GENDER AND LANGUAGE IN COMPUTER-MEDIATED DISCOURSE:
A HISTORICAL ANALYSIS OF USENET NEWSGROUPS

Elli E. Bourlai

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics, Computing, and Engineering
Indiana University
December, 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____
Susan C. Herring, Ph.D.

_____
Pnina Fichman, Ph.D.

_____
Allen Riddell, Ph.D.

_____
Markus Dickinson, Ph.D.

November 30th, 2018

Acknowledgments

This dissertation is the result of a long journey that started in August 2012, and there have been many people along the way who helped me reach my destination.

I would like to thank first my advisor and Chair of my committee, Professor Susan Herring, whose work introduced me to computer-mediated communication as a research area and set a starting point to this journey. She has been a wonderful mentor, providing valuable advice and support whenever I needed it, and helping me grow as a researcher and a scholar; I am truly honored to be one of her students.

I would also like to express my sincere gratitude to my committee members: Professor Pnina Fichman, for all the great advice, suggestions, and help she has offered throughout my doctoral program, but also for being a friendly voice when I struggled with work and life balance; Professor Markus Dickinson for making my first programming experience so positive and for helping me grow all the technical skills needed to achieve my goals; and Professor Allen Riddell, for his willingness to participate in this journey and for offering advice that helped me overcome the statistical challenges of this study.

The completion of this dissertation would not have been possible without the support of the Rob Kling Center for Social Informatics. I am so very grateful for the Rob Kling Social Informatics Fellowship which funded my equipment and research staff expenses. Moreover, I would like to thank Megaputer Intelligence for kindly providing their software for the annotation and analysis of my data.

Special thanks to Stephanie Dickinson, who graciously offered her time and valuable advice on the appropriate statistical models to use for the purposes of my study, and to Lilian Golzarri Arroyo, who was an incredible help in walking me through every step of the statistical models and R code.

In addition, I would like to thank other faculty members and colleagues in the Department of Information and Library Science, as well as in the Department of (Computational) Linguistics: Professor Ying Ding, who provided guidance during the first years of my doctoral program and helped with my first graduate course teaching experience; Professor Howard Rosenbaum, who read and provided feedback on my Qualifying paper and always offered valuable advice for my program, funding, and professional development; Professor Noriko Hara, who served as a reader for my Qualifying paper; Professor Sandra Kuebler, who helped me become more confident in my programming skills and understanding of Computational Linguistics; Professor Damir Cavar, who allowed me to participate in one of his courses where I learned important technical skills; and past and present doctoral students Madelyn Sanfilippo, Guo Zhang Freeman, Chun Guo, Muhammad Abdul-Mageed, Zheng Gao, Brad Demarest, Ashley Dainas, JT Wolohan, Ali Ghazinejad, Tim Bowman, Lois Scheidt, Grant Simpson, Ewa Zegler-Poleska, Satoshi Tsutsui, Chenwei Zhang, Wei-Chu Chen, and Andreas Bueckle who offered their friendship, collaboration, support, and advice.

To the wonderful Corey Tarbell, Katie Novak, Jessica Meyer, and Rhonda Spencer: thank you for helping me navigate the bureaucratic labyrinth of being an international doctoral student studying and working in the United States. Moreover, a big thank you to Mike Gallant for helping and offering advice for all the IT issues related to my dissertation, and to my colleagues at Megaputer Intelligence, especially Rebecca Hale and Margaret Glide, for being so supportive during the last months of my program.

Last but certainly not least, Miguel, Lilian, and Alex (a.k.a. The Γαμάτο Mesa): thank you for keeping me sane and being my family away from home. To my incredible parents and sister: thank you for encouraging me to follow my dreams and for offering so much unconditional love and support every single day of this journey, even if there was an ocean

between us. And, finally, to my wonderful husband who met me at the craziest part of this journey: thank you for loving me, for supporting me in so many ways, for putting up with my stress and my work schedule, and for picking me up every time I fell with encouraging words and emergency ice cream.

Elli E. Bourlai

GENDER AND LANGUAGE IN COMPUTER-MEDIATED DISCOURSE: A HISTORICAL ANALYSIS OF USENET NEWSGROUPS

This study focuses on the expression of gender identity through language from a historical perspective. Specifically, it explores the linguistic styles of women and men in computer-mediated discourse (CMD) diachronically, in order to identify possible changes in gender patterns over time, as well as the manner in which they evolved. An automated Computer-Mediated Discourse Analysis (CMDA) was conducted to analyze the words in approximately four million posts from 47 newsgroups representing all major USENET hierarchies and covering a 25-year time period.

All the linguistic features that were examined exhibited change over time in their overall use, and the change was statistically significant for 63 out of the 72 features modeled. Similarly, there was change in the patterns of the female and male frequencies for all the features over time, and it was statistically significant in 24 out of the 72 features. Using a language contact metaphor, the concepts of convergence and divergence were used to evaluate the manner of change by examining the trend lines of the features. Even though the language used in USENET newsgroups became less gender-differentiated over time, it shifted toward a more masculine linguistic style overall, with cases of "hypermasculinization" by women of linguistic features associated with status and power. Since the majority of newsgroups in the dataset were male-predominant, the newsgroup net.kids/misc.kids is presented as a case study of the language used in a female-predominant newsgroup.

Theoretically, this study continues the tradition of studying gender and its nature, as well as examining the role of technology in gender expression and gender (in)equality; it also

provides new insights by approaching the issue diachronically. The study's main contributions are as a first step toward identifying and understanding the causes of changes in the discourse of men and women online. Consequently, this study helps fill gaps in the literature not only in the field of computer-mediated communication research, but also in the fields of gender studies and historical linguistics.

Methodologically, the study provides useful insights for conducting diachronic CMC research, an area that is severely understudied due to challenges in the collection and analysis of historical data. It also serves as an example of applying an approach traditionally used in historical linguistics to study offline communication to CMD, since there is currently no methodological framework for analyzing CMD diachronically. Finally, this study contributes through the creation of the *Historical USENET Newsgroups Corpus* (Bourlai & Gao, 2017), a large-scale diachronic corpus of CMD that will become publicly available for use upon its completion.

_____

Susan C. Herring, Ph.D.

_____

Pnina Fichman, Ph.D.

_____

Allen Riddell, Ph.D.

_____

Markus Dickinson, Ph.D.

Table of Contents

List of Tables

List of Figures

Introduction

1.1. Background

Gender is one of the most important aspects of human social identity; as such, it has been the focus of much research in a variety of fields, including anthropology, linguistics, psychology, and sociology. Based on long-standing gender assumptions and stereotypes, men and women have systematically been treated differently, especially in terms of social status and power. This gender inequality is evident in the use of language, both in same-sex interactions and mixed-sex interactions (Coates, 1993; Tannen, 1994).

The arrival of computers and, specifically, the Internet created an early optimistic belief that, due to the anonymity and the reduction of social cues when using this technology, communication would be more equal and democratic (e.g., Hauben & Hauben, 1997). *Computer-mediated communication* (CMC), the communication that takes place among individuals via networked computers or mobile devices (Herring, 1996a), was also seen as the ideal environment for individuals to experiment with and explore their identity (Bruckman, 1993), removed from physical constraints imposed by their gender, age, social class, or ethnicity (Danet, 1998; cf. Herring, 2003a; Yates, 1997). This technologically deterministic view predicted more equal and democratic interactions, and that CMC would be especially beneficial in leveling the linguistic inequality women faced (Graddol & Swann, 1989). Such a "technological utopianism" vision, however, is based on the premise that a technology works in an ideal way, and it ignores the requisite social conditions for a technology to be effective as an agent of social change (Kling, 1994).

It was not long before scholars started analyzing online communication, only to discover that the early predictions were far from becoming a reality (Gregory, 1997). While a few early studies (Bruckman, 1993; Danet, 1998; Rodino, 1997) provided evidence that

people (especially females) interacting online made use of the anonymity and the absence of physical cues in CMC to address social inequality issues or experiment with their gender identity, most of the research on gendered interaction has found that traditional gendered linguistic norms (Tannen, 1993, 1994, 2003) have transferred from offline to online communication (Gregory, 1997; Hall, 1996; Herring, 1992a, 1992b, 1993, 1994, 1996b, 1996c, 2000, 2003a).

