shorter and do not reach the 10,000 limit.
- Channels of communication: most content on Reddit is text, but users can link to other sites, to videos and to images.
- Identity: anonymity is the rule on Reddit. Users use pseudonyms, and do not need an email address to register.
- Message format: comments do not necessarily appear chronologically on a thread page. Their order is determined by the number of upvotes and downvotes they get, the most popular comments appearing at the top of the page. The Reddit Enhancement Suite add-on, which was developed independently from the site, allows users to customize the way comments are displayed by showing the newest comments at the top or by hiding "child comments" that do not directly reply to the opening message.
- Quoting: it is possible to quote the entirety or part of a comment posted by another user, although this practice does not seem as widespread on Reddit as on some other forums.

Social factors

- Participation structure: most subreddits are public, but some are private and Redditors need to meet certain criteria to be admitted.

- Participant characteristics: gender, country of origin and socio-cultural background are unknown in the vast majority of cases. Reddit is not a social network like Facebook or Twitter, where users can post pictures and information about themselves in their profile. Reddit's interface is minimalistic. Users' profiles only contain a few pieces of information, apart from the comments they contributed to the site: the amount of time they have been Redditors, their link and comment "karma", which reflects the popularity of their contributions, or their "trophies", which are awarded by some Reddit communities. The only way to gather information about Redditors is to investigate what they say about themselves, or to look at the Reddit "flair" that is used in some subreddits.
- Purpose: Reddit was created for users to share and discuss news stories. Its purpose has somehow shifted, and Redditors now ask for advice and share personal stories.
- Topic: each subreddit focuses on a specific topic. Moderators can remove comments that are off topic.
- Tone: language is informal. Profanity, abbreviations and Internet slang are used.
- Norms or organization: all subreddits have moderators. Moderators can change the settings of subreddits, edits the rules, ban users from submitting and remove comments (Moderation, 2016).
- Norms of social appropriateness: often criticized for its sometimes violent, misogynistic and racist content, Reddit has created a "Reddiquette", a set of guidelines that express the values of many Redditors. It includes rules such as "Don't insult the other", "Don't troll", or "Remember the human". In addition to these guidelines, the various subreddits have their own rules, which are enforced by the moderators (Rediquette, 2015).
- Norms of language: English is the most widely used language on Reddit, most users being from the United States and from other English-speaking countries (Site overview: Reddit.com, n. d.)

3.3. Reddit: A Literature Review

Over the past five years, Reddit has attracted the attention of researchers in different fields. The first studies of the site seem to date back to 2011; they focused on Reddit's voting system (Van Mieghem, 2011; Mills, 2011), used content analysis to determine the characteristics of top-ranked articles on the site (Wasike, 2011), and explored the community's approach to online identity (Bergstrom, 2011).

Many papers originated from the field of computer science; they either sought to understand the mechanisms that characterize Reddit, or used the site as a source of data. Olson and Neal (2013) used the site to demonstrate how they were able to map the primary interest of a social network. In a short paper, Singer, Flöck, Meinhart, Zeitfogel and Strohmaier (2014) examined how submissions to Reddit have evolved over the course of five years by analyzing voting and commenting patterns, while Haralabopoulosa, Anagnostopoulosa, and Zeadallyb (2015) focused on the diffusion and lifespan of information flows on Reddit. Zhang and Setty (2016) developed a method to identify sub-topics in Reddit's comments. Kiene, Monroy-Hernandez and Hill (2016) showed how a Reddit community was able to handle a massive flux of newcomers. Group identity, social feedback and discussions about health have also been studied by researchers (Howard & Magee, 2013; Wang et al., 2015; Cunha, Weber, Haddadi & Papa, 2016), as well as trolling (Mikkelsen, Jensen, Tarding, Haasum & Rasmussen, 2016) and controversies (Centivany & Glushko, 2015). Other papers have investigated the use of Reddit as a learning tool (Ovadia, 2015; Gibeault, 2016).

Reddit has attracted the attention of computational linguists; for instance, Al Rfou et al. (2016), computational linguists at Google, used 2.1 billion Reddit messages to devise a conversational system, and in their working paper Danish and Talukdar (2015) proposed to conduct an empirical analysis in order to identify underlying response-eliciting questions, using data from Reddit. Noguti (2016) used statistical models to uncover relationships between posts language and user engagement. McEwan (2016) investigated the linguistic markers characteristic of the most active and interactive online communities, using fifteen subreddits, and a linguistic analysis was conducted to investigate reply bias on a subreddit (Huang and Bashir, 2016). One study reports on the building of a German corpus from Reddit comments (Barberesi 2015).

Despite the recent flourish of studies and working papers about the site, research about Reddit is still in its infancy. The accelerated rate of publication in 2015 an 2016 suggests that a lot more studies are yet to come. As of this writing, no study seems however to have investigated the lexical characteristics of Reddit, or used the website to verify the existence of gender differences in CMC.

II. Methods

1. Criteria Used to Build the Corpus

In order to build a corpus that would suit the needs of this study, the following set of criteria was established.

1. The corpus would need to be on the bigger side, since the scarcity of Netspeak features has been noted by many researchers (Witmer & Katzman, 1997; Crystal; Hancock, 2004; Baron, 2004; Tagliamonte & Denis, 2008). A two-million word corpus was used as a goal.

2. Only comments that were produced in 2015 and 2016 would be compiled, in order to have a screenshot of the language of Reddit at a given time, since, on the Internet, language change can happen quickly, and since Netspeak has been described as "transient" (Crystal, 2004, p. viii).

3. Only content created by Redditors who clearly specified their gender would be used.

4. The corpus would be composed of two balanced subcorpora. These two subcorpora would each contain about as much data, produced by the same number of contributors. In order to reach the two-million word goal, it was decided to compile 10,000 words worth of content produced by 200 Reddit users, 100 males and 100 females.

Criteria 1. and 2. proved not to be difficult to apply, since Reddit is a huge source of data, easy to access and to navigate. Comments are not precisely dated, but the amount of time that has passed since they were posted is always specified; messages are prefaced with the user name of the Redditor, the number of votes that they gathered, and an expression such as "1 month ago", "1 week ago" or "7 hours ago".

Criterion 3. was more problematic, since Redditors use pseudonyms. In the vast majority of cases, these pseudonyms do not reveal the gender of the users, as is shown by the screen names *directconnection*, *BoostJunky87*, *sudomv, workingtimeaccount* or *syuk,* randomly taken from the /r/news subreddit. Some pseudonyms do reference gender, but they still can be ambiguous or misleading. It was thus necessary to find another source of information about users. What is called "flair", in Reddit lingo, proved to be extremely useful. "Flair" is a short text or symbol that is displayed next to a username. It is not used by all subreddits, and is subreddit specific. A Reddit user who comments on several sub-forums can have

different flairs, but only one flair will be displayed at a time, depending on the subreddit they comment on. In some subreddits, flair is assigned by moderators, for instance to reward Redditors who contribute often. In other cases, subscribers can choose their own flair, while always following rules set by the moderators. For instance, /r/WeddingPlanning subscribers can specify the date and the location of their wedding and /r/GirlGamers members the name of their favorite gaming platform.

A very small number of subreddits include information about gender in their flairs. This is for example the case of /r/AskMen and /r/AskWomen, which give the possibility to their subscribers to use gender symbols to specify their gender, as shown in Figure 1. To do so, both chose the Mars and the Venus symbols; in addition to these, r/AskWomen uses symbols for transgender and gender neutral. It was decided to choose Internet users in /r/AskMen and /r/AskWomen, which, with respectively 241,044 and 242,896 subscribers, ranked 168th and 164th on the most popular subreddits list according to the Reddit Metric website in May 2016. Choosing Reddit users in this way did not mean, of course, that only content they contributed to the /r/AskMen and /r/AskWomen was included, since most Redditors comment on several subreddits.

Building balanced subcorpora did not prove to be a problem, even if it was time-consuming. When clicking on a username, it is possible to scroll down a page to view all the comments written by a given user; content was selected in a way that would allow to achieve a 10,000-word word count per user.

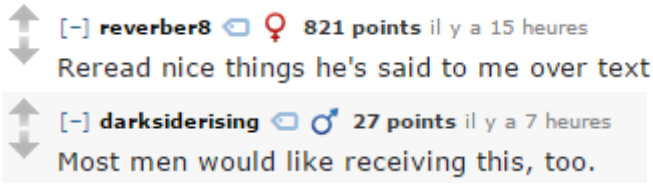
2. Compiling the Corpus

The corpus was compiled manually. Using a site ripper such as HTTrack was briefly considered. However, since the data had to be carefully selected, copying Reddit pages and pasting them into a word processor seemed simpler. A hundred female and 100 male Reddit users were chosen in the /r/AskMen and /r/AskWomen subreddits. Male Redditors were chosen in the /r/AskMen subreddit, and females were picked in the /r/AskWomen subreddit. Only Reddit users who had contributed at least 10,000 words worth of content to the site were included in the corpus. As many files as they were Reddit users were created. They were named F_1 to F_100 for female Redditors, and M_1 to M_100 for male users.

Figure 1. Examples of flair in r/AskMen and r/AskWomen

r/AskWomen

r/AskMen



Once pasted into a word processor, the materials had to be prepared because they contained a number of unwanted items. Boilerplate text was removed by using the "search and replace" function in Word. Pictures were deleted. Names of users, dates of comments and other types of data were not erased, in order to keep as much information as possible on each comment. They were tagged, so that they would not appear during the analysis of the corpus. Quotes of previous messages were also deleted; they were not a lot of them, and they were easy to spot because of the different font color used. Then, the .doc files were converted to .text files, since it is the format supported by AntConc, the software that was used to analyze the corpus. A word list search was performed on each document individually, in order to reach the word count desired. Word counts between 9,000 and 11,000 were considered acceptable. If the document was too short or too long, content was removed or added. When all 200 files were compiled, the final word count of each same-gender corpus was adjusted by removing content, until the two corpora were about the same size. The final make-up of the corpus is synthesized by Table 1. The table does not indicate the number of comments retrieved, since, because of the layout of Reddit user pages, comments would have had to be counted manually. Length of comments is highly variable on Reddit.

Table 1. The Reddit corpus

	Female sub-corpus	Male sub-corpus	Total
Number of users	100	100	200
Word count	1,009,996	1,010,984	2,020,980

3. A Corpus-Based Study of Reddit

Conducting a corpus-based study means establishing a list of the items to investigate. In the case of emoticons, the task was challenging, due to the sheer number of emoticons. In his *Glossary of Netspeak and Textspeak,* Crystal's list of emoticons contains several hundred items; the list was non-exhaustive and, with the rapid growth of CMC, is bound to be obsolete or incomplete. Using findings of previous studies as a starting-point was considered. However, many studies analyzed smaller corpora and did not always report a great diversity of emoticons. (Wolf, 2000, Garrison et al., 2011; Baron, 2004; Thompson and Filik, 2016). The lists of emoticons they provide are thus limited; Garrison et al. list for example 19 items, while Baron (2004) found 49 tokens and 7 types. Thompson and Filik (2016) found 50 and 72 different emoticons in the two experiments they conducted, but did not provide a list for them. Since it is very likely that a 2-million word corpus contains a lot more emoticons and a greater variety of items, taking past findings as a starting-point would probably give only an incomplete picture of the use of emoticons in the Reddit corpus. Park et al. (2013) found a much larger number of emoticons in their corpus of 1.1 billion tweets, but only provided a list of the 23 items (p. 470).

The goal was thus to find another resource. Researching the web for non-academic lists of emoticons and websites only confirmed the sheer diversity of emoticons. The Cool Smileys website, for instance, lists 681 emoticons (List of text emoticons, n. d.). The Wikipedia page "List of emoticons" was chosen, as it seemed to provide a large variety of items, while still being of a manageable size for analysis. The Wikipedia page provides three tables of emoticons. Only the Western list was used, and not the "Eastern" or the "*2channel*" lists, since Park et al. (2013) reported, in their study of emoticon use in fifteen countries, that the Eastern style of emoticon in only marginally used in the US and in other English-speaking countries (p. 470). Prior to finalizing the list, searches were conducted in the Reddit corpus in order to check if Eastern emoticons, or horizontal emoticons, were used; the most frequent

Eastern emoticons of Park et al. (2013) corpus, ^^ = ; ^-^ -__- -.- , were only found a handful of times. The 2channel emoticons are characteristic of a Japanese textboard and thus were not included in the emoticon list either.

The Wikipedia list classifies emoticons by "families", and indicates meanings for each of them**.** It was slightly adapted; emoticons were further classified into logical categories. The final list, seen in Table 2, features a total of 156 emoticons, split into 34 families. The potential meanings are those proposed by the Wikipedia list.

Other lexical Netspeak features presented the same types of challenges as emoticons. The goal was to build a list that would not be exhaustive, of course, but as representative as possible of the use of Netspeak on Reddit. Past research did not provide extensive lists of items, and it was decided, once again, to use Wikipedia as a starting point, with its inventory of "English Internet slang" terms. Other lists of acronyms were considered as well, but they often contained too many items to make the search process manageable. The Netlingo page about chat abbreviations and acronyms, for instance, lists 2,500 items.

However, because it seems that Reddit has its own language, derived from the name of some popular subreddits, it was felt that the Wikipedia list, not being focused on Reddit or on Internet forums, may not feature some prevalent lexical items of Reddit. Google searches were also performed with keywords such as "Reddit slang, "Reddit lingo" and "Internet slang", in order to find lists of items. In the end, a variety of sources were used to supplement the Wikipedia list, including "The TL; DR Guide to Reddit Lingo", published in 2014 on Mashable.com, a list of "30 trendy Internet slang words and acronyms", and articles published on *The Toast*, *The Guardian* and on *The Atlantic* websites. Of course, these sources are not academic; they do not provide any information about frequency, and are only based on the intuitions of their authors. The goal, in the absence of serious research about the lexical characteristics of the language of Reddit, was to compile a list that would be as up-to-date as possible, and that would enable the corpus study to give a clear picture of the Netspeak phenomenon on Reddit in spite of the fact that it would not be exhaustive.

Table 2
List of emoticons with their potential meanings, adapted from Wikipedia

Emoticon	Potential meaning	Variants
:)	"happy face"	:-) :-)) :-) :o) :] :3 :) :> =] 8) =) :} :^) :ↄ)
:D	"laughing, big smile"	:-D 8-D 8D x-D XD =-D =D =-3 =3 B^D
:(	"sad face, frown"	>:[:-(:-c :c :-< :ↄC :< :-[:[:{
;)	"wink"	;(;-) *-) *) ;-] ;] ;D ;^) :-,
:p	"playful"	>:P x-p XP :-p =p :-Þ :Þ :þ :-þ :-b :b d:
:-||	"angry"	:@ >:(
:'-(	"crying"	:'(
:'-)	"tears of happiness"	:')
D:<	"horror, disgust"	D: D8 D; D= DX v.v D-':
O_o	"surprise"	>:O :-O :O :-o :o 8-0 O_O o-o o_O o_o O-O
:*	"kiss"	:-* :^* ('}{')
:/	"skeptical, annoyed"	>:\ >:/ :-/ :-. :\ =/ =\ :L =L :S >.<
:|	"straight face, no expression"	:-|
:$	"embarrassed"	
:X	"sealed lips"	:-X :-# :#
0:)	"angel"	O:-) 0:-3 0:3 0:-) 0;^)
>:)	"evil, devilish"	>;) >:-) }:-) }:) 3:-) 3:)
o/\o	"High five"	^5 >_>^ ^<_<
|;-)	"Cool"	
|-O	"Bored"	
:-J	"tongue-in-cheek"	
:-&	"tongue-tied"	:&
#-)	"partied all night"	
%-)	"drunk"	%)
:-###..	"being sick"	:###..
<:-|	"dumb"	
□_□	"look of disapproval"	
(͡° ͜ʖ ͡°)	"Le Lenny Face"	(͡°)
<*)))-{	"something's fishy"	><(((*> ><>
\o/	"cheer"	*\0/*
@}-;-'---	"rose "	@>--->--
<3	"heart"	
</3	"broken heart "	
	Characters	
~(_8^(I)	"Homer Simpson"	
5:-)	"Elvis Presley "	~:-\
//0-0\\	"John Lennon"	
*<|:-)	"Santa Claus"	
=:o]	"Bill Clinton"	
,:-) 7	"Ronald Reagan"	:^]

Eleven items were thus added to the Wikipedia list. A first set of features that are, according to a source (Koerber, 2014) frequently used on Reddit, or are typical of the site, contained acronyms such as *AMA* "ask me anything", *DAE* "does anyone else", *ELI5* "explain to me like I'm five", *TIL* "today I learned", *OC* "original content", *SO* "significant other", and *ITT*, "in this thread". A second set of items includes terms that seem, according to the authors of the sources used, emblematic of Internet slang; the terms and phrases *bae*, *I can't even*, *This*., and *All the feels* and the construction *because* + noun were included.

One last item, which was noticed while compiling the corpus, was also added: */s*, which was, at first, thought to be an emoticon. According to a Reddit users, */s* stands for sarcasm and was inspired by HTML tags (What does /s mean?, 2014). Since it was noticed randomly many times during the building of the corpus, it was decided to add it to the list. Of course, this means that it is extremely likely that other frequent Netspeak items were omitted. The discussion of the results takes this limitation into account.

The Reddit list was also modified. Terms used to talk about the Internet, like *bot*, *crawl* or *homepage*, were removed, as well gaming terms like *MMORGP* or *PL*, "Powerleveling" and non-English items, such as *IZI*, which is French according to Wikipedia. Even though it is sometimes difficult to make the distinction between Netspeak and items that are used in other communication situations, some items, such as *BBQ* "barbecue", were removed, because they did not feel specific to CMC. The features compiled were distributed in several categories: non-standard spellings and letter homophones; clippings; Netspeak words and phrases; and acronyms and initialisms. This last category is the largest, and contains 233 of the 270 items of the finalized list. The list was also modified during the research process. The search for the acronym *GF*, for instance, revealed that it was used in the corpus, but not with the "good fight" meaning indicated by Wikipedia. In the corpus, it stood for "girlfriend". Therefore, the meaning was adapted, and the acronym *BF* "boyfriend, was added to the list. Table 3 shows a few examples of the Netspeak lexical features that were included to the list. The complete list features in appendix.

Table 3. List of some of the Netspeak lexical features that were searched for

Non-standard spellings and letter homophones (alpha-numerical substitutions)
2 “too” or “to”; 4” for; 10q “thank you”; 10x “thanks”; A C? “AH! Si?” B4 “before”; CU “see you”; F9 “fine”; KTHX “OK, thanks”; kthxbye “OK, thanks, goodbye”; k or kk “OK”; l8r “later”; M8 “mate”; O RLY “Oh, really?”; OIC “Oh, I see”; n0rp “porn”; NE1 “anyone”; pr0n “porn”; pwnd/pwn intentional misspelling of “owned”; U “you”;
Clippings/abbreviations
Dafuq “what the fuck”; gratz “congratulations”
Phrases/words
Bae “babe”;Facepalm “sign of frustration or agitation”; Headdesk “extreme frustration” ONOZ “Oh, no”; Lulz, corruption of lol; Rehi “Hello again”; I can’t even “I am overwhelmed with emotion, I am unable to face the situation”; This. “affirmation of agreement”; because (+ noun); All the feels “strong emotional response”
Acronyms and abbreviations
AAF “as a friend” ADAD “another day another dollar”; “ADIH “another day in hell”; ADIP “another day in paradise”; AEAP “As Early As Possible”; AFAICR “as far as I can recall / remember”; “AFAICS: As far as I can see; AFAICT “as far as I can tell”; AFAIK “as far as I know”; AFAIR “as far as I remember”; AFAP “as far as possible”; “AFK: Away from keyboard” DM “direct message”; DND “do not disturb”; IHT “I had to”; IONO “I don't know”; IIRC “if I recall / remember correctly”; IIUC “if I understand correctly”; ILY I love you”; IMAO “in my arrogant opinion”; IMO “in my opinion”; IMHO “in my humble / honest opinion”; IMNSHO “in my not so humble opinion”; IOW “in other words”; IRL “in real life”; ISTM It seems to me”; ITT “in this thread”; ITYM” I think you mean”; TIMWTK “this inquiring mind wants to know”; TL;DR “too Long; Didn't Read”; TMI: “too much information”; TTYL “talk to you later”; TTYN “talk to you never”; UTFSE “use the fucking search engine”; UGO “you got owned”; URS “you Really Suck”; , W/ or W/O: With or without”; WB “welcome back”; WTB “want to buy”; WTF “what the fuck”; WTG “way to go”; WTH “what the hell”

Searches were performed with AntConc in each sub-corpus to look for all of the items in the two lists. Results were checked in order to remove false positives, such as the conjunction “so” instead of the acronym *SO*. Overall frequency of each item was taken into account, as well as the dispersion of results among users.

III. Results

1. Emoticons

1.1. Overall Emoticon Use

Searches were performed to look for the 156 items on the emoticon list adapted from Wikipedia. Overall, 2,430 emoticons were used in the Reddit corpus, which represents 0.12% of the whole corpus, or 1,202 emoticons per million words. Reddit users produced 43 different emoticons. The 20 most frequent are shown in Figure 2. There is a strong imbalance in the way emoticons are used; the ten most frequent emoticons, presented in Table 4, account for 91.8% of emoticon use in the whole corpus. Only six of them, :), :(, :D, ;), :p, and :/, were produced by more than 20% of Redditors. Thirty-seven emoticons were used by less than 7% of Redditors. Most emoticons produced belong to the "midget smiley" category. Only 9 emoticons with a "nose" element were used in the corpus.

Figure 2. List of the twenty most frequent emoticons

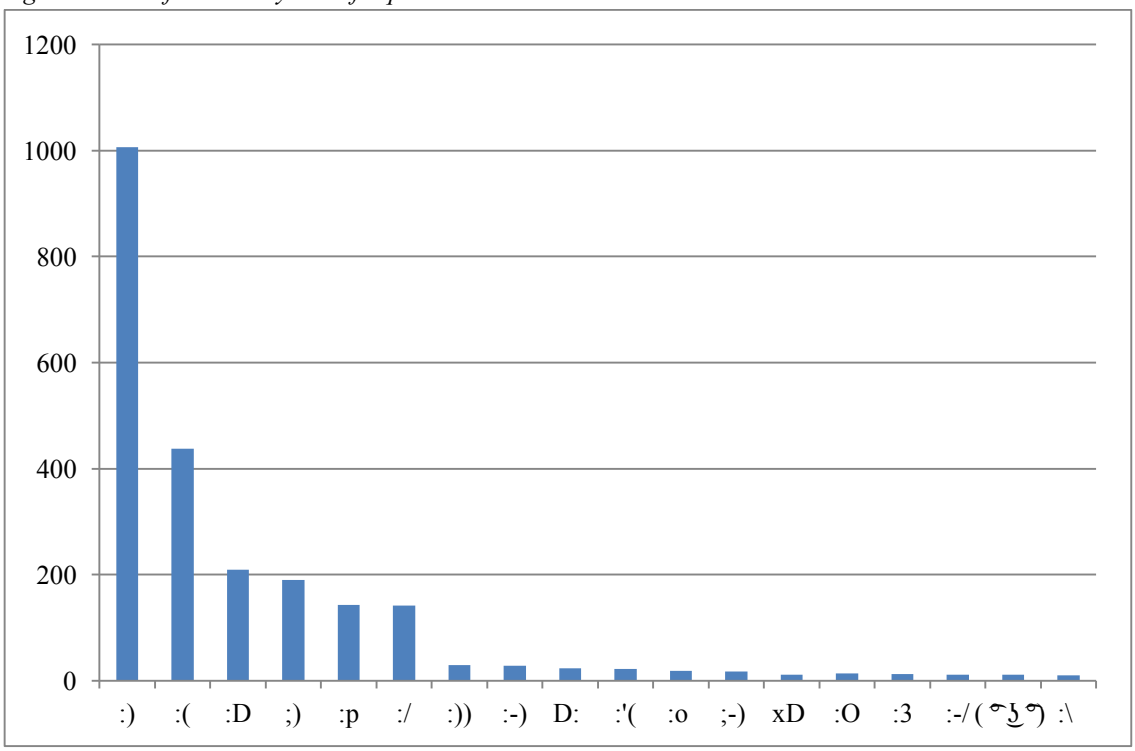


Table 4. The ten most frequent emoticons

Emoticon type	Frequency	Percentage of users
:)	1,006	58.5%
:(	437	58%
:D	209	27.5%
;)	190	32.5%
:p	143	33%
:/	142	23.5%
:))	30	6%
:-)	28	3%
D:	23	5%
:'(	22	5.5%
Total	2,230	

1.2. Emoticon Use and Gender

The male and the female sub-corpora contain about the same number of emoticon types. Thirty-six different emoticons were produced in the male corpus, and 35 in the female corpus, as shown in Table 5. There is, however, a great disparity between the number of tokens used. Male Redditors produced 689 tokens. Female Redditors used more than twice as many, with a total of 1,741 emoticons.

Table 5. Distribution of emoticons in each sub-corpus

	Male	*Female*	*Percentage difference*
Types	36	35	2.8%
Tokens	689	1,741	86.6%

Even if the number of emoticon types in each corpus is remarkably similar, the lists of emoticons produced differ. Twenty-five types of emoticons are present in both sub-corpora. The most frequent emoticons are also the most similar. Out of the ten most frequent emoticons in each sub-corpus, six are the same: the “smiley face” :), the “sad face” :(, the “laughing” emoticon :D, the “wink” emoticon” ;), the “playful” emoticon :p and the “annoyed” emoticon :/. The order is which these emoticons appear is also similar in both corpora, with the exception of the :p and the :/ emoticons, which are respectively the fifth and sixth most frequent emoticons in the female corpus, and the sixth and fifth most frequent in the male corpus.

Perhaps the most unexpected emoticon to appear in the male list is the “Le Lenny Face”, or (͡° ͜ʖ ͡°), which differs considerably from the other emoticons in the way it is created. A “Le Lenny face” do not only use Basic Latin ASCII characters,

but also characters from other character sets such as the degree sign ° from the Latin extensions, or the inverted glottal stop ʖ from the IPA extensions. It should also be noted that it was used exclusively by male Redditors, although it only appears 11 times.

The distribution of emoticons, shown in Table 6, is very unequal in both sub-corpora. In the female corpus, the ten most frequent emoticons represent 94.2% of overall emoticon use, while in the male corpus, they account for 88.7% of emoticon production. Twenty-four emotions were produced ten times or less in the female corpus, and 26 emoticons had a frequency of 10 or less in the male corpus. Not only did female Redditors use more emoticons than males, but more females used emoticons than males. The basic smiley face was produced by 72% of female Redditors, and by only 45% of male Redditors; its frequency is more than three times higher in the female than in the male corpus. The "sad face" emoticon :(was also used three times more often by females; it was produced 329 in the female corpus, versus 108 in the male corpus. It was used by 76% of females and 40% of males. The "laughing" emoticon :D, which ranks third on both lists, was used almost twice as many times by females as by males, and by almost twice as many users. The "winking" emoticon ;) has a frequency of ;) in the female corpus, and of 70 in the male corpus, although it was employed by about the same number of Redditors in both corpora. The "playful" emoticon :p was used 116 times in the female corpus and 27 times in the male corpus.

1.3. Individual differences

Table 7 shows the number of users who employed a number of emoticons in a given range. It reveals some striking differences. Only 3% of female Redditors did not use any emoticon, and 24% of males produced no emoticons. Nineteen percent of females used 21 or more emoticons while only 8% of males did. No male produced more than 50 emoticons, but 4 female users did. The ten most prolific female Redditors produced 645 tokens, or 37.05% of the overall emoticon use. The top five users produced 427 emoticons, or 22.53% of emoticon use. Distribution was even more skewed in the male corpus, with ten Redditors producing almost half (44.27%) of all emoticons, and the five most prolific users being responsible for 27.58% of overall emoticon use. In spite of these discrepancies, it should be noted that around half of the Redditors in each corpus used between 1 and 10 emoticons: 47% of the

female group and 54% of the male group. These groups produced about the same total number of emoticons, 205 in the male corpus and 212 in the female corpus. Thus, about half of the Redditors behaved in the same way.

Table 6. The ten most frequent emoticons in each corpus

Female corpus

Emoticon type	Frequency in corpus	Proportion of total emoticon use	Percentage of users
:)	783	44.97%	72%
:(	329	18.90%	76%
:D	138	7.93%	36%
;)	120	6.89%	34%
:p	116	6.66%	32%
:/	87	5%	30%
:))	21	1.21%	3%
D:	17	0.98%	7%
:o	16	0,92%	10%
xD	13	0.75%	6%
Total	1,640	94.21%	

Male corpus

Emoticon type	Frequency	Proportion of total emoticon use	% of Redditors who used the emoticon
:)	223	32.36%	45%
:(	108	15.67%	40%
:D	71	10.30%	19%
;)	70	10.16%	31%
:/	55	7.98%	24%
:p	27	3.92%	17%
:-)	21	3.05%	3%
:'(	14	2.03%	4%
;-)	11	1.60%	5%
(͡° ͜ʖ ͡°)	11	1.60%	8%
Total	611	88.77%	

Table 7. Frequency of emoticon use in the two sub-corpora

Number of emoticons used	Female users	Male users
0	3	24
1-10	47	54
11 -20	17	14
21-30	14	3
31-40	11	4
41-50	3	1
51-60	0	0
61-70	2	0
71-80	1	0
81-90	0	0
91 or more	2	0

1.4. Comparison of Two Smaller Corpora

Two sets of smaller corpora were created in order to see how great the differences between the two female corpora and the two male corpora were. This analysis was conducted, following Baker's caution that any comparison of two groups of speakers will produce differences (2014, p. 41), to check if contrasting two same-gender corpora would yield a significant less difference. If it was the case, it would give more weight to the results of the first analysis, and contribute to the argument that females and males do use emoticons in a significantly different way. Each corpus was randomly divided into two corpora; one corpus was created with the contributions of female users numbered 1 to 50, and another with contributions of users 51 to 100. The same was done with male Redditors. The results of this analysis are summarized in Table 8. They show that there are indeed differences between each set of corpora. There is a 20.9% percentage difference between the two male corpora, and a 52.7% difference between the two female corpora. This is a significant difference, even though it is still less great than the 86.6% percentage difference found between the male and the female corpus.

Table 8. Total number of emoticon used in the two male and the two female sub-corpora

	Sub-corpus 1	Sub-corpus 2	Percentage difference in frequency
Female	1,100	641	52.7%
Male	380	309	20.9%

2. Netspeak Lexical Items

2.1. Overall Use of Acronyms and Other Netspeak Features

The results obtained when comparing the overall use of Netspeak features in both sub-corpora did not reveal significant differences. A hundred and two of the 270 items of the list established were found in the corpus, as shown in Table 9. The sub-corpora contained about the same number of Netspeak types; males used 83 different items, and females 81. Females used slightly more tokens than men: 2,013 were found in the female corpus, and 1,866 in the male corpus.

Table 9. Results of the corpus analysis: overall frequency

	Total	Male	Female	Percentage difference
Types	102	83	81	2.4%
Tokens	3,969	1,866	2,103	11.9%

Only 9 Netspeak features appeared more than 100 times in the corpus; all were acronyms or initialisms, as seen in Table 10. Overall, the most frequent item was *LOL*, which appeared 729 times and was used by 62% of Redditors. The acronym *OP* had a lower frequency, but was used by a higher proportion of Redditors; it was used 649 times, by 71.5% of Redditors. Next comes *SO*, or "significant other", which has a frequency of 458 and was used by almost half of the Redditors. *IMO*, *IDK*, *WTF*, *BF*, *TBH* and *OMG* feature lower down on the list. All occurred significantly less than the three most frequent items; they were used between 227 and 100 times. Other items that have a rather high frequency but are not acronyms were found, notably the non-standard spelling *U* for "you", which was used by 14% of Redditors, and the "*because* + noun" construction, used by 15% of Reddit users in constructions such "I've seen photographs because Reddit, but that's about it", or "Doug has strange eyebrows and now I think it's because botox.".

While checking the results to ensure that no false positive was included, it was noticed that several acronyms of the Wikipedia list had several meanings; *TBF*, for instance, was not used in the Wikipedia meaning of "time between failures". In Netspeak, it seems to have several meanings, including "to be frank" and "to be fair", according to the not very reliable source Urban Dictionary (TBH, 2003.). In the Reddit corpus, *ETA* did not stand for "estimated time of arrival". Typically, it was added by a user at the end of a comment they had previously posted, and thus more likely meant "edited to add" (ETA, n.d.).

2.2. Netspeak Lexical Items and Gender

As it was observed with emoticons, the lists of the most frequent features in the male and the female corpora have a lot in common. Table 11 shows that 7 out of 10 items appear on both lists: *LOL, OP, SO, IMO, IDK, WTF,* and */s. BF* and *GF* could arguably be added to this list, if they are considered to be equivalents. *LOL* was slightly more used by females; it appears 397 times in the female corpus, and 332 times in the male corpus. About two-thirds of females Redditors used it, while 57%

of males did. Moreover, *LOL* features at the top of the female list, whereas males used it less than the acronym *OP*, which was used 338 times. Again, however, the frequencies of *LOL* and *OP* are quite similar in both corpora, with frequencies higher than 300. The most significant difference is the use of the acronym *SO*, or "significant other". Female Redditors used it more than four times more often than male Redditors; it has a frequency of 89 in the male corpus and of 369 in the female corpus. It was employed by 62% of female Reddit users, and by only 35% of males. *IMO*, on the other hand, appears twice as many times in the male corpus as in the female one. There is a less significant difference in the usage of *IDK*; it has a frequency of 95 in the female corpus, and of 69 in the male corpus. Usage of */s* is comparable in both corpora. The non-standard spelling *U* was produced 43 times in the male corpus and 22 times in the female corpus. *TBH* was favored by females rather than males, with a frequency of 58 in the female corpus, and of 52 in the male corpus.

Table 10. List of the 30 most frequent Netspeak items found in the whole corpus

Item	Frequency	Percentage of users
LOL	729	62%
OP	649	71.5%
SO	458	48%
IMO	227	58.5%
IDK	164	24.5%
WTF	121	35%
BF	115	21%
TBH	100	25.5%
OMG	100	29%
/s	98	23%
TIL	93	27%
GF	76	20.5%
LMAO	73	11%
BTW	66	19%
U	65	14%
IIRC	52	22%
IRL	47	16%
ASAP	38	13.5%
Because	35	15%
TL;DR	35	14%
FYI	32	14%
ETA	28	8.5%
FTFY	27	8.5%
STFU	24	6%
FFS	21	7.5%
FAQ	20	6%
FWIW	19	7.5%
ITT	18	7.5%
bae	17	7%
WOOT	16	6.5%

Table 11. The ten most frequent Netspeak items in each sub-corpus

Female corpus

Rank	Items	Frequency	Number of users	Proportion in total emotion use
1	LOL	397	67%	18.89%
2	SO	369	61%	17.55%
3	OP	311	67%	14.80%
4	IDK	95	29%	4.52%
5	BF	94	31%	4.47%
6	IMO	75	33%	3.57%
7	OMG	74	41%	3.52%
8	WTF	59	37%	2.81%
9	TBH	58	33%	2.76%
10	/s	49	26%	2.33%

Male corpus

Rank	Items	Frequency	Number of users	Proportion in total emotion use
1	OP	338	76%	18.15%
2	LOL	332	57%	17.82%
3	IMO	152	42%	8.16%
4	SO	89	35%	4.78%
5	IDK	69	20%	3.70%
6	TIL	66	35%	3.54%
7	WTF	62	33%	3.33%
8	GF	50	26%	2.68%
9	/s	49	20%	2.63%
10	**U**	43	17%	2.31%

2.3. Individual differences

Analysis of dispersion indicates that, in both corpora, a small portion of users are responsible for the production of a large number of Netspeak features. The 25 most prolific male Redditors used 1,079 tokens out of the 1,866 found in the corpus. Similarly, 25% of female Redditors produced 1,002 Netspeak features, a little less than half of the 2,103 items found. In the female corpus, the ten most prolific users produced 504 features, or almost a fourth of the total production of Netspeak items. The top 10 users of the male corpus had an even more significant impact on the total production of acronyms and other Netstpeak features; they employed 651 items, almost a third of the overall production. Table 12 summarizes the results per user. It shows that no Redditor produced less than 1 feature. Seventy-nine percent of females used from 1 to 30 features and 82% of males used 1 to 30 features. There are a few heavy users in each corpus, who produced more than 41 Netspeak lexical features. The main differences, in terms of distribution, are found in the least prolific users,

who used less than 10 of the items on the list, and who are more numerous in the male group. The male group also contains more heavy users: 8 males produced more than 50 items while only 2 females did.

Further analysis was also conducted in order to investigate the item *SO;* the goal was to better understand how it was used in each corpus and to see to what extent the results may be generalized. It revealed that the high frequency of the item in the female corpus was not due to the presence of a few outliers. Of the 61 female Redditors who used *SO*, 44 produced it less than 10 times, and 17 used it between 10 and 22 times.