In addition to stylistic differences found in the language of males and females that resemble those in face-to-face communication, power asymmetries have also persisted: Males are still the dominant gender in most online environments (with the exception of a few female-predominant mailing lists, discussion groups, and, more recently, social media sites such as Pinterest, as Herring and Stoerger [2014] note), and they exercise power over females in mixed-sex interaction by dominating discussions, as well as by imposing their interaction values as the majority group on the general internet population (Gregory, 1997; Hall, 1996; Herring, 1996b). This dominant attitude can be seen in behaviors ranging from dominating the floor in online discussions (Herring, 1993, 2010; Hert, 1997) to flaming and trolling (Gregory, 1997; Herring, 2003a) to more extreme situations of sexual harassment and death threats against women (Guzzetti, 2008), as seen recently in the phenomenon of Gamergate.[1]

It is apparent that gender is as an important aspect of social identity in online communication, as it is in offline communication. Gender information is the most shared profile field publicly (Las Casas et al., 2014), and the popular "a/s/l" (age/sex/location) acronym has been one of the first lines to appear in interactions among unknown users in public chat platforms (Guiller & Durndell, 2007). Gender influences and shapes interaction in online environments in terms of participation, politeness, sexuality, and identity expression, which in turn affects users' offline or 'real' life personally, socially, and professionally, as the

[1] https://en.wikipedia.org/wiki/Gamergate_controversy

role of CMC in everyday life is becoming increasingly more prominent. Thus, the study of gender in *computer-mediated discourse (CMD)*, the language and language use in CMC environments (Herring, 2001), provides useful insights into and a better understanding of the intricate workings of gender communication and expression in CMC and in society more generally.

1.2. Problem Statement

The use of corpora has played an important role in studies of gender in CMD, as they provide empirical evidence from naturally-occurring data collected from a variety of textual CMC modes (Androutsopoulos & Beißwenger, 2008; Herring, 2003b). While previous corpus-based research on gender in CMC has provided important insights into the linguistic preferences of internet users in constructing their online gender identity, as well as the interaction dynamics among users, the extent to which CMC affects gendered language (including both its social and technological affordances) on gendered language use is still a matter of debate (D'Urso, 2009; Kapidzic & Herring, 2011). Scholars have approached the issue almost exclusively from a synchronic point of view, studying a variety of modes and user groups at different points in time during the past 25 years, and providing evidence for possible emergent changes in the linguistic styles of different genders.

This approach to evaluating the extent to which CMC affects gendered language can be problematic for two reasons. First, differences in the discourse styles of men and women that deviate from traditional norms may be temporary trends rather than linguistic changes in progress. Variable linguistic phenomena are often trends that become more or less popular until they either are standardized or become obsolete with the passing of time (Baker, 2010; Nevalainen, 2011). Second, the technological and social contexts of data sources play an important role in the analysis of CMD data (Herring, 2007): If the sources differ in those

contexts in any degree, comparisons may not be valid. Language change has traditionally been studied with diachronic corpora in historical sociolinguistics, using systematic sampling to control for certain variables (Baker, 2010). Consequently, the effect of CMC on gendered language may be better evaluated through a diachronic study using a homogeneous technological and social context for the entire period studied.

CMC has existed for almost half a century, and CMD provides an ideal source of data for diachronic studies of language change. A number of archives have been preserved that cover a continuous time period since the era of *ARPANET* and *USENET,* in contrast to the sporadic data that are usually available for offline historical studies (Androutsopoulos, 2011; Herring, 2003b). Moreover, gender has been found to play a role in the diffusion of linguistic innovations (Cameron, 2003; Labov, 1963, 1966). In spite of these facts, there are very few historical CMC studies, especially from a linguistic viewpoint. While several publicly available synchronic CMD corpora exist, there is a glaring lack of diachronic CMD corpora. Even though archives of older CMD data exist, they are usually difficult to access and retrieve. This may have discouraged researchers from conducting historical research on CMD, thus resulting in the gaps in the literature. Currently, the oldest available CMD data are academic mailing lists and USENET Newsgroups. Most academic mailing lists are not accessible publicly, but their archives may be acquired upon request. However, user content in the messages of academic mailing lists may be edited, which could affect findings in studies of naturally occurring language. This makes USENET Newsgroups a more suitable choice as a data source for a diachronic CMC corpus, since they are not moderated in their majority, and they span almost four decades of existence.

Given the issues described in this section, this study attempts to fill in research gaps in the areas of gender and language change in CMD by analyzing the linguistic styles of males and females diachronically using data from a newly constructed historical

asynchronous CMD corpus, the Historical Usenet Newsgroups Corpus (HUNC), presented in Section 2.4.4.

1.3. Definition of Terms

Gender is a collection of discourses and practices constructed by society "around the biological differences of sex" (Yates, 1997, p. 284). Nonetheless, gender should not be confused with biological sex: We are not born with it, but rather it is socially constructed as we learn behaviors and attitudes that are considered appropriate to our sex (Graddol & Swann, 1989). Moreover, gender is a continuous variable: A person may have varying degrees of "masculinity" and "femininity" (Graddol & Swann, 1989). With regard to language, the term *gendered language* is used to refer to gender preferences in language use, *masculine* or *feminine linguistic styles* to refer to a set of linguistic characteristics traditionally associated with preferential use by males or females, and *gender markers* to refer to specific linguistic features that are "marked" as belonging to a masculine or feminine linguistic style (Coates, 1993). Previous literature has used the term "the (sociolinguistic) gender pattern" to refer specifically to the difference in use of standard or prestige language variants (Cheshire & Gardner-Chloros, 1998); however, according to ELLO (2018), a more general use of the term is "a typical sociolinguistic pattern, a characteristic type of sex-graded linguistic variation" (n.p.). In this study, a *gender pattern* is defined as the variation in the use frequency of a linguistic feature between males and females.

Language change or **linguistic change** is the variation in linguistic features that occurs over time and may involve different aspects of the language (syntax, morphology, phonology, lexicon, etc.). In this study, the term refers to diachronically measured language change (between time points), as opposed to emergent change evidenced through synchronic variation. According to Bybee (2015), "[c]hange occurs when new patterns arise, when

patterns change their distribution, or when they are lost" (p. 10): This study explores possible changes in the distribution of gender discourse patterns in CMD. Language change may occur through *language contact*, when speakers of two or more languages or language varieties interact and influence each other, which may lead to convergence or divergence of the languages or language varieties. Branmüller and House (2009) define *convergence* as "the various ways in which language or language varieties become more similar to each other in all grammatical subsystems," which can be both a process and product, as well as conscious or unconscious (p. 4). While convergence includes linguistic unification and homogenization, *divergence* leads to diversification and heterogenization: "[L]anguages or language varieties become more distinct from one another" (Branmüller & House, 2009, p. 4). In this study, the term *convergence* is used to refer to the frequency of use of a linguistic feature becoming more similar between females and males over time, and the term *divergence* to refer to the frequency of use of a linguistic feature becoming more different between the genders over time. Additionally, the term *reversal of a gender pattern* refers to when the frequency of use of a linguistic feature becomes more similar over time (convergence), but then starts to differentiate again (divergence), resulting in "flipped" use frequencies between the genders compared to the beginning of the change.

A **corpus** is a collection of machine readable, authentic data (usually in textual form), sampled in such a way that it is representative of a particular language or language variety (Baker, 2010).[2] In this study, the term is used in reference to a collection of naturally-occurring textual data in CMD, and the term *corpus-based studies* refers to studies that employ a systematic (usually quantitative) analysis of such corpora, as opposed to purely interpretive/theoretical approaches and studies that use only surveys/questionnaires. Corpora may be either *synchronic*, comprised of current data from a single time point and used for

[2] The latter part mostly refers to the more traditional *general* or *reference* corpora (such as the BNC or COCA), as opposed to the *project-based* or *specialized* corpora that are more popular in CMD, due to the sampling limitations imposed by the environment (Beißwenger & Storrer, 2008).

exploring language variation, or *diachronic*, comprised of data gathered from different time periods and used for exploring language change (Baker, 2010). Diachronic corpora are further subdivided into *Dynamic Corpora*, corpora that are constantly updated with new data, and *Sample (Static) Corpora*, corpora that are not further updated with new data after their initial collection (Kennedy, 1998). This research project is a corpus-based study that uses a Sample Diachronic Corpus.