Table 12. Frequency of emoticon use in the two sub-corpora

Number of tokens	Female users	Male users
1-10	26	42
11-20	27	30
21-30	26	12
31-40	13	6
41-50	6	2
51-60	0	2
61-70	0	2
71-80	1	2
81-90	0	2
91 and more	1	0

2. 4. Contrasting two male and two female corpora

The two sub-corpora were each divided into two smaller corpora, and the total number of emoticons produced in each sub-corpus was compared. The same analysis, applied to emoticons, revealed a greater difference between the two female corpora. This time, the greater gap was found between the male sub-corpora; the first male sub-corpus contained 724 Netspeak features, and the second 1,124, or a percentage difference of 44.8%. The difference in usage between the two female sub-corpora was equivalent with what was found when contrasting emoticon use in the two male sub-corpora.

Table 13. Total number of Netspeak used in the two male and the two female sub-corpora

	Sub-corpus 1	Sub-corpus 2	Percentage difference
Male	724	1,142	44.8%
Female	1,163	941	21.1%

IV. Discussion

1. Discussion of Results

1. 1. Emoticons: Discussion of Overall Results

Comparing the frequency of emoticons in the Reddit corpus in overall data with previous research about emoticons is not possible in most cases, since not all researchers included the word count of their corpora in their studies, but rather the number of comments, posts or messages. The 2-million word Reddit corpus contained 2,430 emoticons, and accounted for only 0.12% of the total token count. Emoticons were significantly four times less frequent in the Reddit corpus than in Baron's corpora of 11,718-word corpus of IM conversations; emoticons accounted for 0.41% of her corpus. Garrison et al. (2011), another study of IM use among college students, found 301 emoticons in a 32,000 corpus, or a frequency of 0.94% (p 119). In Lyddy et al. (2012), emoticons were forty times as frequent as in the Reddit corpus, and represented a little more than 5% of the total data. These discrepancies may be explained by the nature and the constraints of the various outputs, even if other factors, like age, cannot be entirely excluded, since these studies focused exclusively on college students while it is extremely likely that the Internet users in the Reddit corpus have a much wider variety of backgrounds. Text messages, studied by Lyddy et al (2014), also seem a lot more conducive to emotion use than IM and Reddit. The shorter length of messages, the small size of cell phones' screens and the way messages are typed, using one or two fingers, may explain this feature of text messages. Even if Lyddy et al.'s study did not make a distinction between emoticons and emojis and included them both in their frequency count, probably since many mobile interface automatically convert emoticons into emojis, it suggests that message length may be correlated with production of emoticons. On web platforms that restrict the number of characters, such as Twitter, which has been described as the "SMS of the Internet" (Crystal, 2011, p. 36), emoticons may be significantly more frequent. Crystal's study of Twitter seems to support this hypothesis, even if he used a small corpus of tweets (2011).

The range of emoticons (43) produced in the Reddit corpus is comparable to the results of some previous studies, such as Vidal, Ares, and Jaeger (2016, which found 50 emoticon types in their corpus of 12,260 tweets, or Thompson & Filik (2016),

who found 72 and 50 different emoticons in the corpus of text messages produced by students in their two experiments. It is significantly more than Skovholt et al. (2014), who found only four different emoticons in a corpus of work emails (p. 786), or than Garrison et al. (2011), who found 18 emoticon types in a corpus of IM. It is significantly less, however, than the 15,049 different kinds of emoticons captured by Park et al. (2013). Park et al.'s corpus, which contained tweets written by 56 million users, was much larger than the Reddit corpus. It may be responsible in part for the huge difference in types of emoticons found. The discrepancy is also very likely explained by the method chosen for this dissertation, since using a pre-established list of emoticons inevitably limited the number of emoticons found, and there might be a number of variations not accounted for in the results.

In the Reddit corpus, a small subset of emoticons makes up most emoticon usage, with 10 emoticons accounting for 91.8% of overall emoticon use. This type of imbalance has been reported by other studies (Tossell et al., 2012; Skovholt et al., 2014; Chang, 2016). The ratio between the most frequent and the less frequent emoticons in the Reddit corpus is on par with Park et al.'s (2013) findings; the researchers noted that 76% of emoticons appeared in their corpus fewer than ten times (p. 469); in the Reddit corpus, 72% of emoticons had a frequency of 10 or lower.

Many studies of emoticons have found the "smiley face" to be the most widely used emoticon (Rezabeck & Cochenour, 1994; Wolf, 2000; Baron, 2004; Huffacker & Calvert, 2005; Garrison et al. 2011; Park et al., 2013; Skovholt et al., 2014; Vidal et al., 2016). In the Reddit corpus, the 1,006 instances of smiley emoticons represent almost half of overall emoticon usage, which is consistent with previous findings.

The "midget" version of the smiley or :), which consists of two characters, was used about 500 times more often than the "classic" smiley with a nose or :-). It may seem surprising, since the emoticon with a nose was invented first and is considered to be the basic emoticon form, the smiley without being merely a variant. In several previous studies, classic emoticons outnumbered emoticons without noses. In Rezabeck & Cochenour's study of listservs (1994), for instance, the smiley with a nose was produced 23 times out of 58 emoticons. The Reddit corpus's tendency to favor midget emoticons, with 34 out of 43 emoticon types being midget emoticons, is however consistent with other research such as Baron's study of Usenet newsgroups (2004), which found the midget smiley :) to account for 93% of overall emoticon use

(p. 830). A more recent study of emoticon use has investigated the use of emoticons as markers of sarcasm in text messaging (Thompson & Filik, 2016). The authors conducted two experiments. In the first one, participants had to edit an existing conversation in order to make its intent clear. The authors reported that out of 1,184 instances of "nose-optional emoticons", the variant without a nose was used in 76% of cases (p. 111). In the second experiment, participants were given more freedom and could produce their own utterances. The variant without a nose was used in 95% of cases (p. 115). The authors proposed several possible explanations. The first was the "monetary economy of producing fewer characters", a hypothesis they quickly rejected based on the fact that message length no longer has an impact of cost in most messaging services. They also suggest that the nose does not add any expressive content to emoticons, or that producing midget smileys involves pressing fewer keys, and thus less effort on the part of users (p. 116). It could be argued that midget forms are abbreviations of emoticons, and that they have become the norm on Reddit, and probably on other platforms as well. The decline of the classic emoticon may also be linked to the rapid growth of emojis. Pavalanathan & Eisenstein (2015) have tried to measure the causal effect of the introduction of emojis on Twitter, and have found the increasing use of emojis on the social network to be linked to a high rate of decrease in the production of emoticons, especially of "happy" and "playful" emoticons. (p. 3)

The most frequently used emoticons in the Reddit corpus, and their rank in terms of frequency, are similar to Park et al.'s (2013) findings on many points (p. 469). The top five emoticons are the same in both cases, even though Park et al. found more instances of the "happy" emoticon :D than of the "sad" emoticon :(, while the opposite was true in the Reddit corpus. Other studies also found the sad emoticon to be the most frequent emoticon other than the smiley emoticon (Baron, 2004; Huffackert & Calvert, 2005).

1. 2. Emoticons and Gender

The Reddit corpus seems to indicate that there is a correlation between gender and emoticon usage, since females used more than twice as many emoticons as males. It affirms the findings of several studies, which have reported that females tend to use more emoticons than men (Witmer & Katzman, 1997; Wolf, 2000; Lee, 2003; Baron, 2004; Tossell et al. 2012, Fox et al. 2007). Not only was the overall

frequency of use of emoticons much higher in the female corpus, but only 3 out of 100 female Redditors did not use any emoticons, while 24% of male Redditors produced no emoticons. More males used emoticons, however, than in Baron's corpus of instant messages, where only one male in six produced emoticons (2004). The results of the analysis of the Reddit corpus do not support Huffacker and Calvert's (2005) findings that males tended to use more of what they call flirty emoticons or :p, referred to as "playful" emoticons in the Wikipedia list, as well as more sad emoticons. On the contrary, in the Reddit corpus, females tended to favor this emoticon, using it four times more than men.

Given the striking results of the analysis of the Reddit corpus and the existing body of research suggesting that gender differences do exist in the use of emoticons, it would be tempting to conclude that production of emoticons seems linked to gender. Further analyses, conducted in order to have a clearer view of the distribution of emoticons in the corpus, gave mixed results. When the female corpus was randomly divided into two smaller corpora, the results also showed a significant difference of usage between the two corpora (1,100 vs 641, or a percentage difference of 52.73%). The same analysis, performed on the male corpus, revealed a smaller discrepancy, with a percentage difference of 20.93%. Baker (2014, p. 41) warned that the process of comparing two corpora is bound to reveal differences and gaps. In the case of the Reddit corpus, it seems that caution must indeed be exercised when interpreting the findings. Both the male and the female corpora exhibit large differences in individual usage of emoticons, which largely account for the differences found when the large corpora were compared against each other, and when the two sets of smaller corpora were analyzed. The five most frequent emoticon users in the male and in the female corpora may be identified as outliers, since they were responsible for the production of about a fourth of the overall number of emoticons. Two explanations may thus be given for the differences of frequency of use noted between the male and the female corpus. There could be a correlation between gender and emoticon use, but it is also possible that, because the sample size was relatively small, with 100 male and 100 female users, there simply happened to be more heavy users of emoticons in the female corpus. Analyzing a bigger corpus, which would include the contributions of a large number of Reddit users, would probably provide more generalizable results.

The analysis performed on the Reddit corpus stresses the differences between male and female usage, but it also reveals similarities. Forty-seven percent of the female Redditors and 54% of male Redditors behaved in the same way, producing about the same number of emoticons, which indicates that there are probably more similarities than differences between the two sub-corpora. The "similarity hypothesis" is reinforced, on the other hand, by the fact that females and males produced about the same number of emoticon types, respectively 35 and 36. It does not support the hypothesis that women use emoticons more creatively than men because emoticons "have aesthetic quality and emotional expressiveness" (Witmer & Katzman, 2006, para 8) and does not affirm Tossell et al.'s (2012) findings that males used a wider vocabulary of emoticons than females (2011, p. 662), or Wolf's finding that women produced a greater diversity of emoticons (2000, p. 830).

A perhaps surprising finding is the use of the Le Lenny Face emoticon, or **(͡° ͜ʖ ͡°)**, which was found in the male corpus; it should be noted, however, that even if it appears in the list of the ten most frequent emoticons in the corpus, it was only produced 11 times. The Le Lenny Face differs from other emoticons found in the corpus in that it is a vertical emoticon, and not a horizontal one. Moreover, it is a quite complex emoticon; it is created by using six different characters from different Unicode sets, including degree signs, parentheses, a breve and two upside-down breves, and an inverted glottal stop (How to make a Lenny face, n.d.). No study of this type of emoticons seems to have been conducted, and the only sources of information there are may not be reliable. A source claims that it is often used to spam discussion forums and that it is related to the practice of "shitposting", the intentional spamming of message board (Shitposting, n.d). Another states that Le Lenny Face is typical of the 4chan discussion board (Le lenny face, n. d.), known or its profane and sexist content and mostly used by men (Knutila, 2011; Marwick, 2013, p. 66). It is possible that the emoticon Le Lenny Face and other "shitposting" practices used on 4chan were imported to Reddit primarily by male 4chan users, have continued to be used by these same users, and have been adopted by a number of other Redditors.

Creating and contrasting two smaller female and male corpora produced results that did not firmly affirm or infirm the existence of gender differences in emoticon use. It showed that there were differences in each set of sub-corpora. Even if these differences were less great than then differences found when comparing the male to

the female corpus, they were still significant. These results, together with the fact that about half of Redditors used about the same number of emoticons, put into question the generalizability of the results found when contrasting the two corpora.

2. Acronyms and Netspeak Lexical Features

2.1. Netspeak Lexical Features: Discussion of Overall Results

Out of the 270 items on the list, 102 were found in the two-million word corpus. Many features were not used at all. On the other hand, other items not present on the list, and thus not investigated, were probably used, and it is difficult, if not impossible, to draw conclusions about the overall frequency of Netspeak lexical items based on the corpus-based analysis conducted.

Searches performed to look for Netspeak lexical items confirmed previous studies that found acronyms and abbreviations to be the most frequent Netspeak features, together with emoticons. Out of the 30 more frequent features, 25 are acronyms and initialisms. The presence, at the top of this list, of the acronyms *LOL* and *OMG* does not come as a surprise, as they have been shown to be among the most commonly used Netspeak features (Baron, 2004; Tagliamonte & Denis, 2008). Maybe more surprising is the high frequency of *OP*, which was used only slightly less than *LOL*. This acronym stands for "opening poster" and is used to refer to the person who posted the link or the text other Redditors react to. It seems to indicate, together with the other frequently use acronym *TIL* or "Today I learned", which derives from the name of a popular subreddit, the existence of a "Reddit lingo" described by non-academic sources (Koerber, 2014). This "lingo" may share characteristics with other forums and message boards. It may also be characterized by other lexical elements, such as the construction "because + noun". Non-standard spellings were not found to be prevalent in the corpus; even though there were only 20 such spellings on the list used as a starting-point, it seems that acronyms are favored by Reddit users over non-standard spellings. Only *U* was found to have a frequency higher than 20. The finding that several acronyms, such as *TBF*, *ETA* and *GF,* had different meanings in the Reddit corpus than those specified on the Wikipedia list suggests that a number of Netspeak acronyms have homonyms.

The relatively high frequency of the feature /s, which was produced 98 times and was used by 23% of Redditors, shows the limitations of the corpus-based method, since it was included to the items to be searched for uniquely based on a

chance spotting. It does not seem to have been described by the literature, and it is unclear where it originated. According to Reddit itself (What does /s mean?, 2014), it mimics an HTML end tag, and indicates that what was written previously was intended as sarcasm although its usage does not seem to be restricted to Reddit ([/s] Urban Dictionary, 2007). This finding suggests that items derived from the field of coding may be a staple of Netspeak on Reddit, and possibly on other sites and forums as well.

2.2. Netspeak Lexical Items and Gender

When compared, the overall frequencies of the Netspeak items found in the corpus only show slight gender differences. As was the case with emoticons, the number of types of features produced is remarkably similar, with 83 items used by males and 81 by females. Frequencies of some of the most frequent features, such as *LOL*, *OP*, *IDK*, *WTF*, and */S*, do not show significant differences. On the contrary, they reveal great similarities of use. A few items, however, were used at least twice as many times by a group over another; males tended to favor *IMO* and TIL, and females *OMG* and *SO*. It could thus be argued that the similarities outweigh the differences, especially since a few heavy users in each corpus were responsible for the production of a large number of the items found. If these heavy users are excluded from the corpus, usage of Netspeak lexical items seems equivalent in each corpus. Dividing each corpus into two sub-corpora and contrasting usage of Netspeak features among each group also showed that the greater differences were found within each group, and not between the two gendered corpora. Differences in the dispersion of items among different users thus seem more significant than gender differences.

The most striking finding is probably the large discrepancy noted in the use of the acronym *SO*, "significant other". Several hypotheses may be advanced to explain this difference. It is possible that male Redditors, when they refer to their significant other, favor other terms over *SO*, such as *BF*, *GF*, *girlfriend*, *boyfriend*, *wife*, *husband* or *partner*. It seems unlikely though, since the analysis showed that the frequency of *GF*, "girlfriend", is more than two times lower in the male corpus than the frequency *BF*, "boyfriend," in the female corpus, while females used *GF* slightly more than males used *BF*. It could also be that, on Reddit, males talk less about their significant others, and maybe about their personal lives in general, than female

Redditors. Formulating this hypothesis would mean agreeing with the gender difference paradigm, which probably explains the heavier use of *SO* in the female corpus by the focus on interpersonal relationships that has been found to be one of the properties of female style. However, there may be a third explanation. Female Redditors whose contributions have been used in this particular corpus may tend to gravitate towards subreddits focusing on advice giving and personal life, while men may comment on sub-forums dealing with other types of topics, such as sports, news or politics. In the absence of studies examining the gender make-up of the different subreddits, it is impossible to know if that hypothesis is correct.

The gender differences found in the use of emoticons also raise several questions. As Wolf pointed out, they certainly appear to reinforce the "stereotype of the emotional female and the inexpressive male (Wolf, 2000, p. 827). However, the analysis of individual usage in the Reddit corpus shows that gender may be a factor in the production of emoticons, but that it is probably not the only one. Since gender was the only variable that was controlled for in the study, it impossible to know what these factors are. The best way to find answers is to take a look at previous research. Baron (2004) has stated that variations in online usage reflect gender, but also age, education level, personality, cultural background and years of experience with CMC platforms (2004, p. 29). Age was shown to impact the use of online writing style, with the use of emoticons, acronyms, slang and capitalization decreasing as age increased (Rosenthal & McKeown, 2011, p. 763). Tagliamonte and Denis came to a similar conclusion when examining the use of *LOL*, which, in their study, declined systematically according to age (2008, p. 13). While gender was not shown to have an effect on standard language in a study of Flemish instant messages, Peersman et al. (2016) found that the use of Netspeak features peaked from age 13 to 15. Interacting with friends is also conducive to emoticon production (Derks et al., 2008, p. 99). Thus, it cannot be excluded that the female Redditors selected for this dissertation were younger than males. They may also gravitate towards smaller subreddits, where users have friendly online relationships.

Cultural differences have also been found to impact the use of Netspeak items. Two studies about emoticons focused on cultural differences and showed that Japanese and South Korean Internet users produced more emoticons than US Internet users (Park et al. 2013; Kavanagh, 2016). Apart from personal preferences, there may also be an effect of the idiosyncracies that are bound to develop when a group of

people interact on the sample platform (Crystal, 2004, viii). Thus, they may very well be stylistic differences between the various subreddits; this is suggested by Driscoll's study of multi-user dungeon chatrooms, which showed what she called Internet "dialects" could be confined to a particular chatroom.

It was shown that the gender make-up of the setting where online interactions take place can have an impact of the number of emoticons produced. In a study of dyadic IM exchanges, Lee found that males rarely use emoticons with other males, but do use them when interacting with females, while the gender of the recipient had no influence with female users. Wolf also noted a pattern of change in her study, reporting that males produced more emoticons while interacting with women, and vice versa. In mixed-gender newsgroups, she found that the differences between male and female frequency of use were not statistically significant (2000, p. 831). Fox et al.'s study of email showed that "sex of the target, rather than that of the speaker, affected differences in talk time" (2007, p. 395). Email is a one-to-one communication channel while Reddit is a one-to-many, but gender make-up of a subreddit may also influence the style of the comments posted. Again, in the absence of serious research about gender on the various subreddits, it is impossible to evaluate the possible impact of this factor. As observed but not measured during the compiling of the corpus, it is likely that more female Redditors comment on the /r/AskWomen subreddits, and that men constitute the majority of participants in the /r/AskMen. However, these subreddits were merely taken as starting-point to find users who specified their gender, and the corpus contains comments from dozens or hundreds of different subreddits.

3. Limitations and Suggestions for Further Research

There are several limitations to the present study and to its findings. First, there is no way of knowing that the Reddit users chosen are being truthful and have indicated the gender they really identify with. Since information about users was also limited, gender was the only variable used when building of the corpus, which did not allow to study other factors, and the ways in which they might interact with gender.