1.4. Organization of the Dissertation

This dissertation is organized as follows. In chapter 2, previous literature on gender and CMD and on language change and CMD are presented, along with several issues in the construction of CMD corpora that affect their availability and are related to gaps in the literature. Chapter 3 presents the theoretical framework used in the study with regard to gender and language change. After formulating the hypotheses and the research question (4.1), the methodology chapter first provides background information about the source of the data, as well as the corpus that is used (4.2); then, the selected data from the corpus are described in detail (4.3). The analytical and statistical methods are described in sections 4.4 and 4.5, respectively. The results of the analysis are presented in chapter 5: the descriptive statistics for the overall change in variables over time (5.1) and the change in the gender patterns of variables over time (5.2), the results of the linear regression analysis (5.3), the evaluation of gender pattern trends (5.4), and the findings of the case study of the female-predominant newsgroup net.kids/misc.kids. Chapter 6 discusses the results with regard to the overall change in discourse patterns in USENET newsgroups (6.1) and in relation to traditional gender patterns in previous literature (6.2), and attempts to interpret the changes and trends identified in the gender patterns (6.3). Finally, the concluding chapter (7)

discusses the implications of the study for different research fields (7.1), addresses the limitations of the study (7.2), and suggests directions for future research (7.3).

Literature Review

2.1. Gender and CMD

The study of gender in CMC dates back to the late 1980s and is part of the "first wave" of CMC studies (Herring, 2003a). Intrigued by the notion of a democratic communication medium where anyone could take on any identity they desired in the absence of physical cues due to anonymity (Danet, 1993; Graddol & Swann, 1989), scholars sought to investigate whether such claims were true: Was the communication of men and women different in this new environment than in face-to-face (F2F) interaction, as liberal/postmodern cyberfeminists proposed (Bucholtz, 2001; cf. Hall, 1996)? In need of more substantial evidence than anecdotal observations, researchers decided to investigate the absence or existence of traditional gender patterns in offline communication. The underlying purpose of most studies was to explore gender in this new communication medium, the Internet, or to identify gender inequalities for theoretical or policy-making reasons.

Recently, however, much research on gender in CMD comes from the fields of Computer Science and Computational Linguistics (Natural Language Processing), which looks at gender characteristics for author identification and opinion/sentiment mining. Such interest has partly arisen because of the growth of social media and the abundance of user-generated content they have created, which calls for automated processes to analyze latent demographic characteristics of users (Burger et al, 2011). The findings of corpus-based research on gender and CMD are organized in synchronous and asynchronous studies following the organization in Herring (2003a); a separate section for social media has been added, since social media platforms typically include both modes.

2.1.1. Asynchronous CMD

Asynchronous CMD occurs in online environments that allow a user to read messages at a later time, without having to be logged in at the same time as the other user(s) (Herring, 1996a). Examples of such environments include academic and non-academic mailing lists, discussion newsgroups, forums (Herring, 2003a), and blogs. Much research on CMD uses data from asynchronous environments, due to the fact that they are easier to acquire: The messages persist (Herring, 2007) and there are a lot of publicly accessible data in the archives of mailing lists and discussion newsgroups (Baumann, 2015).

One of the first studies to look at gender differences in asynchronous CMD was Selfe and Meyer (1991), who examined 296 messages from 18 male and 15 female participants in the MegaByte University List (MBU-L) discussion group under two different conditions: 107 messages were sent during a 20-day period when the names of the participants were known, and 189 messages were sent in another 20-day period when participants used pseudonyms. The authors' purpose was to test the claim that anonymity in CMD could make exchanges between the genders more democratic. Even though women stated that they felt more comfortable and liberated using pseudonyms, in contrast to men, who felt more uncomfortable, the authors did not find any differences in the amount of participation or the stylistic features used by either males or females in the two conditions; rather, anonymity seemed to affect only the topics discussed. In both conditions, men contributed twice as many messages with 40% more words, initiated three times as many topics, and disagreed twice as often as women. The authors concluded that there was a persistent power and status imbalance on MBU-L.

Herring's early studies (1992a, 1992b) of messages on the LINGUIST list reported similar findings. In her analysis (1992a) of 72 messages centered around the term "cognitive linguistics" and the responses to a related survey, she found that women were overshadowed by men's participation, both in terms of number of contributors and contributions, and in

length of messages. She interpreted these findings based on her survey results: Women produce less and participate less in adversarial discourse based on their communicative preferences; thus, environments where a male rhetoric of adversarial exchanges is common may inhibit women from participating. In her 1992b study, Herring analyzed all messages posted to the LINGUIST list for a random two-week period and two extended threads. In that corpus, men also participated more by posting more messages that were twice as long as those of women; moreover, men mostly discussed issues and were more likely to engage in debates, in contrast with women who were more personal in their messages. Another interesting finding emerged when she analyzed the frequency of members' use of "male styles" and "female styles" based on a set of features associated with masculine and feminine discourse styles: She found that women were three times more likely than men to use a "mixed style." She attributed this finding to the fact that LINGUIST-L is an academic list, and "male" stylistic features have been traditionally associated with academic professionalism.

Herring then turned her attention to politeness (1994, 1996d). In an analysis of data from nine non-academic and academic discussion lists, she coded posts for violations of negative and positive politeness, as well as observances of negative and positive politeness (cf. Brown & Levinson, 1987). As hypothesized, women made more use of observances of positive politeness and negative politeness in their effort to be considerate of others, as well as to ratify others and be liked. In contrast, men regularly violated both positive and negative politeness, and were more likely than women to resort to flaming. In order to better understand the results of her analysis, Herring compiled a questionnaire to assess whether men and women evaluated politeness differently. Indeed, according to the responses of the participants, it seems that not only do men and women understand politeness differently, but they seem to be adhering to different politeness ethics: an ethic of politeness-based

communication (women) versus an ethic of agonistic debate and freedom from rules or imposition (men). The latter ethic emerged with hacker culture, and will likely be the dominant ethic as long as males dominate the internet (Herring, 1996d).

Herring (1996b) also tested the popular stereotype that women are not as information-oriented as men but are more socially-oriented, by analyzing the schematic organization of 33 male and 31 female messages collected from two extended discussions from the LINGUIST and WMST lists. Two different styles emerged from her analysis: the *aligned* versus the *opposed* message style. Female messages indeed contained more interactional features but disproved the stereotype, in that their messages were actually more informational than those of men. However, male messages expressed more critical views. The most interesting finding, however, was what she calls the "list effect": The minority group in each list adjusted their style to that of the majority group of the list. Consequently, males in the WMST list tended to shift toward a more feminine style, whereas females in the LINGUIST list shifted their style to be more masculine.

Herring, Johnson, and DiBenedetto (1992, 1995) analyzed a lively discussion from the MBU-L where women's participation was temporarily slightly higher than men's, invoking a strong reaction from some of the men on the mailing list. By looking at the participation rates and the number of responses each post received, the authors (1992) found that men dominated the conversation except for a three-day period when women's contributions exceeded those of men on a topic introduced by women; women also were responded to more on that topic. Except for that three-day period, both men and women mostly responded to men's posts; even women responded less to other women. Another revealing finding was that hedges, a feature associated with powerlessness and generally preferred by women, were mostly used by males in the three-day period when women participated more, indicating that they felt a power shift. The authors (1995) also identified

several mechanisms on the part of men for silencing women: (1) *avoidance* through lack of responses to female messages, diversion from their original messages, or dismissal of their messages as trivial and unimportant via patronization and humor, (2) *confrontation*, and (3) *co-optation* by reformulating women's ideas as their own and thereby (re)gaining control of the conversation. Women, however, reacted to these silencing efforts by participating more and persisting, which the authors suggested is a good strategy to redistribute power in CMD.

The findings in Savicki, Lingenfeltr, and Kelley (1996) seem to contradict certain patterns found in the previous studies, however. Even though their corpus also exhibited male dominance in both membership and participation in the discussion groups studied, with more fact-oriented language and calls for action in groups with higher male membership, there was a significant lack of challenges, argumentative language, coarse and abusive words, or status indicators in their data. Similarly, in groups with high female membership, the authors confirmed that women used more self-disclosure and attempted to prevent and reduce tension; however, there were not many apologies, questions, or first person pronouns used. Moreover, users in general participated less in groups where the proportion of males was low. The authors attribute these differences to the gender composition and the context of the groups they studied.

Bucholtz's findings (2001) also presented differences compared to previous literature when she applied the same analysis as Herring (1996b) to 739 messages from two discussion threads on the community blog Slashdot, a community where "geeks" discuss technology news. Even though there were fewer women participants, they produced approximately the same number and even lengthier messages than men, and gave/received an equal number of responses to and from both genders. Moreover, there were no differences in terms of the aligned or opposed message styles identified in Herring (1996b). The author suggests that

there are different versions of female gender identity, since the female users of Slashdot also identify as "geeks."

Later studies of gender in CMD turned to blogs, after blogs became a popular environment for men and women to express their identity. Huffaker and Calvert (2005) analyzed 35 male and 35 female blogs of teenagers and, while some of their results adhere to traditional gendered language, parts of their findings contrast with previous research. They found some gender differences consistent with previous literature, but they also found similarities in gender styles. Both boys and girls discussed personal issues in their blogs and revealed personal information. Moreover, there was no overt aggression in male blogs or overt passivity in female blogs, even though males used more active and assertive language. There was also no difference regarding accommodating and cooperative language, and, surprisingly, males used more emoticons in their blogs than females, especially flirty and sad emoticons.

This latter finding is inconsistent with previous research on the use of emoticons by gender: Witmer and Katzman's study (1997) found that the use of graphical accents by females was significantly higher than that by men, who flamed and challenged others more, in accordance with previously-identified patterns. Wolf (2000) also analyzed the use of emoticons in 251 posts from four different USENET newsgroups: one female-predominant group, one male-predominant group, and two gender-balanced groups. She found that emoticon use was higher in the female-predominant group than in the male-dominant group; however, instead of females adopting the linguistic style of males in the balanced groups, she found that males made higher use of emoticons in those groups, adjusting their style to that of females. Baron's study (2004) of IM messages described in the next section also found higher use of emoticons by females.

Huffaker and Calvert's findings may indicate a genre bias in the corpus of the authors: The blogs they analyzed are diary blogs, which have been found to be used more by females and to use more of a personal discourse style, as suggested by Herring and Paollilo (2006). In their study, Herring and Paollilo analyzed 65 female entries and 62 male entries from 44 different weblogs, and they controlled their multivariate analysis for the variables of gender and genre (diary blogs or filter blogs). Based on a set of features comprising male and female discourse styles as suggested by Argamon et al. (2003), the authors found that genre is a stronger predictor of "gendered" stylistic features than author gender.

Relatedly, Thomson (2006) tested the effect of topic on gendered language in 60 threads from public discussion lists and group discussions by university students. Using an extensive set of features including male and female gender markers, his multivariate analysis showed that gendered language is strongly dependent on the context or topic of discussion. Perhaps the most comprehensive work on bloggers' expression of gender and age identity through language is that of Argamon et al. (2007): The authors' analysis of 21,682 female and 25,065 male blogs indicates that males talked more about religion, politics, business, and the internet, and use more articles and prepositions. Females made higher use of words in the categories of conversation, home, fun, romance, and swearing; they also used more personal pronouns, conjunctions, and auxiliary verbs. Similarly to Herring and Paollilo (2006), Argamon et al. emphasized the importance of content and blog genre in gender style preferences.

Another study explored masculine and feminine linguistic styles in student posts (Guiller & Durndell, 2007). Guiller and Durndell used 21 task codes, 12 linguistic codes, 16 stylistic codes, and eight paralinguistic codes (e.g., capitalization) to analyze 699 posts from 48 males and 148 female students in the Netherlands. Even though they found no significant differences in individual features, with the exception of more intensifiers used by females,

they found different patterns when clustering features into "supercodes": Males were more likely to use authoritative language and respond negatively in interactions, whereas females were more likely to explicitly agree with and support others, as well as to be more personal and emotional.

Exclamation marks, another gender marker traditionally associated with females' discourse, were analyzed by Waseleski (2006) in public discussion lists. Although exclamation marks were not used extensively by either gender (only 343 exclamation phrases/sentences in 3,000 messages) and did not usually express excitability, females made greater use of exclamation marks in order to thank others and appear friendly. This may also be the reason why females tend to use more emoticons, as suggested by Witmer and Katzman (1997) and Baron (2004).

De Oliveira (2007) examined gender preferences in politeness as expressed by Portuguese address forms in a study of university bulletin boards. While traditional patterns of male dominance in terms of participation and authority found in previous studies were confirmed, the women in her corpus were not as concerned with politeness; rather, it was actually the men who took on the role of "politeness adjudicators."

Herring (2010) returned to studying the issue of power and control in online conversations by testing the applicability of Edelsky's *floor types* (1981) in 313 messages from the LINGUIST, MBU, and WMST lists. She found systematic differences between samples with higher male presence and samples with higher female presence: Male-predominant discussions exhibited the features of the F1 floor as in Edelsky's findings, whereas female-predominant discussion had a mixed floor type that combined features of F1 and F2.[3]

[3] According to Herring (2010), F1 floors in CMC have long duration, longer messages, are interactionally "sparse," have a single thematic focus, are usually contentious, and are hierarchically dominated by a minority of participants. F2 floors have shorter duration, shorter messages, are interactionally "dense," have multiple thematic foci, are usually supportive and collaborative, and are more egalitarian in terms of participation.

Finally, a more recent study (Laniado et al., 2012) that analyzed the comments and profiles of 2,613 male and 165 female Wikipedia users found style differences that agree with most of the previous research. Female editors tend to be more assertive but not aggressive, in that there is higher positive valence in their messages with a more affectionate and welcoming tone, which indicates an aversion to criticism and conflict; in contrast, male editors tend to maintain a more neutral tone. Women also seem to prefer participating in topics that have more positive valence and use more links to Wikipedia policies, possibly because they are sometimes addressed in a paternalistic or condescending tone. Regarding their profiles, female editors express much more emotion than male editors and also receive more positive messages left on their pages, expressing less dominance than the messages left for female users without administrative power. These findings suggest that female administrators on Wikipedia tend to assume a traditional female style of cooperativeness and aversion to conflict, but that because of their administrative role, they also enjoy higher status privileges that are usually associated with men (cf. Herring, 1992a).

2.1.2. Synchronous CMD

Synchronous CMD refers to discourse in environments where users have to be logged in at the same time in order to communicate (Herring, 1996a), including public chat rooms on platforms such as Internet Relay Chat (IRC), virtual text-based environments such as Multi-User Dungeons (MUDs) and MUDs Object-Oriented (MOOs), Multi-Modal Online Role Playing Games (MMORPGs), and private messaging applications mostly known as Instant Messengers (IM). One important difference between these environments and asynchronous environments is their ephemeral nature (Herring, 2007): Log chat files are usually not archived by the platforms or the users, except for IM logs that can be automatically archived by the messenger applications. Thus, CMD scholars need to either be virtually present in

these environments in order to collect data by saving the chat logs, or they need to request logs archived by the users. This may explain why there are fewer studies of synchronous CMD.