The corpus-based methodology that was used has serious limits when investigating open-ended sets of items. The lists established were intended to be possibly representative of the most frequent items in the corpus, but were by no

means exhaustive, since a vast number of emoticons, abbreviations, acronyms and nonstandard spellings exist. The list of lexical items has other shortcomings. First, the line between lexical features typical of Netspeak and features and spoken or written features is sometimes blurry; the Wikipedia list contains items that could arguably not considered as Netspeak, such as the military term *AWOL*, "absent without official leave". Second, the fact the several items, *GF*, *TBF* and *ETA* were found to have different meanings in the corpus suggests that the list is not necessarily representative of the type of Netspeak used on Reddit.

Therefore, this dissertation does not claim that have uncovered all the instances of emoticons, acronyms, abbreviations and other Netspeak features in the Reddit corpus. It is extremely likely that many lexical items were left behind by the study. The example of the sarcasm marker */s* is telling, because it showed how a widely used lexical feature was not even considered when the list of items was compiled, but was uncovered serendipitously, while browsing users' comments. Expressive lengthening, especially, was overlooked in the study, as well as other non-standard spellings. A corpus-driven study of Netspeak would probably yield different results, certainly in the number of types of features found. The question is, is it possible that the frequency of these items is so great that it would give dramatically different results, especially when contrasting the two corpora?

It should also be noted that graphical items typical of Netspeak were not examined in the study of the Reddit corpus. Because AntConc, the software that was used, can only process .text files, the formatting of the comments was lost during to conversion from Word files to plain text files. The corpus was not annotated, and some information that could have been relevant in this study just disappeared. Even though Reddit does not allow users to pick a font and only uses Verdana, users can personalize their messages by using markdown syntax (Reddit comment formatting, 2011) or by downloading the Reddit Enhancement Suite or RES, a browser extension that was developed independently from Reddit. Even if only basic text formatting is possible, Reddit users are able to display a great deal of creativity, as shown in Table 14. The practice of "vertical posting", which is when a message is written horizontally as well as vertically and which is believed to have originated on the message board 4chan (Vertical posting, 2015), was noticed in the corpus during the compiling process but was not studied due to the difficulty to search for it in AntConc.

Table14: Examples of graphical elements used on Reddit

Formatting	I NEED ALL POSSIBLE WEIGHTS to convey my *FULL RAGE*
Vertical posting	C U R R E N T Y E A R U R R E N T Y E A R
Emojis	Edit: She's fishing and he bit. "Average" lol

Another phenomenon that was not investigated is the use of emojis. Emojis are ready-made symbols that are represented by Unicode characters, while emoticons are created from a sequence of ASCII characters. Despite the similarity between the two terms, they have different origins. Emoticon is a blend of "emotion" and "icon", where as emoji is a loanword from Japanese (Oxford Dictionaries Word of the Year 2015). It is composed of "e", which means "picture", et "moji", which stands for "letter, character". Use of emojis has rapidly increased over the past few years, and has been linked to a decrease in use of emoticons in a corpus of Tweets (Pavalanathan & Eisenstein, 2015). Even if they do not seem used to a great extent on Reddit, it is possible that emojis have an impact on emoticons. Other phenomena could be investigated in further studies, such as the use of hashtags and asterisks, or the expressive lengthening of words. Lexical features derived from the field of coding, like /s or /rant, to signal the end of an unusually long comment or message, would be another possible avenue of research.

Research about Reddit is still scarce and more information about its users is needed in order to better understand the language of the sites and the gender

interactions that take place on the platform. This information could for instance be obtained by conducting surveys, or by close analysis of a few subreddits. Focusing on one or on a few subreddits may reveal the existence of speech communities. Contrasting the language of Reddit with other forums, like 4chan or Metafilter, or social media sites such as Twitter, Facebook, and Instagram, would allow more insight about language variation on the Internet.

Conclusion

This dissertation did not give a definitive answer to the problem it sought to investigate. The corpus-based study of emoticons and other Netspeak lexical items has revealed some significant differences, which may be correlated to gender. Emoticon usage, in particular, was found to be asymmetrical. Females produced more than twice as many emoticons as males, which is consistent with the findings of most studies of emoticon use in computer-mediated communication, and seems to support the traditional view of female communication style as expressive and supportive. It is unclear what might explain this discrepancy; gender may be one of the factors that come into play, with age, dialect and socio-cultural backgrounds. On the other hand, the number of emoticon types found in the two sub-corpora was almost identical. The list of the most frequent emoticons in each corpus was also extremely similar.

No significant gap was found in usage of most acronyms and other Netspeak lexical features. Males and females used about the same number of tokens. The number of Netspeak lexical types found in both corpora and the list of the most frequent items were also very similar. The only significant result was the vast difference in the frequency of the acronym *SO* or "significant other", which females used four times more often than males. Since gender was the only variable taken into account, this finding cannot be explained by this dissertation. Further research about gender and Reddit may uncover some of the complex dynamics of gender, language, and computer-mediated communication.

References

Al-Rfou, R., Pickett, M., Snaider, J., Sung Y., Strope, B. & Kurzweil, R. (2016). Conversational contextual cues: The case of personalization and history for response ranking. Retrieved from https://arxiv.org/abs/1606.00372v1

Anderson, M. (2015). *Men catch up with women on overall social media use*. Retrieved from http://www.pewresearch.org/fact-tank/2015/08/28/men-catch-up-with-women-on-overall-social-media-use/

Andrews, P. (1994) User friendly: Put on a happy face, but not in my email!. Retrieved from http://www.mit.edu/people/cordelia/smileys_edit.html

Baheri, Tia. (2013, November 20). our Ability to Can Even: A Defense of Internet Linguistics. Retrieved from http://the-toast.net/2013/11/20/yes-you-can-even/

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

Barbaresi, A. (2015). Collection, description, and visualization of the German Reddit corpus. Proceedings of the *2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media*.

Baron, N. S. (1998). Letters by phone or speech by other means: The linguistics of email. *Language and Communication, 18*, 133-170.

Baron, N. S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology, 23*(4), 394-423.

Barthel, M., Stocking, G., Holcomb, J. & Mitchell, A. (2016). *Nearly eight-in-ten Reddit users get news on the site.* Retrieved from http://www.journalism.org/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/

Bergstrom, K. (2011). Don't feed the troll: Shutting down debate about community expectations on Reddit.com. *First Monday, 16*(8). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/3498/3029#author

Bucholtz, M (2004). Editor's introduction. In Lakoff, R. T. (2004), M. Bucholtz (Ed.), *Language and woman's place*. *Text and commentaries* (3-14). Oxford: Oxford University Press.

Caffrey, C. (2016). Reddit (website). *Salem Press Encyclopedia.* Research Starters.

Centivany, A. & Glushko, B. (2015). “Popcorn tastes good”: Participatory policymaking and Reddit’s’ “Amageddon”. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’16)*, San Jose, CA. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2753144

Chang, C. Y. EFL reviewers’ emoticon use in asynchronous computer-mediated peer response. *Computers and Composition, 40*, 1-18.

Cheng, N., Chandramouli, R., & Subbalakshmi, K.P. (2010). Author identification from text. *Digital Investigation, 8*, 78-88.

Coates, J. (1996). *Women talk*. Oxford: Blackwell.

Colley, A. & Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology, 21(*4), 380-392.

Complete Bboard Post Sequence in which the :-) Was Invented. (n. d.). Retrieved from http://research.microsoft.com/en-us/um/people/mbj/Smiley/BBoard_Contents.html

Crystal, D. (2001). *Language and the Internet* (2nd ed). Cambridge: Cambridge University Press.

Crystal, D. (2004). *A glossary of Netspeak and Textspeak.* Edinburgh: Edinburgh University Press.

Crystal, D. (2011). *Internet Linguistics*. London: Routledge.

Cummings, J. N., & Kraut, R. (2001). Domesticating computers and the Internet. Retrieved from http://repository.cmu.edu/hcii/94/

Cunha, T.O., Weber, I., Haddadi, H., & Papa, G. L. (2016). The effect of social feedback in a Reddit weight loss community. *ACM Digital Health 2016*. Retrieved from http://arxiv.org/abs/1602.07936

Cushing, P. J. (1996). Gendered conversational rituals on the Internet: An effective voice is based on more than simply what one is saying. *Anthropologica*, *38*(1), 47-80. http://www.jstor.org/stable/25605819

Danish, Y. D. & Talukdar, P. (2015). Want answers? A Reddit inspired study on how to pose questions. Retrieved from http://arxiv.org/abs/1512.01768?

Davis, B. H., & Brewer, J. P. (1997). *Electronic discourse: Linguistic individuals in virtual space*. New York; State University of New York Press.

Davis, B. H., & Brewer, J. (1997). *Electronic discourse*. Albany, NY: State University of New York Press.

Derks, D., Bos, A. E. R., & Von Grumbkow, J. (2008). Emoticons in computer-mediated communication: Social motives and social context. *CyberPsychology & Behavior, 11*(1), 99-101. doi: 10.1089/cpb.2007.9926

Dery, Mark. (1993). Flame wars. *Southern Atlantic Quarterly 92,* 559–68.

Dewey, C. (2014, September 5). Meet the unashamed 33-year-old who brought the stolen celebrity nudes to the masses. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-intersect/wp/2014/09/05/meet-the-unashamed-33-year-old-who-brought-the-stolen-celebrity-nudes-to-the-masses/

Dewey, C. (2015, June 10). These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy — and why it banned them. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-intersect/wp/2015/06/10/these-are-the-5-subreddits-reddit-banned-under-its-game-changing-anti-harassment-policy-and-why-it-banned-them/

Dobrow, S. B. (2015) Rise of the Internet and the World Wide Web. *Salem Press Encyclopedia.*

Driscoll, D. (2002). The ubercool morphology of Internet Gamers: A linguistic analysis. Retrieved from http://www.kon.org/urc/driscoll.html

Duggan, M., & Smith, A. (2013). 6% of online adults are reddit users. Retrieved from http://www.pewinternet.org/2013/07/03/6-of-online-adults-are-reddit-users/

Durndell, A. & Haag, Z. (2002). Computer self efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample. *Computers in Human Behavior, 18,* 521-535.

Eckert, P. & McConnell-Ginet, S. (1992). Communities of practice: Where language, gender and power all live. *Locating Power: Proceedings of the Second Berkeley Women and Language conference*. Women Language Group.

Emoticon. (n. d.). Retrieved from http://www.merriam-webster.com/dictionary/emoticon

ETA. (2006). Retrieved from http://fr.urbandictionary.com/define.php?term=ETA

FAQ Writing Prompts. (n. d.). Retrieved from https://www.reddit.com/r/WritingPrompts/wiki/faq

Ferrara, K., Brunner, H., & Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Communication, 8(1)*, 8-34.

Fishman, P. (1978). What do couples talk about when they're alone. In Douglas Butturff and Edmund L. Epstein (Eds.) *Women's Language and Style,* (pp. 11-22). Akron, Ohio: L & S.

Fox, A. B., Bukatko, D., Hallahan, M., & Crawford, M. (2007). The medium makes a difference gender similarities and differences in instant messaging. *Journal of Language and Social Psychology, 26*(4), 389-397.

Garber, Megan. (2013, November 19). English Has a New Preposition, Because Internet. *The Atlantic.* Retrieved from http://www.theatlantic.com/technology/archive/2013/11/english-has-a-new-preposition-because-internet/281601/

Garrison, A., Remley, D., Thomas, P., & Wierszewski, E. (2011). Conventional faces: Emoticons in instant messaging discourse. *Computers and Composition*, *28*, 112-125. doi:10.1016/j.compcom.2011.04.001

Internet Society global Internet report (2015). Retrieved from http://www.internetsociety.org/globalinternetreport/

Isaac, M. (2015, July 16). Reddit changes content rules as Steve Huffman takes charge. The New York Times. Retrieved from http://www.nytimes.com/2015/07/17/technology/reddit-steve-huffman.html

Gibeault, M. (2016). Embracing geek culture in undergraduate library instruction: The TIL subreddit for resource evaluation and qualitative assessment. . *University Libraries Faculty Publications and Presentations. Paper 8*.

Gibbs, M. (2006, November 15). Alexa and Alexaholic, useful, inaccurate, free stats. Retrieved from http://www.networkworld.com/article/2300804/software/alexa-and-alexaholic--useful--inaccurate--free-stats.html

Goodwin, M. H. (1990). *He-said-she-said: Talk as social organization among black children*. Bloomington: Indiana University Press.

Gray, J. (1992). *Men are from Mars, women are from Venus*. New York, NY: HarperCollins.

Hafner, K. (2002, September 19). Happy Birthday :-) to You: A Smiley Face Turns 20. *The New York Times*. Retrieved from http://www.nytimes.com/2002/09/19/technology/circuits/19SMIL.html

Hankes, K. (2015). The most violently racist internet content isn't found on sites like Stormfront and VNN any more. Retrieved on https://www.splcenter.org/fighting-hate/intelligence-report/2015/black-hole

Haralabopoulosa, G., Anagnostopoulosa, I., & Zeadallyb, S. (2015). Lifespan and propagation of information in on-line social networks: A case study based on Reddit. *Journal of Network and Computer Applications, 56*, 88-100.

Hass, M. R. (1944). Men's and women's speech in Koasati. *Language*, *20(3)*, 142-149.

Herring, S. C. (1996). Bringing familiar baggage to the new frontier: Gender differences in computer-mediated communication. In V. Vitanza (Ed.), *CyberReader* (144-154). Boston: Allyn & Bacon.

Herring, S. C. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet,* http://www.languageatinternet.org/articles/2007/761. Accessed June 4, 2015.

Herring, S. C., Johnson, D. A., & DiBenedetto, T. (2011). Participation in electronic discourse in a "feminist" field. In J. Coates & P. Pichler (Eds.), *Language and gender: A reader*, (2nd ed.) (pp. 171-182). Oxford: Wiley-Blackwell. http://www.wiley.com/WileyCDA/WileyTitle/productCd-1405191279,descCd-tableOfContents.html

Herring, S. C, & Stoerger, S. (2013) Gender and (a)nonymity in computer-mediated communication. In Holmes, J., M. Meyerholl & Ehrlich (Eds.), *Handbook of language, gender and sexuality* (2nd ed.). doi: 10.1002/9781118584248.ch29

Herring, S. C., & Zelenkauskaite, A. (2008). Gendered typography: Abbreviation and insertion in Italian iTV SMS. In J. F. Siegel, T. C. Nagle, A. Lorente-Lapole & J. Auger (Eds), *IUWPL7: Gender in Language: Classic Questions, New Contexts* (pp. 73-92). Bloomington, IN: IUL Publications.

History. (n.d.). *Reddit Metrics*. Retrieved May 7, 2016, from http://redditmetrics.com/history

Howard, M. C. & Magee, S. M (2013). To boldly go where no group has gone before: An analysis of online group identity and validation of a measure. *Computers in Human Behavior, 29*(5), 2058-2071.

How to make a Lenny face. (n.d.). Retrieved from https://www.reddit.com/r/howto/comments/3bxba5/how_to_make_a_lenny_face/

Huang, H. Y. & Bashir, M. (2016) Online community and suicide prevention: Investigating the linguistic cues and reply. Retrieved from http://chi2016mentalhealth.media.mit.edu/wp-content/uploads/sites/46/2016/04/sui_final_2016CHI.pdf

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10(2*), p 00. Doi **:** 10.1111/j.1083-6101.2005.tb00238.

Hyde, J. (2005). The gender similarity hypothesis. *American Psychologist*, *60(6)*, 581-592.

Isaac, M. (2015, July 16). Reddit changes content rules as Steve Huffman takes charge. *The New York Times*. Retrieved from http://www.nytimes.com/2015/07/17/technology/reddit-steve-huffman.html

Jespersen, O. (1922). *Language: Its nature, development and origin*. London: G. Allen & Unwin, ltd.

Kaye, L. K., Wall, H. J., & Malone, S. A. (2016). "Turn that frown upside-down": A contextual account of emoticon usage on different virtual platforms. *Computers in Human Behavior, 6,* 463-467. http://dx.doi.org/10.1016/j.chb.2016.02.088

Kiene, C. Monroy-Hernandez, A. & Hill, B. Surviving an "Eternal September" - How an online community managed a surge of newcomers. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems.* ACM, New York, NY. Retrieved from http://arxiv.org/abs/1605.08841

Kimbrough M., Guadagno R. E., Muscanell N. L. & Dill J. (1993). Gender differences in mediated communication: Women connect more than do men. *Computers in Human Behavior*, 29, 896-900.

Knutila, L. (2011). User unknown: 4chan, anonymity and contingency. *First Monday, 16(10)*. http://dx.doi.org/10.5210/fm.v16i10.3665

Koerber, Brian. (2014, March 10). The TL;DR Guide to Reddit Lingo. *Mashable.com*. Retrieved from http://mashable.com/2014/03/10/reddit-lingo-guide/#2YeRw_vB4Zqn

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17(4)*, 401-412.

Internet English Slang. (n. d.). Retrieved from https://en.wiktionary.org/wiki/Appendix:English_internet_slang

Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Lakoff, R. (1975) *Language and Woman's Place*. New York: Harper & Row. Reprinted 2004, ed. by Mary Bucholtz.

Lam, S. K., Anuradha Uduwage A., Dong Z., Sen S., Musicant, D. R., Terveen, L., & Riedl, J. (2011). WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance. *WikiSym'11*, October 3–5 2011, Mountain View, California.

Lee, C. (2003). *How Does Instant Messaging Affect Interaction Between the Genders*? Stanford , CA : The Mercury Project for Instant Messaging Studies at Stanford University. Retrieved from http://web.stanford.edu/class/pwr3-25/group2/pdfs/IM_Genders.pdf

Le lenny face. (2016) Retrieved from http://fr.urbandictionary.com/define.php?term=le+lenny+face

List of emoticons. (n. d). Retrieved from https://en.wikipedia.org/wiki/List_of_emoticons

List of texte emoticons. (n. d.). Retrieved from http://cool-smileys.com/text-emoticons

Lyddy, F., Farina, F., Hanney, J., Farrell, L., & Kelly O'Neill, N. (2014). An Analysis of language in university students' text messages. *Journal of Computer-Mediated Communication*, *19*(3), 546-561. http://dx.doi.org/10.1111/jcc4.12045

Maltz, D. N. & Borker, R. A. ([1982] 1983). A Cultural approach to male-female miscommunication. In J. J. Gumperz (Ed.), *Language and social identity* (196-215). Cambridge: Cambridge University Press.

Marwick, A. Gender, sexuality, and social media (2013). In J. Hunsinger & T. M. Senft (Eds.) *The Social Media Handbook* (pp. 59-74). London: Routledge.