Encouraged by the ephemeral status of the environments, the majority of synchronous CMD occurs under the guise of pseudonyms rather than the real (offline) identity of the users, which tends to be more salient in asynchronous communication. As such, it seemed like an ideal context for the postmodernist cyberfeminist approach that focused on the democratic and liberating potentials of technology on gender. An early study embracing this view was Danet's (1998) analysis of 260 IRC nicknames from four IRC channels. She assessed the degree to which the nicknames conveyed gender information and found that only one-fifth of the nicknames in her data suggested the gender identity of the users. Based on her analysis of unconventional gender (and accompanying 3rd person pronoun) choices in MediaMOO and LambdaMOO, Danet argued that people use the anonymity of CMD to escape the social constraints of gender. Rodino's (1997) study of 414 lines of chat text from IRC makes similar claims: Despite the fact that some participants adhered to binary gender behaviors, a number of individuals enacted multiple gender performances, suggesting that CMD facilitates breaking out of gender binaries.

Herring and Panyametheekul (2003) also found some results that showed different interaction patterns than those found in previous research on gender in asynchronous CMD (cf. Herring, 1992a, 1992b). In their study of 917 messages from the Thai chat room #jaja, females were actually more dominant and empowered in terms of turn allocation, contrary to certain cultural expectations regarding Thai women. The authors conclude that their findings illustrate how culture interacts with gendered CMD in complex ways.

Some of the findings in Baron's (2004) study seem to diverge from gender participation styles found in asynchronous CMD, as well. By analyzing 23 IM conversations

between American undergraduate students, she found that females used longer turns than males, especially in closing conversations. However, some of her results are in keeping with findings on asynchronous CMD: Females used more emoticons, a higher number of standard forms, more capitalization and punctuation, and were more socially-oriented than males. It should be noted that the data of this study comprised private CMD, as opposed to public CMD in the rest of the studies in this chapter.

Herring and Martinson's (2004) analysis of 1,622 messages from gender games at the Turing Game site found that participants chose nicknames appropriate to the gender they were performing in the game, but that the stylistic features of their messages resembled those of their real life gender identities more than their performed gender in the game. Moreover, the judges asked stereotypical questions, and judgments about gender identity were based on cultural stereotypes about women and men, rather than on the players' discourse behaviors. A different game environment using chat (among other modes of communication) is MMORPG games such as Everquest II. Huh and Williams (2009) analyzed the operating logs (including chat logs) and survey responses of 1,622 players of the game regarding their gender-swapping practices. Their findings indicate that gender swapping is very rare, employed mostly by males – especially homosexual males. They also found that female non-swappers engaged in more male-associated activities (such as fighting monsters) and more chatting than female swappers. The authors conclude that gender-swapping occurs mostly for practical, game-related reasons rather than exploring one's gender identity; they also suggest that the female identity may have more than one version, since certain of their findings did not seem in accordance with traditional patterns of female behavior.

Scholars have also paid attention to the linguistic styles of male and female teenagers. Subrahmanyan, Greenfield, and Tynes (2004) analyzed 815 lines of a conversation in a monitored teen chat room and found that gender identity is important in that online

environment, as seen by the use of "a/s/l" and the gender connotations in nicknames. But teenage girls seem to be more liberated with regard to initiating online relationships with the opposite sex, since CMD offers them implicit ways to do so without being stigmatized as they might be offline. In a similar study of monitored and unmonitored teen chat sites, Subrahmanyan, Smahel, and Greenfield (2006) found that males communicate more explicitly about sex and are more active. Females in contrast are more passive in sexualized interaction, communicate more implicitly about sex, and prefer nicknames with (female) gendered identities; the authors note that gender roles seem to be complementary but traditional.

Kapidzic and Herring (2011) found similar patterns in their multi-level linguistic analysis of 1,000 chat messages from five different teen chat sites. Even though the differences at the level of individual word choice were not striking, teen boys tended to use more articles, and teen girls expressed more emotion. At the discourse-pragmatic level, boys used more manipulative acts, whereas girls used more reactive acts. Moreover, boys adopted a more flirtatious and overtly sexual tone, in contrast to girls, whose messages tended to be friendlier and not as sexual. The authors' additional analysis of profile pictures also supported traditional gender roles, in that the girls presented themselves as more seductive and submissive, whereas the boys appeared more remote and dominant in their photos.

2.1.3. Social Media

The term *social media* refers to "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" (p. 61, Kaplan & Haenlein, 2010). This category can be considered separate from other modes of asynchronous and synchronous CMD, because the concept of (a)synchronicity is blurred in social media due to the different communication

options and modalities offered on the same platform, and sometimes in the same CMC tool (e.g., Facebook Inbox messages, which can be both synchronous and asynchronous).

The first gender and CMD studies in social media used corpora with data collected from MySpace, a social networking site that was very popular before Facebook. Thelwall (2008) analyzed the use of swear words in relation to gender, age group, and region (US and UK) in 767 MySpace profiles. His results show that almost all younger users' profiles and about half of the middle-aged users' profiles contain some swearing. Regarding the strength of swearing, UK and US showed some gender differences: Males in the US used more strong and moderate swearing than females, whereas there were no significant differences in strong swearing between men and women in the UK, especially among younger users. The author concludes that the assimilation of traditional male swearing by UK females in MySpace suggests deeper changes in gender roles in UK society related to the rise of "ladette culture." Another study (Thelwall, Wilkinson, & Uppal, 2010) of 1,000 public comments in MySpace forums focuses on emotion. The findings suggest that MySpace is a very emotive environment overall, with negative emotion appearing in a minority of comments (20%). Females were more likely to give and receive a greater amount of positive comments than males, but no gender differences were found regarding negative comments. The authors suggest that females are more successful in their use of social network sites because of their greater ability to harness positive affect through language.

In their analysis of 162 MySpace forum comments from threads with different gender participation distributions, Fullwood et al. (2011) found gender differences similar to those of most studies in asynchronous CMD: Females used more emoticons, repeated punctuation, hedges, tag questions, and abbreviations than males, while males swore more, referred to taboo subjects more frequently, and used more slang than females. However, analysis of the same features in the "About Me" sections showed different results: Except for abbreviations,

which were used more by females than males, there were no significant differences in the use of other features. The authors conclude that the more "androgynous" style in the "About Me" sections may be explained in terms of appeal, as a gender-neutral style of communication in one's self-representation in MySpace may appeal to a wider range of users. Taking into account that the data in this study are from UK users, the findings might also be explained by the rise of "ladette culture" in the UK, as suggested by Thelwall (2008).[4]

The microblogging site Twitter has also attracted a lot of attention, mainly from Natural Language Processing (NLP) scholars, because of the ease of collecting vast amounts of data using the Twitter API for use as training sets in machine learning algorithms. Rao et al. (2010) used a corpus of 405,151 tweets from 500 female and 500 male users to test two different models for gender identification: a sociolinguistic model and a purely lexical n-gram model. They found that females are more likely to use emoticons (especially the heart emoticon, whereas males prefer the grin and wink emoticons), ellipses, letter repetition, multiple exclamation marks, puzzled punctuation ("??!?!"), abbreviations like "OMG," disfluencies like "oh, hmm, ugh," and agreement than males. Their experiments show that the sociolinguistic model in combination with bigrams beginning with "my" performs best: Bigrams such as *my zipper, my wife, my gf, my nigga, my want, my beer, my shorts, my jeep* are male markers, whereas bigrams such as *my zzz, my yogurt, my yoga, my husband, my bf, my prof, my daddy, my research, my gosh* are female markers. The authors note, however, that their sample is heavily biased toward a younger population.

Burger et al. (2011) use a much larger dataset (4.1 million multilingual tweets from approximately 200,000 users) in order to test which single features or combination of features perform better in gender prediction. Their experiments indicate that the single most informative field is the user's full name, followed by their screen name and description; the

4 *Ladette culture* is a group term that describes females with gender reversal behavior and is related to the rise and eventual acceptance of binge drinking among UK females (Thelwall, 2008).

most accurate classification algorithm is one using a combination of the user's full name, screen name, description, and tweets. Their corpus analysis produced more strong female markers (*hair*, several variants of *love*, letter repetition, and emoticons) than strong male markers (*http* and *Google* variants). The authors note that the higher use of *http* by males does not indicate a higher use of links compared to females; upon closer analysis, they found that females also include links in their tweets, but prefer bare URLs.

Walton and Rice (2013) analyzed 3,751 tweets for valence, disclosure, and stage functions. According to their findings, tweets by females have more positive valence and disclose more overall and more backstage material (self-expression of an intimate nature) than tweets by males, affirming females' societal gender role as nurturing and emotional. Males did not have many disclosures in their tweets, which the authors explain as a greater need to manage their impression.

Even though Facebook has become the most popular social networking site, there are not many CMD studies using Facebook corpora. A reason for that may be Facebook's strict policy about using their API for data collection. One of the very few CMD Facebook gender studies is that of Wang, Burke, and Kraut (2013). They analyzed one million English status updates to explore topic preference by gender, taking age into account as well. They found that women older than 25 write much more about relationships and personal details than do men, whereas men are more likely to write about sports and abstract concepts; these findings are consistent with those of Herring and Paollilo (2006) for weblogs. In the 13-17 age group, though, their findings are more homogeneous: Both boys and girls complain a lot, write about their boyfriends/girlfriends, and use slang; but girls are still less likely to write about sports, while boys are less likely to write about family events. It is suggested by the authors that the common life stages of teenagers regarding relationships and school, as well as the use of

slang as a community identifier, may explain this homogeneity in their discourse styles, as reported also by Huffaker and Calvert (2005).

The *Open Vocabulary Approach* study of Schwartz et al. (2013), using a corpus of 19 million Facebook wall posts and messages, supports traditional gender patterns in terms of preferred lexicon: Females used more emotion words, first-person singular pronouns, and mentioned more psychological and social processes, whereas males used more swear words and object references. Additionally, the authors found that males tended to precede references to the opposite sex with the first person possessive singular pronoun ("my"), whereas females used other possessive pronouns and adjectives for similar references to the opposite sex, suggesting that males do not talk as much about other people's partners and are more possessive of their own.

Finally, the study by Ottoni et al. (2013) of another social media platform, Pinterest, also shows different gendered language preferences. The authors analyzed 550,436 female and 29,644 male Pinterest profiles for a variety of discourse and non-discourse features. They found that females convey more affection and positive emotion in their user descriptions, whereas men tend to interact in ways that assert their power and status, as indicated by their higher use of words connected to work, achievements, and money.

As can be concluded from the above sections, the majority of research in CMD supports the argument that traditional gender patterns in face-to-face communication carry over into online environments. At the same time, context – both the technological affordances (e.g., synchronous versus asynchronous) and the social dynamics of specific environments – plays an important role in understanding gendered language, especially in groups where one gender is more numerically predominant than another. Context, for example, may explain findings that appear contradictory to traditional gender patterns, such as in Huffaker and Calvert's (2006) study: The genre of diary blogs has been found to influence discourse style

more than the actual gender of the users (Herring & Paollilo, 2006), similarly to the topic of discussion in discussion lists (Thomson, 2006). Cultural context also interacts with gendered language; some of the different findings in Herring and Panyametheekul's (2003) study and de Oliveira's (2007) study may be explained in terms of Thai and Portuguese cultural norms, respectively. Similarly, Thelwall's (2008) and Fullwood et al.'s (2011) reports of a seeming merger of male and female styles may be explained in the context of a rising "ladette culture" in the UK. Moreover, context can also explain results such as the temporarily greater posting activity of women in otherwise male-predominant groups (Herring, 2010; Herring, Johnson & DiBenedetto, 1992, 1995), as well as the shifting of one gender's linguistic style towards that of another (Herring, 1996b).

Table 1 provides a summary of the main findings regarding male and female discourse styles from studies of asynchronous and synchronous CMD, based on the organization of findings in Herring (2003a), with the addition of a social media CMD section by the author.

Table 1

Summary of Gender Differences in Asynchronous CMD, Synchronous CMD, and Social Media.

	Males	Females
	Participation	
Asynchronous	more messages longer messages receive more responses based on ethic of agonistic debate and freedom of rules or imposition	fewer messages shorter messages receive fewer responses based on ethic of cooperation and politeness
	Discourse Style	

	active	passive
	less polite (violations of positive and negative politeness)	more polite (observations of positive and negative politeness)
	strong assertions	attenuated assertions
	absolute and exceptionless adverbials	hedges and qualifiers
	impersonal	adjectives
	use of presuppositions	emotional
	rhetorical questions	personal
	self-promotion	apologies
	disagreement with others	true questions
	flaming	support and agreement with others
	exclusive 1st person plural pronouns	inclusive 1st person plural pronouns
		graphical accents
		exclamation marks
	Participation	
	more messages	fewer messages
	get fewer responses	get more responses
	Discourse Style	
	explicit sexual references	implicit sexual references
	sexual, flirtatious tone	friendly and supportive tone
Synchronous	dominant	submissive
	manipulative	reactive
	articles	personal pronouns
	violent verbs/profanity/swearing	neutral/affectionate verbs
	evaluative judgments	graphical accents/laughter
	sarcasm/insults	attribution of feelings to self and others
	deviation from standard	emotional
	typography/orthography	adherence to standard typography/orthography
	Participation	
	greater participation in certain social media platforms (e.g., YouTube, Reddit)	greater participation in certain social media platforms (e.g., Pinterest, Tumblr)
	Discourse Style	
	1st person singular possessive pronoun ("my")	1st person singular personal pronoun ("I")
	object references	mention of psychological/social
Social Media	swear words	processes
	reference to taboo subjects	emotional
	slang	positive
	negative	personal
	abstract concepts	tag questions
	expression of power and status	abbreviations
	disagreement	ellipsis
		agreement
		repeated letters
		repeated punctuation

Note. Adapted from "Gender and (A)nonymity in Computer-Mediated Communication", by S. C. Herring and S. Stoerger, 2014, *The handbook of language and gender* (2nd Edition).

2.2. Language Change and CMD

A number of review articles address the issue of language change in CMD, focusing mostly on lexis and orthography/typography, adducing evidence from synchronic studies (Androutsopoulos, 2011; Baron, 1984, 2009), but there are very few purely diachronic studies evaluating language change in CMD. Herring (1998) made the first diachronic study of CMD to test anecdotal claims about language change empirically, using a corpus of 50-message samples from MsgGroup covering 11 years, sampled at roughly 2-year intervals. She analyzed syntactic complexity, formality, politeness, and variance, and based on her findings, dismissed any strong technological deterministic explanations. Herring concluded that social and ecological influences, such as user demographics, communicative purpose, the degree of social accountability in the relationships among participants, as well as concurrent changes in the culture of the Internet in general, had more influence on the observed changes in CMD than the medium.

Another diachronic study by Gao (2008) used the *apparent time construct* framework to study language change in Chinese Internet Language (CIL), using a diachronic corpus comprising data from five different internet modes from October 2002 to December 2007. Gao argued that what he refers to as "Chinese Internet Language" (CIL) is mostly used by younger individuals online and may spread to other age groups and beyond CMC, thus leading to changes in the Chinese language.

A smaller-scale diachronic study is that of Rowe (2011). The author studied the rapid evolution of a private language code in the email communication of two sisters, using a diachronic corpus of 245 messages collected January through May 1996. Although the time period is shorter than for traditional diachronic corpora, the study looks at linguistic features at regular time points. Rowe's findings suggest that the features of email communication, the

sibling relationship, and general sociolinguistic factors played an important role in the rapid development of the code.

Berdicevskis (2013, 2014) attempted to study change in the Russian language, adopting a technological deterministic approach that assigns CMC a central role, based on the "written turn" concept of Baron (1984): With the re-invigoration of written language by CMC, certain linguistic changes that are more easily diffused via written speech than oral speech would be more successful, such as a new neuter form in Russian. Using a diachronic corpus of 975 webpages produced by 729 authors covering a time period between 2001 and 2011, he also studied the emergence and fall of a new variety of online Russian, known as Olbanian language or Padonki (literally, 'scum'), that eventually fell out of fashion (2013).