McEwan, B. (2016). Communication of communities: linguistic signals of online groups. Information, *Communication & Society, 19*(9), 1233-1249. doi:10.1080/1369118X.2016.1186717

Men are from Mars, women are from Venus. (n. d). Retrieved from https://en.wikipedia.org/wiki/Men_Are_from_Mars,_Women_Are_from_Venus

Mikkelsen, Jensen, Tarding, Haasum & Rasmussen (2016). Trolls on social media. Retrieved from http://rudar.ruc.dk/handle/1800/27299

Mills, R. (2011). Researching Social News – Is reddit.com a mouthpiece for the 'Hive Mind', or a Collective Intelligence approach to Information Overload? *ETHICOM 2011 Proceedings*. Sheffield Hallam University, Sheffield.

Moderation. (2016). Retrieved from https://www.reddit.com/wiki/moderation

NetLingo List of Chat Acronyms & Text Shorthand. (n. d.). Retrieved from http://www.netlingo.com/acronyms.php

Nevalainen (2000). Gender differences in the evolution of standard English: Evidence from the Corpus of Early English Correspondence. *Journal of English Linguistics, 28(1*), 38-59.

Newman, M. L, Groom, C. J., Handleman, L. D & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*(3), 211-236.

Nie, N. H., & Erbring, L. (2002). Internet and society: A preliminary report. *IT&Society, 1(1),* 275-283. Retrieved from http://www.nomads.usp.br/documentos/textos/cultura_digital/tics_arq_urb/internet_society%20report.pdf

Noguti, V. (2016). Post language and user engagement in online content communities. *European Journal of Marketing*, *5*(5/6).

Odell, P. M., Korgen, K. O., Schumacher, P., & Delucchi, M. (200) Internet use among female and male college students. *CyberPsychology & Behavior*, *3*(5), 855-862.

Olsen, M. (2005) *Ecriture féminine*: Searching for an indefinable practice? *Literature and Linguistic Computing, 20*, 147-164. doi:10.1093/llc/fqi020

Olson, R. S., & Neal, Z. P. (2013). Navigating the massive world of reddit: Using backbone networks to map user interests in social media. Retrieved from https://arxiv.org/abs/1312.3387

Ovadia, S. (2015). More than just cat pictures: Reddit as a curated news source. *Behavioral and Social Sciences Librarian, 34*, 37-40. doi: 10.1080/01639269.2015.996491

Oxford Dictionaries Word of the Year 2015 is… (n. d.) http://blog.oxforddictionaries.com/2015/11/word-of-the-year-2015-emoji/

Park, J., Barash, V., Fink, C., & Cha, M. (2013). Emoticon Style: Interpreting Differences in Emoticons Across Cultures. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Patkar, M. (2015). 30 Trendy Internet Slang Words and Acronyms You Need To Know To Fit In. Retrieved from http://www.makeuseof.com/tag/30-trendy-internet-acronyms-slang-need-know-fit/

Pavalanathan, U. & Eisenstein, J. (2016). Emoticons vs. emojis on Twitter: A causal inference approach. *AAAI Spring Symposium 2016*. Retrieved from http://arxiv.org/abs/1510.08480

Peersman, C., Daelemans, W., Vandekerckhove, R., Vandekerckhove, B., & Van Vaerenbergh, B. (2016). The effects of age, gender and region on non-standard linguistic variation in online social networks. Retrieved from https://arxiv.org/abs/1601.02431

Pilkington, J. (1992). "Don't try and make out that I'm nice!": The different strategies women and men use when gossiping. *Wellington Working Papers in Linguistics. 5*, 37-60.

Press, L. (1993). Before the altair: The history of personal computing. *Communications of the ACM*, *36*(9), 27-33. http://dx.doi.org/10.1145/162685.162697

Provine, R. R., Spencer, R. J., & Mandell, D. L. (2007). Emotional expression online: Emoticons punctuate website text messages. *Journal of Language and Social Psychology, 26*(3), 299-307.

Reddit comment formatting. (2011). Retrieved from https://www.reddit.com/r/raerth/comments/cw70q/reddit_comment_formatting

Reddiquette. (2015.). Retrieved from https://www.reddit.com/wiki/reddiquette

Reddit FAQ (2015). Retrieved from https://www.reddit.com/wiki/faq

Reddit Metrics (n.d.). *Reddit Metrics*. Retrieved May 7, 2016, from http://redditmetrics.com/

Rezabeck, L. L., & Cochenour, J. J. (1994). Emoticons: Visual clues for computer-mediated communication. In *Imagery and visual literacy: Selected readings from the Annual conference of the International Visual Literacy Association*, 371-383.

Riva, G. (2002). The sociocognitive psychology of computer-mediated communication: The present and future of technology-based interactions. *CyberPsychology & Behavior, 5*(6), 581-598.

Romaine, S. (2003). Variation in language and gender. In J. Holmes & M. Meyerhoff (Eds.). *The handbook of language and gender* (pp. 98-118). Malden, MA: Blackwell.

Rosenthal,S., & McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. Proceedings of the *49th Annual Meeting of the Association for Computational Linguistics*, pages 763–772, Portland, Oregon, June 19-24, 2011.

[/s] Urban Dictionary. (2007). Retrieved from http://www.urbandictionary.com/define.php?term=%5B%2Fs%5D

Samadder Rhik. (2010, January 30). “This.” sucks: why you need to stop using the internet's worst one-word sentence. *The Guardian*. Retrieved from http://www.theguardian.com/media/2015/jan/30/this-sucks-why-you-need-to-stop-using-the-internets-worst-one-word-sentence

Shaw, L. H., & Gant, L. M. (2002). Users divided? Exploring the gender gap in Internet use. *CyberPsychology & Behavior*, *5*(6), 517-527.

Shepherd, R. P. (2016). Men, women, and Web 2.0 writing: Gender difference in Facebook composing. *Computers and Composition, 39*, 14-26. http://dx.doi.org/10.1016/j.compcom.2015.11.002

Sherman, R. C., End, C., Kraan, E., Cole, E., Campbell, J., Birchmeier, Z., Klausner, J. (2000). The Internet gender gap among students: Forgotten but not gone? *CyberPsychology & Behavior, 3*(5), 885-894.

Shitposting. (n. d.). Retrieved from http://knowyourmeme.com/memes/shitposting

Singer, P., Flöck, F., Meinhart C., Zeitfogel, E. & Strohmaier, M. (2014). Evolution of reddit: from the front page of the internet to a self-referential community? Proceedings of the *23rd International Conference on World Wide Web,* pp. 517-522. Retrieved from http://dl.acm.org/citation.cfm?id=2576943

Site overview: Reddit.com. (n.d.). Alexa. Retrieved May 20, 2016 from http://www.alexa.com/siteinfo/reddit.com

Skovholt, K., Gronning, A., & Kankaanranta, A. (2014). The Communicative Functions of Emoticons in Workplace E-Mails: :-)∗. *Journal of Computer-Mediated Communication*, *19*, 780-797.

Smith, C. B., McLaughlin, M. L., & Osborne, K. (1997). Conduct control on Usenet. *Journal of Computer-Mediated Communication*, 2(4). doi: 10.1111/j.1083-6101.1997.tb00197.x

Spender, D. (1980). *Man Made Language.* London: Routledge.

Squires, L. M. (2007) Whats the use of apostrophes? Gender difference and linguistic variation in instant messaging. *American University TESOL Working Paper, 4*. Retrieved from http://hdl.handle.net/1961/5240

Sunderland, J. (2006). *Language and gender: An advanced resource book*. London: Routledge.

Tagliamonte, S. A, & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech, 83*(1), 3-34. doi: 10.1215/00031283-2008-001

Tannen, D. (1990). *You just don't understand: Women and men in conversation*. New York: Morrow.

TBH. (2003). Retrieved from http://fr.urbandictionary.com/define.php?term=tbh

The Global Gender Gap Report. (2015). Retrieved from http://reports.weforum.org/global-gender-gap-report-2015/

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

Thompson, D. & Filik, R. (2016). Sarcasm in written communication: Emoticons are efficient markers of intention. *Journal of Computer-Mediated Communication, 21,* 105-120. doi:10.1111/jcc4.12156

Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology, 40*(2), 193-208

Top Subreddits. Retrieved on May 8, 2016, from http://redditmetrics.com/top/offset/100%20May8

Thurlow, C., Lengel, L, & Tomic, A. (2004). *Computer mediated communication*. London: SAGE Publications Ltd.

Tossell, C. C., Kortum, P., Shepard, C., Barg-Walkow, L. H, Rahmati, A., & Zhong, L. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior, 28*, 659-663. doi:10.1016/j.chb.2011.11.012

Trudgill, P. (1974). The co-variation of phonological variables with social parameters. In P. Trudgill (Ed.), *The social differentiation of English in Norwich*, (pp. 90-95). Cambridge: Cambridge University Press

Trudgill, P. (1983). *Sociolinguistics: An introduction to language and society*. Harmondsworth, England: Penguin.

Truss, L. (2004). *Eats, shoots & leaves: The zero tolerance approach to punctuation*. New York: Gotham Books.

Valdes, M. (2013). Innocents accused in online manhunt. Retrieved from http://www.newshub.co.nz/technology/innocents-accused-in-online-manhunt-2013042212#axzz4AvJzibLJ

Van Alphen, I. (1987). Learning from your peers: The acquisition of gender-specific speech styles. In D. Brouwer & D. De Haan (Eds.), W*omen's language, socialization and self-image* (pp. 58-75). Dordrecht: Foris.

Van Der Meij, H. (2007). Differences in CMC and does elementary school children's e-mail fit this picVture? *Sex Roles, 57,* 341-354. doi 10.1007/s11199-007-9270-9.

Van Mieghem, P. Human psychology of common appraisal: The Reddit score. *IEEE Transactions on Multimedia, 13*(6), 1404-1406. doi: 10.1109/TMM.2011.2165054

Varnhagen C. K., McFall, G. P., Pugh, N., Routledge, L. Sumida-MacDonald, H., Kwong, T. E. (2010). lol: new language and spelling in instant messaging. *Read Writ*, *23,* 719-733. doi: 10.1007/s11145-009-9181-y

Vertical Posting. (2015). Retrieved from http://knowyourmeme.com/memes/vertical-posting

Vidal, L. Ares, G., & Jaeger, S.(2016). Use of emoticon and emoji in tweets for food-related emotional expression. *Food Quality and Preference, 49,* 119-128.

Wang L., .Zhan Y., Li Q., Zeng, D. D., Leischow, S. J. & Okamoto, J. (2015). An examination of electronic cigarette content on social media: Analysis of e-cigarette flavor content on Reddit. *International Journal of Environmental Research and Public Health, 12*, 14933-14952. doi:10.3390/ijerph121114933

Wasike, B. S. (2011). Framing social news Sites: An analysis of the top ranked stories on Reddit and Digg. *Southwestern Mass Communication Journal, 27*(1), 57-67.

Wasserman, I. M. & Richmond-Abbott, M. (2005). Gender and the Internet: Cause of variation in access, level, and scope of use. *Social Science Quarterly, 86(1*), 252-270.

Weiser, E. B. (2000). Gender differences in Internet use patterns and Internet application preferences: A two-sample comparison. *CyberPsychology & Behavior, 3(2)*, 167-178.

Weinberger, M. (2015). Reddit's new CEO is willing to sacrifice revenue for freedom of speech. Retrieved from http://uk.businessinsider.com/reddit-new-content-policy-2015-7?r=US&IR=T

What does /s mean? (2014). Message posted to https://www.reddit.com/r/OutOfTheLoop/comments/1zo2l4/what_does_s_mean/

Witmer, D. & Katzman, S. L. (1997). On-line smiles: Does gender make a difference in the use of graphic Accents*? Journal of Computer-Mediated Communication, 2*(4). doi: 10.1111/j.1083-6101.1997.tb00192.x

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing: A corpus based study. In S. C. Herring (Ed.) *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 29-46).

Zhang, H. & Setty, V. (2016). Finding diverse needles in a haystack of comments: social media exploration for news. WebSci ’16: Proceedings of the 8th ACM Conference on Web Science, 286-290. doi: 10.1145/2908131.2908168

Appendix: List of Netspeak lexical features, adapted from Wikipedia and other sources

Non-standard spellings and letter homophones (alpha-numerical substitutions)
2 “too” or “to”; 4” for; 10q “thank you”; 10x “thanks”; A C? “AH! Si?” B4 “before”; CU “see you”; F9 “fine”; KTHX “OK, thanks”; kthxbye “OK, thanks, goodbye”; k or kk “OK”; l8r “later”; M8 “mate”; O RLY “Oh, really?”; OIC “Oh, I see”; n0rp “porn”; NE1 “anyone”; pr0n “porn”; pwnd/pwn intentional misspelling of “owned”; U “you”;
Clippings/abbreviations
Dafuq ”what the fuck; “ gratz “congratulations”
Phrases/words
Bae “babe”;Facepalm “sign of frustration or agitation”; Headdesk “extreme frustration” ONOZ “Oh, no”; Lulz, corruption of lol; Rehi “Hello again”; I can’t even ““I am overwhelmed with emotion, I am unable to face the situation”; This. affirmation of agreement; because (+ noun)All the feels “strong emotional response”
Acronyms and abbreviations
3DPD “3-dimensional pig disgusting”; AAF “as a friend” ADAD “another day another dollar”; “ADIH “another day in hell”; ADIP “another day in paradise”; AEAP “As Early As Possible”; AFAICR “as far as I can recall / remember”; “AFAICS: As far as I can see*;* AFAICT “as far as I can tell”; AFAIK “as far as I know”; AFAIR “as far as I remember”; AFAP “as far as possible”; “AFK: Away from keyboard”; ALOL “Actually laughing out loud”; AMA “ask me anything” ASAP “as soon as possible”; ASL or A/S/L: Age / sex / location”; ASLP or A/S/L/P “age, sex, location, picture”; ATEOTD “at the end of the day”; ATB “all the best”; ATM “at the moment”; AWOL “absent Without (Official) Leave”; BBIAB “be back in a bit”; BBL/BBS “be back later / shortly / soon”; BCNU “be seein' you”; BFF “best friends forever”; BFN “bye for now”; BOFH “bastard operator from hell”; BRB “be right back”; BSOD “Blue Screen of Death”; BTDT “been there done that”; BTTT “back to the top”; BTW “by the way”; CMIIW “Correct me if I'm wrong”; COB “close of business”; CYA “see ya” or “cover your ass”; DAE “does anyone else”; DFTT “don't feed the trolls”; DFTBA “don't forget to be awesome”; DGAF “don't give a fuck”; DIAF Die in a fire”; DILLIGAF/D/S Does it look like I give a flip / fuck / damn / shit”; D/L: Download”; DM “direct message”; DND “do not disturb”; DOA “Dead on arrival”; ELI5 “explain like I’m five”; EOM “end of message”; ESAD “eat shit and die”; ETA “estimated time of arrival” or “edited to add”; FAQ “frequently asked question(s)”; FFS “for fuck's sake”; FMCDH “from my cold dead Hands; “FML “fuck my life”; FOAD “fuck off and die”; FOAF “friend of a friend;FTFY “fixed that for you”; FTR “for the record”; FTL “for the loss”; FTW”for the win”; FTW? “Fuck the what?”, reversal of WTF?; FU “fuck you”; FUBAR “fucked up beyond all recognition / repair”; FUD “fear, uncertainty and Doubt”; FWIW “for what it's worth”; FWP “first world problems”; FYI “for your information”; GBTW “get back to work”; GF “great/good fight or “girlfriend”; BF “boyfriend”; GFU “good for you”; GFY “go fuck ourself”; GMTA “great minds think alike”; GTFO “get the fuck out”; GTG or G2G “Got to go” or “Good to go”; GR “good race”; GIYF “Google is your friend”; HAND “have a nice day”; HF “have fun”; HIFW “how I felt when”; HTH “hope this / that helps”; IANAL “I am not a lawyer”; IBTL “in before the lock”; ICYMI “in case you missed it“; IDGAF “I don’t give a fuck”; IDK “I don't know”; IHT “I had to”; IONO “I don't know”; IIRC “if I recall / remember correctly”; IIUC “if I understand correctly”; ILY I love you”; IMAO “in my arrogant opinion”; IMO “in my opinion”; IMHO “in my humble / honest opinion”; IMNSHO “in my not so humble opinion”; IOW “in other words”; IRL “in real life”; ISTM It seems to me”; ITT “in this thread”; ITYM” I think you mean”; IWSN “I want sex now”; IYDMMA “iIf you don't mind my asking” IYKWIM If you know what I mean; JAS “Just a sec”; JFTR Just for the record”; JK or j/k Just kidding, or joke”; JFGI “jus fucking google it”; JSYK “just so you know)”; KISS “keep it simple stupid”; KOS “kill on sight”; LFG “looking for group”; LFM “looking for more”; LMAO “laughing my ass off”; LMBO “laughing my butt off”; LMFAO “laughing my fucking ass off” LMGTFY “let me Google that for you”; LMIRL “let's meet in real life”; LMK “let me know”; LOL “laughing out loud, laugh out loud”; LTNS “long time no see”; LYLAB “love you like a brother”; LYLAS “love you like a sister”; MFW “my face when”; MIRL “me in real life”; MOTD “message of the day”; MRW “my reaction when”; MTFBWY “may the Force be with you”; MUSH “multi-user shared hallucination”; MYOB “mind your own business”; NGL “not gonna lie”; NIFOC “naked in front of computer”; NM Sometimes written N/M) “Not much”, “never mind” or “no message”; NP “no problem”; NS “nice shot”; NSOH “no sense of humor”;

NSFL “not safe for life)”; NSFW “not safe for work”; NVM, NVMD, or nm “nevermind or “not much”; OC “original content”; OFN
“old freaking news”; OMG “Oh my god”; OMFG “oh my fucking god”; OMGWTF “Oh my God what the fuck”; OMGWTFBBQ “Oh my God what the fuck barbecue”; OMW “On my way” or “Oh my word”; OP “Original poster / Operator / Outpost / Overpowered”; OT “Off topic”; OTB “off to bed”; OTOH “on the other hand”; OTP “on the phone” or “one true pairing or “on the piss”; PAW “parents are watching”; PEBKAC/PEBCAK “problem exists between keyboard and chair”; PITA “Pain in the ass”; PLMK “please let me know”; PMSL “pissing myself laughing”; POS “piece of shit” or “parent over shoulder”; POTS “plain old telephone service”; POV “point of view”; PPL “people”; PTKFGS “punch the Keys For God's Sake”; QFT “Quoted for truthiness”; QWP (texting) “Quit Whining, Please”; RL “real Life”; RMS “ride me sideways”; ROFL/ROTFL “rolling on (the) floor laughing”; ROFLMAO/ROTFLMAO “rolling on (the) floor laughing my ass off”; ROFLMAOWPIMP/ROTFLMAOWPIMP “rolling on (the) floor laughing my ass off while peeing in my pants”; ROFLOL/ROTFLOL “rolling on (the) floor laughing out loud”; RSN “real soon now”; RTFB “read the fucking binary/ book”; RTFS “read the fucking source”; RTFM/RTM “Read the (fucking/fine) manual or reboot the (fucking) machine”; SCNR “sorry, could not resist”; SFW “safe for work”; sk8/sk8r/sk8er “skate/skater”; SMH “shaking my head”; SNAFU “situation normal: all (fucked/fouled) up”; SO “significant other”; SOHF “sense of humor failure”; SOS “someone (watching) over shoulder”; STFU “shut the fuck up”; STFW Search the fucking web”; TANSTAAFL “there ain't no such thing as a free lunch”; TBF “time between failures or “to be frank”; TBH “to be honest”; TFW “that feel when”; TG “that's great”; TGIF “thank god it's Friday”; TIA “thanks in advance”; TIL “today I learned”; TIMWTK “this inquiring mind wants to know”; TINC “there Is No Cabal, a term discouraging conspiracy theories”; TINS “there is no spoon”; TL;DR “too Long; Didn't Read”; TMI: “too much information”; TOS: “terms of service”; TTBOMK “to the best of my knowledge”: TTFN “ta ta for now”; TTYL “talk to you later”; TTYN “talk to you never”; TTYS “talk to you soon”; TY “thank you”; TYT “take your time”; TYVM “thank you very much”; UTFSE “use the fucking search engine”; UGO “you got owned”; URS “you Really Suck”; , w00T or WOOT, backronym for the term "we own the other team"; W/ or W/O: With or without”; WB “welcome back”; WBU “what (a)bout you”; W/E “whatever or “weekend”; WRT “with respect / regard to”; WTB “want to buy”; WTF “what the fuck”; WTG “way to go”; WTH “what the hell”; WTS “want to sell”; “WTT “want to trade”; WUG “what you got?”; WUBU2 “what (have) you been up to?”; WUU2 “what (are) you up to?”; WYSIWYG “What you see is what you get”; W8 Wait”; YAGNI “you’re gonna need it”; YAGTOH: you are going to own him” YGM “you've got mail”; YHBT You have been trolled”; YKW “you know what?”; YMMV “your mileage may vary”YOLO “you only live once”; YOYO “you're on your own”; YTMND “you're the man now, dog”; YW “you're welcome”; ZOMG an intentional misspelling of "oh my god”

Ouverture sur la thèse

Etude diachronique de la langue de Reddit

Marie Flesch

Erudi- Master Mondes anglophones

Session 2 2015-2016

Table des matières

Introduction……..……1

I. Présentation de Reddit……...6

1. Pourquoi Reddit……6
2. Une communauté de pratique…….……..6
3. Etudier la variation linguistique sur le court terme sur Reddit ?..... 8

II. Méthodes……..9
1. Recueillir les données…….……..9
2. Définir les périodes à étudier……..…10
3. Choisir les échantillons……...11

III. Analyse du corpus……..16
1. Outils……...……16
2. « Corpus-driven » ou « corpus-based ? »……..……..16
3. Analyses quantitatives……17
4. Analyses qualitatives……..17
5. Analyse synchronique du corpus……18
6. Résultats : Expliquer les différences……..…18

IV. Considération éthiques……..19

Conclusion……..20

Bibliographie……...21

Résumé : Partant du constat qu'il existe peu d'études diachroniques de la langue d'internet, cette ouverture sur la thèse propose de créer deux ou trois corpus à partir du site web communautaire Reddit. Elle décrit les méthodes qu'il sera possible d'utiliser pour bâtir les corpus et les analyser.

Introduction

La « computer-mediated communication » ou CMC fait aujourd'hui partie intégrante du quotidien de près de trois milliards d'individus. Les emails, forums de discussion et sites web figurent parmi les premiers modes de communication à avoir été utilisés par le grand public sur Internet dans les années 1990. Depuis, ils ont été rejoints par une foule de nouveaux outils et plateformes, qui sont apparus au fur et à mesure des innovations technologiques, notamment grâce à la révolution de l'internet mobile.

Décrite, lors de son émergence, comme un mélange de langue parlée et de langue écrite, la langue que l'on utilise sur Internet a ensuite été considérée comme un « troisième médium » à part entière (Crystal, 2001). Les chercheurs qui sont intéressés à cette langue naissante ont tenté d'identifier ses caractéristiques, et se sont rendu compte qu'elle était éminemment hétérogène. Tout comme il y a des différences de registres et d'usages dans l'oral et l'écrit, la langue du web prend de multiples formes, que les internautes se sont habitués à manier en surfant d'un site et d'une plateforme à l'autre, depuis leur smartphones, leurs ordinateurs ou leurs tablettes. Ils écrivent des emails, envoient des SMS, rédigent des statuts Facebook, twittent leurs découvertes, partagent les publications Instagram de leurs connaissances virtuelles, commentent des vidéos sur YouTube, se donnent rendez-

vous sur Snapchat ou Periscope, s'informent et posent des questions sur les forums, plateformes collaboratives et listes de discussion. Ils fréquentent ainsi une multitude de communautés qui possèdent leurs propres pratiques linguistiques. En passant d'un medium de communication à l'autre, ils apprennent de nouveaux codes et développent de nouvelles compétences. Parmi les signes les plus visibles de ces différences, citons, par exemple, l'omniprésence des hashtags sur Twitter et Instagram, l'abondance de gifs, ou images animées, sur Tumblr, ou encore le phénomène du « green text » propre à 4Chan, utilisé pour raconter des anecdotes sous forme de phrases très concises. Les contraintes techniques propres à chaque outil ou site façonnent en partie la langue qu'on y utilise ; les utilisateurs de Twitter, qui n'autorise qu'un maximum de 140 caractères, doivent ainsi s'adapter au besoin de concision, tandis que la prise en charge des *emojis* par les systèmes d'exploitation mobiles a provoqué un essor de ces caractères Unicode représentant des images ou symboles.

La variation linguistique, telle qu'elle s'affiche sur le web et sur d'autres plateformes de CMC, intéresse tout particulièrement la linguistique de corpus. Des corpus dédiés à différents genres sont aujourd'hui mis à la disposition des chercheurs, journalistes, enseignants, étudiants, et internautes, comme, par exemple, le corpus Enron, qui contient plus de 600 000 emails écrits par 158 employés d'Enron, et qui a été rendu public après l'enquête dont la société a fait l'objet. Plusieurs corpus ont été créés par Mark Davies à la Brigham Young University à partir du web : un corpus de Wikipedia, qui contient près de deux milliards de mots, le Corpus of Global Web-Based English, qui vise à représenter vingt variétés d'anglais, ou encore le NOW Corpus, un corpus de trois milliards de mots issus de journaux en ligne. Une autre approche consiste à considérer le web comme un corpus. Pour faciliter les recherches, l'outil WebCorp a été mis au point ; contrairement aux navigateurs internet classiques, qui fournissent les résultats les plus pertinents, WebCorps effectue des recherches purement linguistiques sur le web et en extrait des concordances.

De leur côté, les chercheurs n'ont pas attendu ces initiatives pour créer eux-mêmes des corpus afin de répondre aux besoins de leurs projets. Les premières études sur la langue d'Internet remontent ainsi aux années 1990, avec, notamment, l'étude exploratoire de Werry sur le chat (1996), ou le travail de Davis et Brewer sur ce qu'ils appellent l'« electronic discourse ». La recherche s'est intéressée à

différents types de plateformes, à commencer par les emails (Baron, 1998 ; Skovholt et al. 2014), les SMS (Tossell et al., 2012 ; Kirsten-Torrado, 2012), la messagerie instantanée (Baron, 2004 ; Fox et al. 2007 ; Tagliamonte & Denis, 2008 ; Varnhagen et al. 2010 ; Garrison, Remley, Thomas & Wierszewski, 2011 ; Sanchez-Moya & Cruz-Moya, 2015), les blogs (Huffaker & Calvert, 2005 ; Rosenthal & McKeown, 2011) et les réseaux sociaux comme MySpace (Thelwall, 2008) ou Twitter (Park et Barash, Fink, & Cha, 2013 ; Bamman, B., Eisenstein, J., & Schoebelen, 2014 ; Pavalanathan & Eisenstein 2015). Certaines études ont examiné les différences entre plusieurs plateformes, comme celle de Lin et Qiu (2013), qui ont comparé Facebook et Twitter. Les forums de discussion ont eux aussi fait l'objet de recherches, tout particulièrement les newsgroups USENET, qui diffèrent des forums du web par leur fonctionnement et par la façon dont y accède (Herring, Johnson & DiBenedetto 1992 ; Rezabek & Cochenour, 1994 ; Witmer & Katzman, 1997 ; Wolf, 2000, par exemple).

Les forums du web paraissent avoir été moins étudiés, même si Claridge (2007) a noté leur potentiel et a exposé sa méthode pour construire un corpus. Reddit, notamment, ne semble avoir pas avoir fait l'objet d'une étude approfondie d'un point de vue linguistique. Le site, qui a été créé en 2005, est aujourd'hui une des plus importantes communautés en ligne du web, surtout aux Etats-Unis où, selon la société Alexa, il est le neuvième site le plus consulté (Site overview, Reddit). Reddit prend la forme d'une multitude de sous-forums, appelés subreddits, sur lesquels les utilisateurs, ou Redditeurs, peuvent publier des liens vers des articles, vidéos ou photos, ou encore écrire des « self-posts » et poser des questions à la communauté. Autorégulé par des modérateurs bénévoles et par les Redditeurs eux-mêmes, Reddit est célèbre pour la liberté de ton qui y règne, qui peut parfois basculer du côté de la violence, du racisme ou de la misogynie : tout y est permis, ou presque, puisque, récemment, les dirigeants du site ont imposé des sanctions aux subreddits les plus violents.

Même si, ces cinq dernières années, la cadence des publications sur Reddit s'est accélérée, la majorité des articles et des projets s'inscrit dans le champ de recherche de la linguistique informatique et de l'informatique. Citons, par exemple, le travail de Haralabopoulosa, Anagnostopoulosa, et Zeadallyb, qui se sont intéressés à la façon dont les informations se diffusent sur le site (2015), l'article de Noguti (2016), qui a mis au point un modèle statistique afin d'étudier la relation entre la langue utilisé

dans les messages et l'implication des utilisateurs, ou encore le travail de McEwan (2016), qui s'est penché sur les marqueurs linguistiques des sous-forums les plus actifs. La langue propre à Reddit, en tant que communauté virtuelle, ne semble pas avoir fait l'objet de peu d'études ; Barbaresi (2015), par exemple, décrit la méthode qu'il a utilisée pour construire un corpus à partir de messages en allemand. Et pourtant, Reddit apparait comme un lieu tout indiqué pour ce genre de travail : le site représente une source de données immense, il est entièrement public, et son interface de programmation est mise à disposition de tous. Sur le plan linguistique, les Redditeurs utilisent bien souvent termes argotiques, acronymes, émoticons et autres caractéristiques lexicales du « Netspeak ».

Par ailleurs, il semble que peu de chercheurs se soient intéressés à la variation diachronique de la langue du web. On ne parle pas ici de textes destinés à d'autres supports et qui ont été numérisés, qui permettent d'étudier la variation sur le long terme, mais de contenus spécifiquement écrits pour le web. Il ne s'agit donc pas d'étudier la variation diachronique sur le long terme, mais sur le court terme. Quelques études ont adopté cette seconde approche ; on peut citer, par exemple, l'article de Kirsten-Torrado (2012), qui s'est penchée sur deux corpus de SMS produits à dix années d'écart, ou encore le travail de Pavalanathan et Eisenstein (2015), qui ont comparé l'évolution de la fréquence des émoticons dans un corpus de tweets produits entre 2014 et 2015. Il parait dommage que ce type d'étude soit rare, car il suffit de comparer les premières recherches effectuées sur la langue d'Internet dans les années 1990 avec les articles publiés depuis les années 2000 pour constater les différences, et parfois les fossés, qu'il y a. Les émoticons par exemple, ont connu une évolution remarquable ; les premières études sur le sujet, comme Rezabeck et Cochenour en 1994, n'en recensaient qu'une poignée ; aujourd'hui, il en existe des milliers (Park et al., 2013).

Dans la continuité du travail entamé en Master 2, ce projet de thèse s'intéresse à langue du web, mais adopte cette fois-ci une perspective diachronique pour étudier les pratiques linguistiques des Redditeurs. Il se propose de dresser le portrait d'un type d'Internet English, le « Redditese », et de cerner son évolution sur le court terme, c'est-à-dire dans les onze ans d'existence du site. Il pourrait tenter de répondre à plusieurs questions différentes. Les analyses statistiques permettraient de mesurer la montée en popularité de certains termes, de visualiser le déclin d'autres phénomènes, et identifier les zones du lexique les plus favorables aux changements.

Elles pourraient également permettre de comprendre comme le « jargon » propre à Reddit, qui semble essentiellement constitué d'acronymes, s'est constitué, ou de cerner le rôle de certaines sous-communautés dans la diffusion de certains phénomènes de « Netspeak ». Il pourrait aussi, dans la mesure où il sera possible de récolter des informations démographiques sur les Redditeurs, étudier les variations synchroniques, entre les genres, entre les différentes variétés d'anglais, ou entre les subreddits. Pour ce faire, la constitution de deux ou trois corpus est envisagée ; dans l'idéal, elle serait très soigneuse, et tenterait autant que faire se peut de donner une image représentative du site.

Cette ouverture sur la thèse commence par présenter Reddit comme une communauté de pratique, intéressante sur le plan sociolinguistique. Ensuite, elle s'intéresse aux méthodes qu'il sera possible d'utiliser pour, d'une part, créer un corpus représentatif, et, de l'autre, l'analyser et en interpréter les résultats.

I. Présentation de Reddit

1. Pourquoi Reddit

Plusieurs caractéristiques de Reddit font de ce site une excellente source de données pour ce projet de thèse. Tout d'abord, Reddit est extrêmement populaire, et contient une énorme quantité de texte. Dans l'article de blog qui célèbre le dixième anniversaire du site, les administrateurs du site précisent que, à la date du 21 juin 2015, quelques 190 millions de posts et plus d'1,7 milliard de commentaires ont été publiés sur Reddit. Même si tous ne sont pas actifs et si bon nombre ont été supprimés, plus de 800 000 subreddits ont été créés, ainsi que plus de 36 millions de comptes (Happy 10th birthday to us, 2015). Evidemment, cela ne signifie pas que Reddit est représentatif de l'ensemble des internautes ; le site reflète surtout les usages d'un public américain, jeune et masculin. Plus de 50 % des Redditeurs sont basés aux Etats-Unis (Alexa) ; l'Inde et le Royaume-Uni arrivent loin derrière, représentant un peu plus de 5 % des visiteurs chacun, suivi par le Canada et l'Australie. Un rapport du think tank américain Pew Research Center indique par ailleurs que, même si de plus en plus de femmes fréquentent le site, quelques sept Redditeurs américains sur dix sont des hommes, et que 56 % des utilisateurs sont âgés de 18 à 29 ans. Non seulement Reddit est une des plus grandes communautés en ligne du monde, mais c'est aussi un espace propice aux échanges détenus et informels, et à l'utilisation d'abréviations, d'acronymes, d'émoticons, et d'autres types de termes, ce qui en fait une matière première idéale de données pour l'étude du phénomène « Netspeak ».

2. Une communauté de pratique

Reddit n'est pas une « speech community », car cette notion, utilisée par des sociolinguistes comme Labov et Trudgill, désigne des populations situées dans un lieu bien précis, un quartier ou un lieu de travail par exemple. Une « speech community » est par ailleurs définie par sa démographie, c'est-à-dire par des catégories comme la classe, le genre, l'âge ou l'ethnicité. Les Redditeurs, par contraste, proviennent de lieux différents et ont des profils divers. On pourrait peut-être en revanche dire que Reddit est une « communauté de pratique », un concept proposé par les anthropologues Jean Lave et Etienne Wenger en 1991 qui en avaient fait le principe de base de leur théorie sociale de l'apprentissage, et qui a été repris

par les sociolinguistes Eckert et McConnell-Ginet (1992). Afin d'étudier la relation entre le genre et la langue, elles définissent la communauté de pratique comme un groupe d'individus réunis par une activité commune, et donnent comme exemple une équipe de bowling, un club de lecture, une famille ou une église. C'est donc en partageant des pratiques qui fait que l'on appartient à ce type de communauté, et non pas en appartenant à une classe sociale donnée, ou en passant du temps dans un lieu donné. Les membres d'une communauté de pratique ont en commun des façons de faire et de parler, des croyances et des valeurs, et de relations de pouvoir (p. 8). On appartient rarement à une seule communauté de pratique, et l'ensemble des communautés que l'on fréquente participe à la construction de l'identité individuelle. La communauté de pratique, ajoute Eckert, est le lieu privilégié du processus de la construction identitaire et langagière (2006, p. 111).

Reddit peut être défini comme une communauté de pratique parce que le site répond aux caractéristiques indispensables définies par Eckert : les internautes y partagent une expérience commune, sur la durée, et se positionnent comme un groupe distinct, comme on peut notamment le constater dans la Foire Aux Questions de Reddit (FAQ, 2015). Tout d'abord, même si le site est public, il est nécessaire d'y avoir créé un compte pour pouvoir y commenter. Les pratiques de Reddit se différencient de celles des réseaux sociaux car l'anonymat y est la règle, et que la création de comptes multiples y est encouragée : la FAQ recommande ainsi de créer des « throwaway accounts », des comptes à usage unique, quand on souhaite partager des informations sensibles. Même si les administrateurs du site sont intervenus dans le passé pour supprimer des subreddits violents et racistes, le site est autorégulé. Les modérateurs des différents subreddits sont des bénévoles, et tous les Redditeurs peuvent participer à la modération du site en utilisant le « downvoting », c'est-à-dire en votant contre un contenu ou un commentaire injurieux, hors-sujet ou promotionnel. Ils peuvent également utiliser le « report button » pour signaler aux modérateurs d'un sous-forum un commentaire qui enfreint les règles du site. Plusieurs mesures ont par ailleurs été mises en place pour pousser les Redditeurs à contribuer de façon positive au site : en partageant un lien qui intéresse la communauté, on obtient du « karma », qui est comptabilisé et affiché sur la page de chaque utilisateur. Certains subreddits récompensent les meilleurs contributeurs par des trophées. Il est également possible de profiter d'une version « Premium » de Reddit en achetant du « Reddit Gold », c'est-à-dire en payant un abonnement.

Evidemment, il ne faut pas oublier que Reddit est avant tout une entreprise, et que son but est de gagner de l'argent, par la publicité et les liens sponsorisés d'un côté, et par les abonnements de ses membres de l'autre.