Only one diachronic study has taken gender into account. Herring (1999a) studied contractions in CMD in an attempt to test the *principle of markedness assimilation,* using a diachronic corpus comprised of data from two discussion groups (MsgGroup for 1975-1986 and the LINGUIST list for 1990-1998). She found an interesting pattern in the use of contractions by males and females in the LINGUIST-L sample: Although contraction use decreased in both genders over time, use by females underwent a much steeper decrease, even though females originally used contractions more than men. Herring attributes the original difference to women's discourse ethic of avoiding contentiousness through use of more "informal" language, and the decline to a policy change that called for more formal interaction by the moderators; women "thereby [became] leaders in an institutionally-sanctioned, standardizing change" (p. 16), in accordance with previous diachronic sociolinguistic findings (Cameron, 2003; Nevalainen, 2011). At the time of this writing, no other diachronic studies exist where the main focus is change in the linguistic styles of men and women.

2.3. Issues in CMD Corpus Construction

Even though the majority of research on CMD has been conducted using naturally-occurring online data, organized efforts to build publicly-available CMD corpora and standardized schemes for their annotation have begun only in the past few years. Issues of representativeness, data processing techniques, delimitation of genres, and amount of contextual information, as well as issues of privacy and anonymity in CMC, require different methodologies from those for offline communication and have posed many complications for researchers working with CMD data (Androutsopoulos & Beißwenger, 2008; Beißwenger et al., 2014; King, 2009, 2015). As Androutsopoulos and Beißwenger note, "much research in the area has been based in small, ad-hoc data sets" because "there is a lack of standard guidelines for CMD corpus design and a lack of publicly available CMD corpora" (2008, n.p).

Beißwenger and Storrer (2008) categorize CMD corpora in terms of design principles, as shown in Table 2.

Table 2

Types of CMC Corpora.

Data edited for purposes of analysis? / The corpus was originally designed to be ...	No	Yes
... project-related	1 corpora of raw data	3 annotated corpora
... for general use	2 corpora of raw data	4 annotated corpora

Note. Reprinted from Corpora of Computer-Mediated Communication, by M. Beißwenger and A. Storrer, 2008, retrieved from http://www.cmc-corpora.de. Copyright 2008 by M. Beißwenger and A. Storrer.

Most corpus-based research on CMD has been conducted using project-related corpora of raw data, which have to be collected by individuals from the web or obtained from users of CMD environments (Beißwenger & Storrer, 2008). Project-based corpora are created for the purposes of studying a specific phenomenon, genre, or topic, and hence they are context-specific, which limits the number of research questions such corpora can be used to answer. Beißwenger and Storrer further note the absence of CMD components in national reference corpora and comment that the larger, publicly-accessible corpora specifically designed for analyzing CMD phenomena are unsatisfactory; seemingly, they are referring to the genres/modes covered. For this reason, recent efforts in Germany (Beißwenger et al., 2012; Margaretha & Lüngen, 2014), France (Chanier et al., 2014), the Netherlands,[5] and Switzerland (Stark, Ruef, & Ueberwasser, 2015) have been focused toward the goal of creating corpora for general use (reference corpora) that will represent CMD in their respective National Corpora. At the time of this writing, no similar projects for English language CMD appear to exist.

Despite the efforts aimed at addressing the lack of organized public CMD corpora, there is a striking imbalance regarding the two traditional types of corpora, synchronic and diachronic: Almost all corpora used in CMD research are synchronic. A possible explanation rests in the challenges of designing and collecting a diachronic corpus in terms of representation, since traditional sampling techniques may not be easily applicable to vast amounts of data. Another issue related to representation is the difficulty of finding data sources; some types of CMD data are very ephemeral, and almost all user-generated contents can be deleted by their creators without notice (King, 2015). Earlier CMD data found in discussion groups and mailing lists may be obtained either from list moderators (if the list is archived) or from public archives such as the Internet Archive or Google Groups (Baumann,

[5] The SoNaR project, cited in Beißwenger et al. (2012): http://www.lt3.ugent.be/en/projects/sonar/

2015); however, such data may be incomplete or difficult to collect. Synchronous CMD is even more difficult to find, since what log files exist are either archived on users' local computers or in public chat systems. Consequently, even though there are some data sources that can be used for historical research in CMD, the difficulties in access and collection of the data – as well as the absence of existing, ready-to use diachronic CMD corpora – may dissuade researchers who wish to conduct diachronic studies, thus leading to gaps in the literature.

Theoretical Framework

3.1. Gender

Studies of gender in CMD began in the early 1990s, when "the field of language and gender research was engulfed in a wave of social constructionism" (Holmes, 2007, p. 51). Social constructionism has been used as a theory of knowledge in a variety of fields and examines how jointly constructed understandings of the world have developed to form the basis of shared assumptions about reality (Leeds-Hurwitz, 2009). By the time gender studies in CMD began, gender studies in offline (spoken) communication had become largely polarized between two traditions within the social constructionist framework: the dominance framework and the difference framework (Freed, 1995). The *dominance framework* considers the different linguistic patterns of men and women to be a consequence of a male-predominant patriarchal society (e.g., Spender, 1998). Popularized by Deborah Tannen's work, the *difference framework* makes no reference to power imbalances, but considers gender differences to be a natural development of a gender-differentiated society where individuals acquire certain behaviors by being members of a particular linguistic community, in this case male or female (Tannen, 1990, 1994). Attempts to reconcile the two theoretical approaches in the 1990s added cultural and situational context as a crucial factor, arguing that gender is a complex collection of social practices within communities. These social practices are sometimes connected to personal attributes and power relations, but in a variety of ways that are subtle and may change (Cameron, 1999; Eckert, 2011). Consequently, *social constructionism* in gender studies takes a binary approach to gender regarding culturally gendered ways of talking ("masculine" and "feminine" interaction styles) and analyzes the dynamic ways that gender identity is constructed in context-embedded discourse (Holmes, 2007).

This study also adopts a social constructionist approach: It analyzes gender identity as it is socially constructed through the use of language in CMD and takes into account both the social and technological contexts where the discourse takes place. Although social constructionism adopts a binary approach to gender, the author recognizes that gender is not strictly binary; however, the use of a binary gender classification in this study is a practical necessity due to the limited gender information available in the data. USENET Newsgroups do not provide any gender information about the users, unlike many social media platforms nowadays, some of which have also expanded their gender options to include non-binary genders; for example, Facebook updated its profile gender options to 71 in 2015 (Facebook Diversity, 2015). We can only infer gender through the name and email information of the USENET users, unless they explicitly state their gender in the post content. Due to the large amount of data it contains, the HUNC corpus used in this study was annotated for gender using an automated method based on the name and email information of users and a list provided by the US Census database with probabilities of a name being male or female (Bourlai & Gao, 2017) – effectively adopting a binary approach to gender. Consequently, gender in this study is classified as **male** or **female**.

3.2. Language Change as a Result of Language Contact

Language contact occurs when speakers of two or more languages or language varieties interact and influence each other. The field of Contact Linguistics dates back to the 19th century and has grown out of historical linguistic research, which argues that language contact may lead to language change (Braunmüller & House, 2009). Language contact may involve face-to-face interaction, but also non-personal contact via written language (Braunmüller & House, 2009). Language contact in CMC has been studied regarding different languages and the effects of their contact online (Gao, 2008; van Gass, 2008), but it

may also be used to explain language change in gendered language in CMD. Following a social constructionist approach, we may consider the male and female linguistic styles as two different varieties of a language produced by social factors such as power inequalities and membership in different linguistic communities in the environment of offline communication. Those linguistic varieties have been used in different ways by both genders depending on cultural and situational context in the environment of offline communication where they have been in contact. CMC, however, is a different environment where the manner of contact, the social structures, and the cultural and situational contexts may differ from those of offline communication. Consequently, the contact of these two linguistic varieties in this different environment may have effects that could lead to language change. This change could be divergence, if more differences appear or existing differences become more pronounced, or convergence, resulting in more gender-similar language. The language could also continue to diverge again after converging, showing a reversal in the linguistic features used by males and females.