Sur le plan linguistique, les éléments les plus aisément identifiables du « jargon » de Reddit sont l'utilisation d'acronymes inspirés par les subreddits les plus populaires, comme *DAE* « Does anyone else », *AMA* « Ask me anything », ou encore *ELI5* « Explain like I'm 5 ». La Foire aux questions de Reddit liste ainsi vingt-et-un termes que les nouveaux venus pourraient ne pas connaitre. Tous ces éléments renforcent l'aspect « communautaire » du site, et justifient l'étude sociolinguistique de la langue de Reddit comme étant la langue utilisée par une véritable communauté.

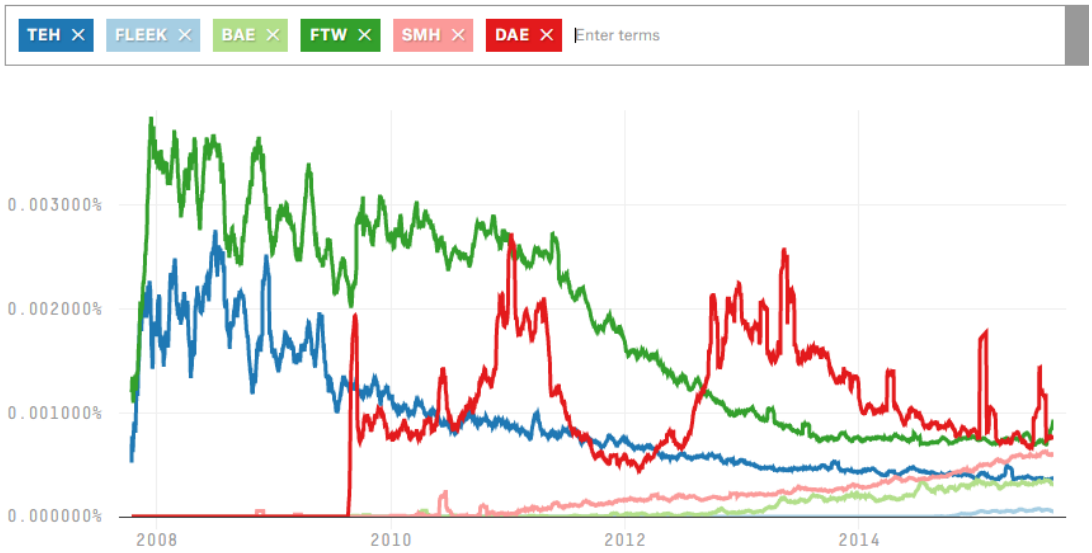
3. Etudier la variation linguistique sur le court terme sur Reddit ?

On pourrait légitimement se demander si Reddit, qui existe depuis 2005, mais dont les commentaires ne sont archivés que depuis octobre 2007, se prête à une analyse diachronique. On a souligné la rapidité des changements linguistiques sur Internet (Crystal, 2001, par exemple), mais qu'en est-il réellement ? Plusieurs études adoptant une perspective diachronique ont été réalisées, qui mettent en évidence des changements significatifs. Kirsten-Torrado (2012) a ainsi comparé deux corpus de SMS britanniques datant de 2000 et de 2010. Même si ces corpus sont de taille réduite et proviennent de sources différentes, ils semblent montrer que le phénomène de l'allongement expressif des mots, désigné par l'expression « stylish talk » par la chercheuse, a pris de l'ampleur depuis 2000. Pavalanathan et Eisenstein (2016) ont quant à eux étudié l'utilisation d'émoticons sur Twitter, dans un corpus de tweets datant de février 2014 à 2015. Sur cette durée relativement courte, ils ont remarqué des différences significatives dans le nombre d'émoticons utilisés : les émoticons sont de plus en plus boudés par les adeptes du site de microblogage, ce que Pavalanathan et Eisenstein expliquent par l'introduction des emojis sur Twitter, qui semblent remplacer les émoticons.

Dans le cas de Reddit, aucune étude diachronique ne semble avoir été réalisée. On peut toutefois mesurer la popularité de certains termes sur le site grâce à l'outil n-gram conçu par Olson et King (2015). La figure 1 montre l'évolution de la fréquence de six termes de « Netspeak », choisis un peu au hasard. On y voit notamment le déclin de *Teh*, une corruption intentionnelle du déterminant « the », et de *FTW* ou « For the win », un acronyme généralement utilisé pour montrer son enthousiasme.

On remarque également la brusque apparition et popularité de *DAE*, ou « Does anybody else ? », suite à la création du subreddit du même nom en 2010. Enfin, la figure 1 illustre la montée en popularité de l'acronyme *SMH*, ou « shaking my head », qui exprime la consternation et du terme *bae*, souvent utilisé comme synonyme de « baby ». Le graphique ci-dessous témoigne aussi de la très récente apparition de l'expression *on fleek*, qui veut dire « looking good », qui trouverait ses origines dans l'anglais vernaculaire afro-américain et a été récemment été popularisé sur les réseaux sociaux (Eyebrows on Fleek, 2015). Evidemment, cet outil ne permet d'obtenir qu'une analyse assez grossière de ces phénomènes. Dans le cas de *DAE*, par exemple, les fréquences affichées semblent prendre à la fois en compte les titres des fils de discussion, et non seulement les messages, ce qui fausse les résultats. Le petit test effectué montre toutefois qu'évolution il y a ; celle-ci n'est sans doute pas cantonnée à l'utilisation d'acronymes et autres termes de « Internet slang ».

Figure 1. Fréquence de six termes de « Netspeak » sur Reddit, d'octobre 2007 à août 2015.



II. Méthodes : Création du corpus

1. Récolter les données

Afin d'éviter de perdre du temps, et d'égarer en cours de route des éléments intéressants l'étude, il sera nécessaire de bien réfléchir à la façon dont les données seront sélectionnées, récoltées et préparées pour l'analyse. Deux approches différentes pourraient être envisagées pour créer un corpus diachronique de Reddit.

Tout d'abord, on pourrait compiler les données manuellement, en sélectionnant uniquement les pages pertinentes et en les copiant. L'inconvénient principal de cette méthode est sa lenteur. En outre, elle peut conduire à faire des erreurs, et à obtenir des échantillons biaisés. L'avantage, c'est qu'elle ne demande aucune compétence technique particulière.

La seconde approche consisterait à utiliser un « scraper », un programme informatique qui « aspire » du contenu des sites internet, ou se servir d'un script. Ce système permettrait toutefois de gagner du temps, et de construire un corpus plus important. Le problème, c'est qu'il faut être capable de programmer l'outil choisi afin de récolter uniquement le contenu désiré, ce qui nécessiterait une phase d'apprentissage. Il faut toutefois savoir que Reddit autorise le « scraping » et que de nombreux tutoriels, mis au point par des Redditeurs, sont disponibles sur le site. La suite de la présentation de la création du corpus ne prend pas en compte ce type de méthode, et se focalise sur des méthodes manuelles, non pas parce qu'elles sont forcément plus adaptées, mais parce qu'elles sont plus aisément maitrisables. Si cette ouverture sur la thèse débouchait sur un projet de thèse, il serait envisageable de s'intéresser aux méthodes de scraping et à la programmation. Dans tous les cas, quelle que soit la méthode retenue, le but est de créer un corpus de qualité, qui seul pourra permettre d'obtenir des résultats valides et de recueillir un maximum d'informations sur les données recueillies.

2. Définir les époques à étudier

Pour mener une étude diachronique de Reddit, il est nécessaire de créer non pas un, mais deux ou trois corpus. Il faut donc définir les périodes correspondantes à ces corpus. On peut immédiatement écarter les commencements du site, c'est-à-dire l'année 2005 et sans doute une partie de l'année 2006 car, à ses débuts, Reddit était purement un site de partage de lien ; on ne pouvait pas publier de commentaire ou écrire de « self-post », c'est-à-dire message sous forme de texte. Au lieu de choisir la période la plus ancienne possible, il faudrait choisir la période où de nombreux commentaires ont commencé à être publiés sur le site, ce qui permettrait de disposer d'un corpus suffisamment important. Un second corpus correspondrait à la période la plus récente possible. Pour permettre une analyse plus fine, il faudrait peut-être également envisager la création d'un troisième corpus intermédiaire.

3. Choisir des échantillons

Un corpus réalisé à partir d'un site comme Reddit, qui prend la forme d'une galaxie de forums, peut être conçu de plusieurs façons différentes. Herring (2004), décrivant une approche qu'elle appelle « computer-mediated discourse analysis » ou « CMDA », liste les différentes techniques d'échantillonnage possible sur le web. Elle explique qu'il est par exemple possible de choisir des messages au hasard, ce qui peut avoir pour avantage de créer un corpus représentatif, mais s'accompagne d'une perte de cohérence et de contexte. Ce n'est donc sans doute pas une méthode intéressante pour ce projet. On peut aussi sélectionner les messages par thème, c'est-à-dire en copiant tous les messages d'un fil de discussion donné, par date de publication, ou par individu ou groupe. Le tableau 1 résume les différents types d'approches possibles, détaillées ci-dessous.

Recueillir tous les messages publiés sur Reddit entre une date A et une date B.

Cette méthode semble tout indiquée pour une étude diachronique. Elle est également possible à mettre en œuvre sur Reddit. Même si les messages sont archivés au bout de quatre années, et n'apparaissent pas dans les subreddits ou dans l'historique des utilisateurs, on peut y accéder en intégrant une syntaxe Cloudsearch dans la barre d'adresse, et en précisant la période souhaitée à l'aide de l'heure Unix, qui indique le nombre de secondes écoulées depuis le 1er janvier 1970 (Use Cloudsearch, 2015). Le site Epoch Converter permet de convertir facilement les dates en heures Unix. On pourrait ainsi rechercher et intégrer au corpus tous les messages publiés entre deux dates sur l'ensemble de Reddit. Cela donnerait sans doute une image représentative de l'activité du site à un moment donné. Le problème, c'est que l'on se retrouverait avec du contenu issu de milliers de subreddits différents, produits par des centaines de milliers de Redditeurs. Il serait donc difficile d'obtenir des informations sur les auteurs des messages, ce qui pourrait limiter les possibilités d'interprétation.

Recueillir les messages d'un ou plusieurs forums entre une date A et une date B.

La méthode de recherche décrite ci-dessus peut également être limitée à un subreddit donné. A la place d'inclure dans le corpus des commentaires et des posts

issus de l'ensemble de Reddit, on pourrait ainsi ne sélectionner que le contenu de certains subreddits. Cette deuxième option permettrait d'avoir davantage de contrôle sur le contenu du corpus, même s'il faudra bien réfléchir à la façon dont les sous-forums sont sélectionnés, afin d'obtenir un corpus relativement représentatif. Les subreddits diffèrent en effet par bien des aspects. Certains sont fréquentés par des millions d'internautes. C'est par exemple le cas de /r/AskReddit, qui compte plus de 11 millions d'abonnés. Des dizaines de posts y sont publiés chaque jour ; ils reçoivent régulièrement des milliers de commentaires et des milliers de votes. Par contraste, quelques 38 000 Redditeurs sont abonnés à /r/askphilosophy, un sous-forum dont les messages récoltent rarement plus de cinquante commentaires. La longueur des messages et des commentaires varie également énormément : sur /r/relationships, par exemple, un subreddit où les internautes partagent les problèmes qu'ils rencontrent dans leur vie privée, de nombreux commentaires dépassent aisément les deux cents mots. Dans d'autres subreddits, comme /r/me_irl la concision est de rigueur. Il existe également des différences de style entre les subreddits. Certains, notamment ceux dédiés à des sujets « sérieux », à l'instar de /r/science, semblent adopter un style moins détendu et moins informel que d'autres subreddits comme /r/funny.

Tableau 1. Compositions possibles du corpus

	Corpus 1 : 2007	Corpus 2 : 2012	Corpus 3 : 2016
Par date	Tous les messages publiés sur le site du 1er janvier 2007 au 1er avril 2007	Tous les messages publiés sur le site du 1er janvier 2012 au 1er avril 2012	Tous les messages publiés sur le site du 1er janvier 2016 au 1er avril 2016
Par subreddits	Tous les messages publiés sur les subreddits suivants du 1er janvier 2007 au 1er avril 2007 : /r/AskReddit /r/AskPhilosophy /r/funny /r/ELI5 /r/zen etc.	Tous les messages publiés sur les subreddits suivants du 1er janvier 2012 au 1er avril 2012 : /r/AskReddit /r/AskPhilosophy /r/funny /r/ELI5 /r/zen etc.	Tous les messages publiés sur les subreddits suivants du 1er janvier 2016 au 1er avril 2016 : /r/AskReddit /r/AskPhilosophy /r/funny /r/ELI5 /r/zen etc.
Par utilisateurs	Tous les messages publiés par 500 utilisateurs actifs du 1er janvier 2007 au 1er avril 2007	Tous les messages publiés par 500 utilisateurs actifs du 1er janvier 2012 au 1er avril 2012	Tous les messages publiés par 500 utilisateurs actifs du 1er janvier 2016 au 1er avril 2016

Recueillir tous les messages mis en ligne par un individu entre une date A et une date B.

En effectuant une recherche sur Reddit à l'aide de la syntaxe Cloudsearch, on peut combiner plusieurs critères, et n'obtenir que les messages et commentaires publiés par un Redditeur donné, à une période donnée, sur l'ensemble de Reddit ou sur un subreddit donné. L'avantage, c'est que l'on pourrait récolter et exploiter des données démographiques plus aisément, en consultant l'historique de chaque utilisateur. Le problème, c'est que l'on ne pourrait pas visualiser les interactions entre les utilisateurs. Il faudrait également définir la façon dont seront sélectionnés les Redditeurs, et on voit difficilement comment bâtir un corpus représentatif de cette façon.

Le *diachronic sampling dilemma*

Lorsque l'on souhaite réaliser une étude diachronique, on se retrouve face à ce que Baker appelle le « *diachronic sampling dilemma* » (2010, p. 60). Pour s'assurer que les changements diachroniques sont bien responsables des différences constatées, il faut essayer de minimiser les interférences causées par d'autres facteurs, dus à la façon dont les corpus ont été construits. Cela implique de créer des corpus qui ont une structure comparable, et qui sont tous bâtis selon les mêmes règles. Si l'on décidait de n'utiliser que le contenu issu de certains subreddits, on pourrait aisément se retrouver confronté à ce problème. Un corpus consacré à l'année 2007, par exemple, qui contiendrait une proportion démesurée de contenu provenant de /r/science et /r/AskPhilosphy, différerait forcément d'un corpus principalement réalisé à partir de /r/relationships et /r/programming. Si on optait pour cette solution, il faudrait donc établir un modèle de corpus, qui serait utilisé pour la création des deux ou trois corpus. On devrait ainsi décider des subreddits que l'on inclurait dans le corpus, et du nombre de mots que l'on récolterait par corpus. On pourrait aussi donner une place plus importante aux forums les plus populaires, ou qui contiennent le plus de commentaires, afin de donner une image plus fidèle de l'ensemble du site. Cette méthode permettrait en outre d'effectuer une analyse synchronique en comparant plus aisément les différents subreddits.

L'autre facette du *diachronic sampling dilemma*, c'est qu'il est intéressant d'essayer d'obtenir une image la plus représentative possible d'une langue donnée à des moments donnés. Ce problème est moins aigu dans le cadre de ce projet que

lorsque l'on étudie la variation diachronique sur plusieurs siècles. On pourrait toutefois être amené à modifier le modèle d'un corpus à l'autre. Dans le cas de Reddit, le modèle pourrait être conçu de façon à inclure dans les corpus 2012 et 2016 des subreddits n'existant pas en 2007, ou à donner moins d'importance dans les corpus les plus récents à des subreddits qui ont joui d'une forte popularité dans les premiers temps du site avant d'être éclipsés par d'autres sous-forums. On pourrait imaginer classer les subreddits par catégories (sérieux, humour, conseils, sports, jeux vidéo, etc.) et choisir les subreddits les populaires. Le but serait de donner une image représentative du site à trois moments donnés dans le temps.

A ce stade préparatoire du projet de thèse, il est difficile de mesurer exactement les avantages et les limites de ces différentes méthodes de récolte de données. Dans l'idéal, pour faire le choix le plus adéquat, il faudra sans doute réaliser plusieurs corpus pilotes. Ainsi, on pourrait évaluer les atouts et les problèmes de chaque méthode, et évaluer le temps requis pour créer les deux ou trois corpus nécessaires.

4. La taille du corpus

De même, il n'est pas évident de répondre à la question de la taille du corpus. Biber (cité par McEnery, Xiao et Tono, 2006), estime que des échantillons de 2000 mots consécutifs suffisent à étudier les phénomènes les plus fréquents, mais cela n'est pas d'une grande aide dans le cas de ce projet. D'un côté, la vaste majorité des messages publiés sur Reddit font quelques centaines de mots tout au plus ; de l'autre, les phénomènes qui seront étudiés restent relativement rares, et il faudra sans doute recueillir une quantité importante de données. Il ne s'agit toutefois pas de créer le corpus le plus important possible ; il semble essentiel de privilégier la qualité des données récoltées sur la quantité.

5. Préparation du corpus

La création d'un corpus à partir d'un forum de discussion pose plusieurs problèmes. Ceux-ci ont été décrits par Claridge (2007), qui prend pour exemple la création d'un corpus capable de rendre compte de différences dans l'usage des pronoms personnels et de l'expressivité dans quatre variétés d'anglais différentes. Tout d'abord, elle explique qu'il convient de choisir une unité de base, qui peut être soit le commentaire, soit le fil de discussion. Cette seconde option lui semble préférable, parce qu'elle permet d'étudier les interactions entre les internautes.

Ensuite vient le problème des citations, c'est-à-dire des commentaires ou des fragments de commentaires précédemment publiés dans un fil de discussion, et qui sont utilisés par d'autres internautes dans leur message. Sur Reddit, cette pratique ne semble pas très répandue, mais il faudra bien entendu en tenir compte. Les deux derniers points tiennent à l'identité des auteurs des commentaires. Claridge conseille d'indiquer le genre des internautes quand c'est possible, soit parce qu'ils l'indiquent clairement, ou à partir de ce qu'ils dévoilent d'eux-mêmes. La question de la variété d'anglais utilisée est également importante. L'anglais américain est dominant sur Reddit, mais près de la moitié des visiteurs du site viennent d'autres pays du monde, anglophones ou non. Encore une fois, il faudra vérifier s'il est possible d'obtenir des informations sur les différents dialectes utilisés par les Redditeurs. Si c'est le cas, il serait sans doute judicieux de l'inclure dans l'étude.

Claridge propose par ailleurs une structure de balisage hiérarchique, qui permet de conserver le maximum d'informations possible sur les messages et leurs auteurs. Le système qu'elle utilise indique le sujet du message, la date à laquelle il a été publié, et, si possible le genre de son auteur et le lieu où il vit. Le balisage a également pour intérêt de pouvoir délimiter les différents paragraphes d'un message, de signaler les portions d'un message qui proviennent d'une citation d'un message précédent, et d'indiquer la place d'un commentaire dans le fil de discussion. Claridge précise qu'il est également possible d'inclure des informations sur les images et les hyperliens (p. 96-97).

Le type d'annotation suggéré par Claridge sera indispensable. Cette étape sera effectuée en en utilisant par exemple l'Extensible Markup Language, qui fournit des informations sur le texte à l'aide de balises encadrées par des chevrons. Il faudra toutefois identifier les types d'informations les plus pertinents pour ce projet. L'annotation des phénomènes graphiques sera nécessaire, car elle seule permettra de rendre compte de l'utilisation de différents formats de police, de « vertical posting », par exemple, ou d'images. Disposer d'informations sur la longueur des paragraphes et les auteurs des messages serait également judicieux. Il faudra par ailleurs supprimer les commentaires écrits dans des langues autres que l'anglais, qui semblent rares dans les sous-forums anglophones, mais qui pourraient interférer avec l'analyse du corpus.

Analyse du corpus

1. Outils.

De nombreux logiciels permettent d'analyser les corpus. On pourrait en tester plusieurs, pour voir lequel est le plus pratique ou le mieux adapté à ce projet. Il serait aussi préférable, pour analyser le corpus, d'utiliser un logiciel prenant en charge les caractères Unicode, afin de pouvoir étudier les emojis.

2. « Corpus-driven » ou « corpus-based ? »

Dans la mesure du possible, l'analyse du corpus sera « corpus-driven » plutôt que « corpus-based ». Il ne s'agira donc pas de réfuter, de valider ou d'affiner des hypothèses formulées en amont de l'analyse, en reprenant la définition des études « corpus-based » de McEnery et Hardie (2012, p. 6), mais d'adopter une approche plus naïve : le corpus sera la source des hypothèses, et on se laissera guider par les résultats pour tenter d'esquisser un portrait de la langue de Reddit et des changements diachroniques. Par ailleurs, comme le conseillent McEnery et Hardie, qui rejettent la distinction binaire entre les méthodes « corpus-based » et « corpus-driven », on combinera à l'occasion les deux méthodologies. Cela semble indispensable dans une certaine mesure, parce que certains termes pourraient sembler inaperçus ; l'acronyme *SO*, par exemple, pourrait être confondu avec la conjonction ou l'adverbe so.

On ne tentera pas d'étudier un maximum de phénomènes linguistiques. On se focalisera ainsi tout particulièrement sur les caractéristiques lexicales de l'anglais d'internet, telles qu'elles s'exprimeront dans le corpus. Une attention toute particulière sera ainsi portée aux éléments ci-dessous. Cette liste n'est pas exhaustive et sera complétée selon les résultats obtenus.

- **Les orthographes non-standard,** comme l'utilisation de lettres homophones, comme *ICU* pour « I see you », ou *U* pour « you », le remplacement de lettres par des chiffres, comme dans le cas de *4*, « for », ou l'utilisation du suffixe *–z* à la place de *–s* comme marqueur du pluriel.
- **Les abréviations**, comme *cuz* à la place de « because ».
- **Les acronymes et initialismes**, qui sont devenus, avec les émoticons, sans doute un des phénomènes les plus visibles de la langue du web. Parmi les incontournables, citons *LOL*, *OMG*, *WTF* ou *BTW*. Reddit semble un espace

particulièrement propice à l'utilisation d'acronymes, dont certains, comme *ELI5* ou *AMA*, sont devenus emblématiques du site.

- **La ponctuation** : elle peut être absente, ou, au contraire, être utilisée de façon expressive, avec une répétition de signes de ponctuation. On s'intéressera également à l'utilisation de symboles « non-standard », comme le caret ^, l'astérisque *, le slash / ou le hashtag #.
- **L'utilisation ou l'absence de capitales**.
- **L'allongement de certains mots**, comme dans *nooooooooo* ou *heyyyy*.
- **Les émoticons**.
- **Les émojis**. Ces caractères Unicode semblent en plein essor ; ils semblent devoir leur succès aux interfaces mobiles, qui permettent leurs insertions dans les messages.
- **La mise en forme des messages** : gras, italiques, petits ou gros caractères, caractères en exposant, etc.
- **Les néologismes**.
- **Les termes et les expressions de l'argot d'internet**, comme *yasss, This., lolz, epic fail,* etc.

Cette liste ne contient que des éléments lexicaux et graphiques, mais on ne peut pas exclure s'intéresser également à des phénomènes grammaticaux, même s'ils sont doute beaucoup moins fréquents.

3. Analyses quantitatives

L'étude des corpus ferait évidemment la part belle aux analyses quantitatives, c'est-à-dire statistiques. On s'intéressera aux mots dont la fréquence varie le plus d'un corpus à l'autre, et qui ont de préférence des fréquences élevées. On réfléchira aux mesures statistiques à utiliser pour rendre compte des différences entre les corpus.

On pourrait également imaginer comparer le corpus global, ou les trois corpus, avec un corpus plus représentatif de la langue du web dans son ensemble, comme le GloWbE ou le Corpus of Global Web-Based English, qui contient environ 1,9 milliards de mots issus de sources diverses : blogs, forums, magazines en ligne, sites internet statiques, et nouveaux médias. Ce corpus permet en outre de faire des comparaisons entre les variétés d'anglais utilisées dans vingt pays.

4. Analyses qualitatives

Baker met en garde contre les conclusions trop hâtives qui peuvent être tirées des résultats d'analyses quantitatives effectuées sur un corpus diachronique (2010, p. 80). Cela peut aisément conduire à la construction d'un « narrative » censé expliquer des changements, séduisant mais parfois erroné. Pour pallier à ce problème, il conseille d'utiliser des méthodes qualitatives, qui permettent d'apporter un éclairage supplémentaire sur les éventuels changements. Ces méthodes visent à replacer le corpus et les éléments étudiés dans leur contexte. Il recommande ainsi, d'une part, d'analyser les lignes de concordances, qui peuvent aider à comprendre pourquoi un élément est utilisé. Cette étape sera sans doute incontournable pour tenter de cerner le sens de nombreux acronymes et termes types du Netspeak, qui ne sont pas encore entrés dans les dictionnaires et qui ont souvent, selon les sources populaires disponibles sur le web, comme le site Urban Dictionary, des sens multiples. Le mémoire de Master a bien montré l'utilité de la lecture des lignes de concordance, dans le cas des acronymes qui ont plusieurs homonymes. D'autre part, Baker recommande de prendre en compte les contextes historiques et sociaux des périodes comparées. Dans le cas de Reddit, les différences de contexte peuvent sembler minces, mais elles sont tout de même bien réelles : il faudra ainsi se pencher sur l'histoire du site, sur l'actualité des époques étudiées, ou l'influence de nouveaux sites, plateformes ou innovations technologiques.

5. Analyse synchronique du corpus

Puisque l'on disposera d'un corpus de taille conséquente, il sera intéressant de procéder à une analyse synchronique, pour mettre en lumière plusieurs types de variation. Cela ne pourrait évidemment réalisée que si l'on arrive à obtenir un maximum d'informations démographiques sur les auteurs des messages : genre, pays d'origine, âge, expérience sur Reddit, notamment. Même en l'absence d'informations sur les Redditeurs, on pourra tout de même effectuer des comparaisons entre les différents subreddits, et identifier, par exemple, lesquels sont les plus novateurs sur le plan linguistique.

6. Résultats : Expliquer les différences

Evidemment, il est impossible de savoir à l'avance ce que cette étude révélera de la langue de Reddit. Si des différences significatives ont été notées, sur le plan

synchronique ou sur le plan diachronique, il sera nécessaire, dans la dernière partie de ce projet de thèse, de tenter de les expliquer. Il s'agit là, selon Baker, d'un des aspects les plus périlleux de l'analyse d'un corpus diachronique (2010, p. 69). Les éventuelles différences ou évolutions pourraient être dues, par exemple, aux innovations technologiques, comme l'utilisation des emojis sur les smartphones, ou à l'influence de « mèmes internet », ces phénomènes repris en masse sur Internet. La création ou la popularité croissante de certaines plateformes, comme Instagram ou Snapchat, pourrait également être source d'innovations linguistiques qui se transmettraient à Reddit. Enfin, on pourrait s'intéresser de plus près à l'histoire de Reddit, qui, de communauté relativement confidentielle et plutôt masculine, est en passe de devenir une plateforme « grand public », fréquentée par de plus en plus d'internautes aux profils divers.

Considérations éthiques

Recueillir des données sur Internet peut poser des problèmes éthiques. L'Association of Internet Researchers, dans un rapport daté de 2012, conseille aux chercheurs de bien réfléchir aux questions éthiques qui se posent quand on utilise des données tirées d'Internet (Markham & Buchanan). Elle explique qu'Internet complique la question de l'identité personnelle ; on peut ainsi se demander si une identité numérique est l'extension du soi, et si travailler avec du contenu issu de blogs, réseaux sociaux ou autres forums revient à travailler avec des sujets humains (p. 7). Dans certains cas, on peut considérer que le fait de récolter une grande quantité de données suffit à faire disparaitre dans la masse les différentes identités digitales, et les personnes qui les utilisent. Le rapport précise par ailleurs que, même dans le cas où les données sont anonymisées, on peut souvent tout de même identifier les individus. Il explique enfin qu'il existe tellement de cas de figure différents qu'il n'y a pas de règle, et qu'il faut toujours bien réfléchir avant d'entamer des recherches. La British Association for Applied Linguistics, ou BAAL, conseille quant à elle de prendre en compte le type de plateforme ou de site étudié (British Association for Applied Linguistics, 2006). Les internautes qui publient du contenu sur des sites accessibles à tous, comme Reddit, devraient normalement considérer que leurs contributions sont publiques. Leur autorisation n'est donc pas requise. De plus, les conditions d'utilisation de Reddit n'interdisent pas l'utilisation des données

du site à des fins de recherche (User Agreement, 2016). Maintenir la confidentialité des utilisateurs est en revanche nécessaire, même sur les sites publics. Pour ce faire, la BAAL recommande d'anonymiser les noms des internautes, qu'ils utilisent leur « vrai » nom ou un pseudonyme. Cela pourrait être fait très simplement, en attribuant un code à chaque Redditeur.

Conclusion

Cette ouverture sur la thèse a ébauché un projet qui ouvre plusieurs pistes pour l'étude de la langue de Reddit. La construction d'un corpus de qualité pourrait se prêter à plusieurs types d'analyses, qualitatives et quantitatives, et permettre une interprétation assez fine des résultats. Cette étude diachronique pourrait ainsi fournir de précieuses informations sur la façon sur la vitesse à laquelle la langue du web évolue, sur ce qui est éphémère et sur ce qui reste, et aider à mieux connaitre et comprendre les pratiques linguistiques d'une communauté virtuelle en pleine expansion. Toutefois, ce n'est évidemment qu'une esquisse, et il sera nécessaire de l'affiner et de la recadrer avant d'entamer le travail de thèse.

Bibliographie

Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh : Edinburgh University Press.

Bamman, B., Eisenstein, J., & Schoebelen, T. (2014). Gender identity and social variation in social media. *Journal of Sociolinguistics*, 18(2), 135-160.

Barbaresi, A. (2015). Collection, description, and visualization of the German Reddit corpus. Proceedings of the *2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media*.

Baron, N. S. (1998). Letters by phone or speech by other means: The linguistics of email. *Language and Communication, 18*, 133-170.

Baron, N. S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology, 23*(4), 394-423.

Barthel, M., Stocking, G., Holcomb, J. & Mitchell, A. (2016). *Nearly eight-in-ten Reddit users get news on the site.* Adresse : http://www.journalism.org/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/

Claridge, C. 2007. Constructing a corpus from the web: message boards. In Hundt, M., Nesselhauf, N., & Biewer, N., (Eds.), *Corpus linguistics and the web*, pp. 87-108.

Corpus of Global Web-Based English. Adresse : http://corpus.byu.edu/glowbe/

Crystal, D. (2001). *Language and the Internet* (2nd ed). Cambridge : Cambridge University Press.

Davis, B. H., & Brewer, J. P. (1997). *Electronic discourse: Linguistic individuals in virtual space*. New York : State University of New York Press.

Eckert, P. & McConnell-Ginet, S. (1992). Communities of practice: Where language, gender and power all live. *Locating Power: Proceedings of the Second Berkeley Women and Language conference*. Women Language Group.

Eckert, P. (2006). Communities of practice. In Mey, J. (Ed.), *Concise encyclopedia of pragmatics* (pp. 109-111). Amsterdam : Elsevier.

Enron Email Dataset. Adresse : https://www.cs.cmu.edu/~./enron/

Epoch Converter. (2016). Adresse : http://www.epochconverter.com/

Eyebrows on fleek. (2015). Adresse : http://knowyourmeme.com/memes/eyebrows-on-fleek

FAQ. (2015). Adresse : https://www.reddit.com/wiki/faq

Fox, A. B., Bukatko, D., Hallahan, M., & Crawford, M. (2007). The medium makes a difference gender similarities and differences in instant messaging. *Journal of Language and Social Psychology, 26*(4), 389-397.

Garrison, A., Remley, D., Thomas, P., & Wierszewski, E. (2011). Conventional faces: Emoticons in instant messaging discourse. *Computers and Composition*, *28*, 112-125. doi:10.1016/j.compcom.2011.04.001

Happy 10th birthday to us! Celebrating the best of 10 years of Reddit. (2015, June 23). Adresse : http://www.redditblog.com/2015/06/happy-10th-birthday-to-us-celebrating.html

Haralabopoulosa, G., Anagnostopoulosa, I., & Zeadallyb, S. (2015). Lifespan and propagation of information in on-line social networks: A case study based on Reddit. *Journal of Network and Computer Applications, 56*, 88-100.

Herring, S. C., Johnson, D. A., & DiBenedetto, T. (2011). Participation in electronic discourse in a "feminist" field. In J. Coates & P. Pichler (Eds.), *Language and gender: A reader*, (2nd ed.) (pp. 171-182). Oxford : Wiley-Blackwell.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In Barab, S. A., Kling, R., & Gray, J. H. (Eds.). *Designing for virtual communities in the service of learning* (pp. 338-376). New York : Cambridge University Press

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10(2*), p 00. doi **:** 10.1111/j.1083-6101.2005.tb00238.

Kirsten-Torrado, U. (2012). Development of SMS language from 2000 to 2010: A comparison of two corpora. *Lingvisticae Investigationes*, *35*(2), 218-236.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Lin, H., & Qiu, L. (2013). Two Sites, Two Voices: Linguistic Differences between Facebook Status Updates and Tweets. *Lecture notes in computer science, 8024*, 432-440.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Methods, theory and practice. Cambridge : Cambridge University Press.

McEnery, T., Xiao, R., Tono, Y. (2006). *Corpus-based language studies : An advanced resource book.* London : Routledge.

McEwan, B. (2016). Communication of communities: linguistic signals of online groups. *Information*, *Communication & Society, 19*(9), 1233-1249.

Markham, A., & Buchanan, E. (2012). *Ethical decision-making and Internet research*. Adresse : http://aoir.org/reports/ethics2.pdf

Nesselhauf, N. (2007). Diachronic analysis with the internet? Will and shall in ARCHER and in a corpus of e-texts from the web. In Hundt, M., Nesselhauf, N., & Biewer, N., (Eds.), *Corpus linguistics and the web*, pp. 287-305. Amsterdam : Rodopi.

Noguti, V. (2016). Post language and user engagement in online content communities. *European Journal of Marketing*, *5*(5/6).

Olson, R., & King, R. (2015). How the Internet talks. Adresse : http://projects.fivethirtyeight.com/reddit-ngram/?&start=20071015&end=20150831&smoothing=10

Pavalanathan, U. & Eisenstein, J. (2016). Emoticons vs. emojis on Twitter: A causal inference approach. *AAAI Spring Symposium 2016*. Adresse : http://arxiv.org/abs/1510.08480

Park, J., Barash, V., Fink, C., & Cha, M. (2013). Emoticon Style: Interpreting Differences in Emoticons Across Cultures. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Rezabeck, L. L., & Cochenour, J. J. (1994). Emoticons: Visual clues for computer-mediated communication. In *Imagery and visual literacy: Selected readings from the Annual conference of the International Visual Literacy Association*, 371-383.

Rosenthal,S., & McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations.

Proceedings of the *49th Annual Meeting of the Association for Computational Linguistics*, pages 763–772, Portland, Oregon, June 19-24, 2011.

Sanchez-Moya, A., & Cruz-Moya, O. (2015). "Hey there! I am using WhatsApp": A Preliminary Study of Recurrent Discursive Realisations in a Corpus of WhatsApp Statuses. Procedia – Social and Behavioral Sciences, 212, 52-60.

Site overview. Reddit.com. Consulté le 20/06/2016 à l'adresse : http://www..com/siteinfo/reddit.com

Skovholt, K., Gronning, A., & Kankaanranta, A. (2014). The Communicative Functions of Emoticons in Workplace E-Mails: :-)∗. *Journal of Computer-Mediated Communication*, *19*, 780-797.

Tagliamonte, S. A, & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech, 83*(1), 3-34. doi: 10.1215/00031283-2008-001

The British Association for Applied Linguistics. (2006). *Recommendations on good practice in applied linguistics.* Adresse : http://www.baal.org.uk/dox/goodpractice_full.pdf

Tossell, C. C., Kortum, P., Shepard, C., Barg-Walkow, L. H, Rahmati, A., & Zhong, L. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior, 28*, 659-663. doi:10.1016/j.chb.2011.11.012

Use Cloudsearch to search for posts on reddit within a time frame. (2015). Adresse : https://www.reddit.com/r/reddittips/comments/2ix73n/use_cloudsearch_to_search_for_posts_on_reddit/

User agreement (2016) Adresse : https://www.reddit.com/help/useragreement/

Varnhagen C. K., McFall, G. P., Pugh, N., Routledge, L. Sumida-MacDonald, H., Kwong, T. E. (2010). lol: new language and spelling in instant messaging. *Read Writ*, *23,* 719-733. doi: 10.1007/s11145-009-9181-y

WebCorp. Adresse : http://www.webcorp.org.uk/live/

Werry, C. (1996) Linguistic and interactional features of Internet Relay Chat. In Sampson, G., & McCarthy, D. (Eds*.) Corpus linguistics :Reading in a widening discipline*, 2005, pp. 340-352.

Wilson, S. M., & Peterson, L. C. (2002). The anthropology of online communities. *Annual Review of Anthropology, 31*, 449-467.

Witmer, D. & Katzman, S. L. (1997). On-line smiles: Does gender make a difference in the use of graphic Accents*? Journal of Computer-Mediated Communication, 2*(4). doi: 10.1111/j.1083-6101.1997.tb00192.x

Wolf, A. (2000). Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior*, 3(5), 827-833.