It should be noted that males and females are in contact both in online and offline environments, and the cause of changes in their language use may reside in factors in either of the two environments, or in both. While there are several theories that explain the causes of convergence and divergence in language (Braunmüller & House, 2009), the focus of this study is not to identify those causes. Instead, the study aims at identifying *if* any changes occurred and if so, *how* they have manifested, as the first step toward identifying and understanding those causes, and, in turn, the effect of CMC on the linguistic styles of men and women and on language in general. In this study, **change in the gender pattern of a linguistic feature** is operationalized as any change in the *distribution* of the female and male frequency of use for that linguistic feature over time. For example, the hypothetical linguistic feature in Figure 1 exhibits change in its overall use and its use by each gender over time, but

not in its gender frequency distribution: The difference between the female and male frequency remains the same or similar over time. Consequently, even if there is overall change or change for each individual gender in the linguistic feature over time, the gender pattern may remain the same.

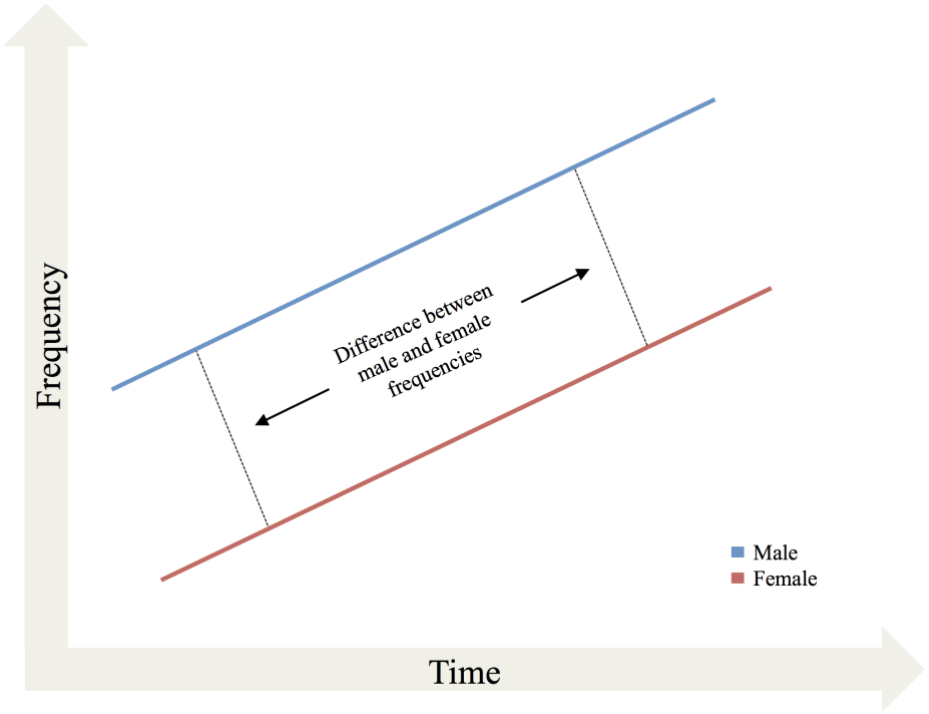


***Figure 1*. Change in linguistic feature overall and for each individual gender over time with no change in the distribution of female and male frequencies.**

The theoretical framework of language change as a result of language contact allows us to use the concepts of convergence and divergence in order to understand the manner in which changes in the language use of men and women manifest in CMD. Convergence, divergence, and reversal of gender patterns are operationalized in this study in terms of the change in the frequencies of male and female use of each linguistic feature studied. If the frequencies of use of a linguistic feature become more similar over time, this is considered a **convergence**, as illustrated, hypothetically, in Figure 2.

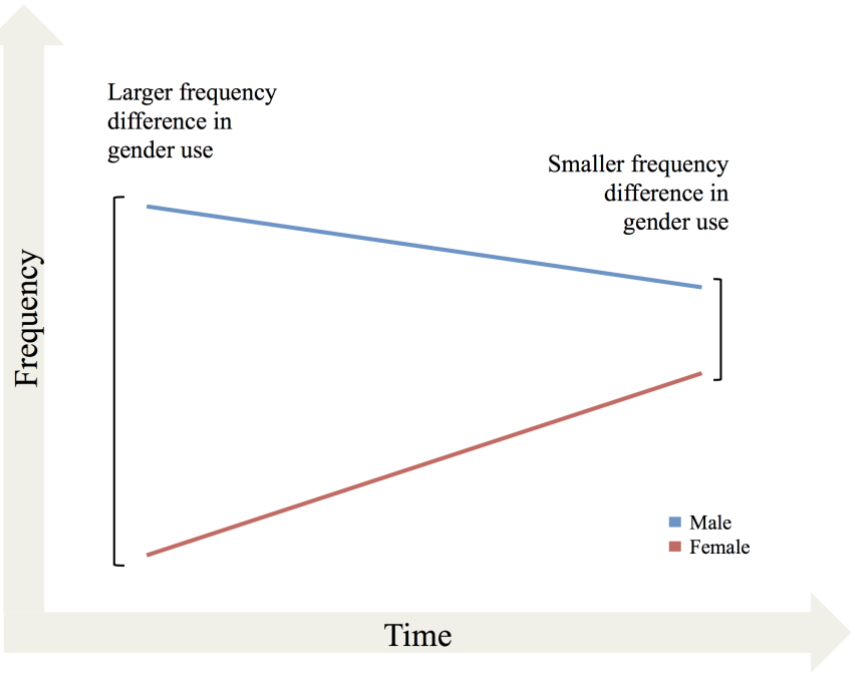


***Figure 2*. Convergence of female and male frequencies over time.**

In contrast, if the difference in frequencies becomes greater over time, it is considered a **divergence,** as shown, again with hypothetical data, in Figure 3.

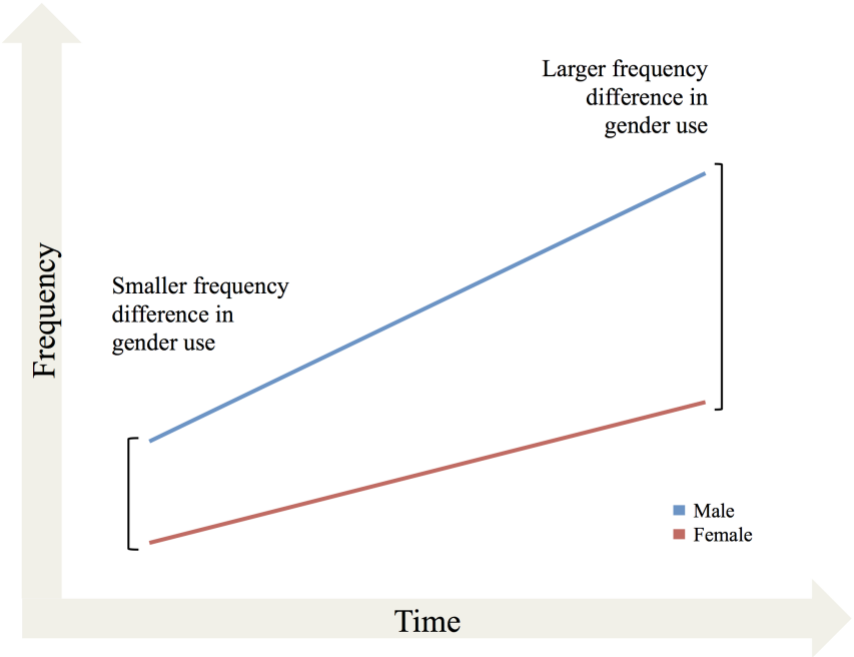


***Figure 3*. Divergence of female and male frequencies over time.**

A **reversal of a gender pattern** occurs when the initial relative frequencies associated with the male and female use of a linguistic feature are "flipped": For example, if females make lower use and males make higher use of a linguistic variable at the beginning of the time period studied, but females make higher use and males make lower use of that feature at the end of that time period, this is considered a reversal of the gender pattern for that linguistic feature. As illustrated in Figure 4, during the process of the pattern reversal, the female and male frequencies converge and then diverge again in the opposite direction.

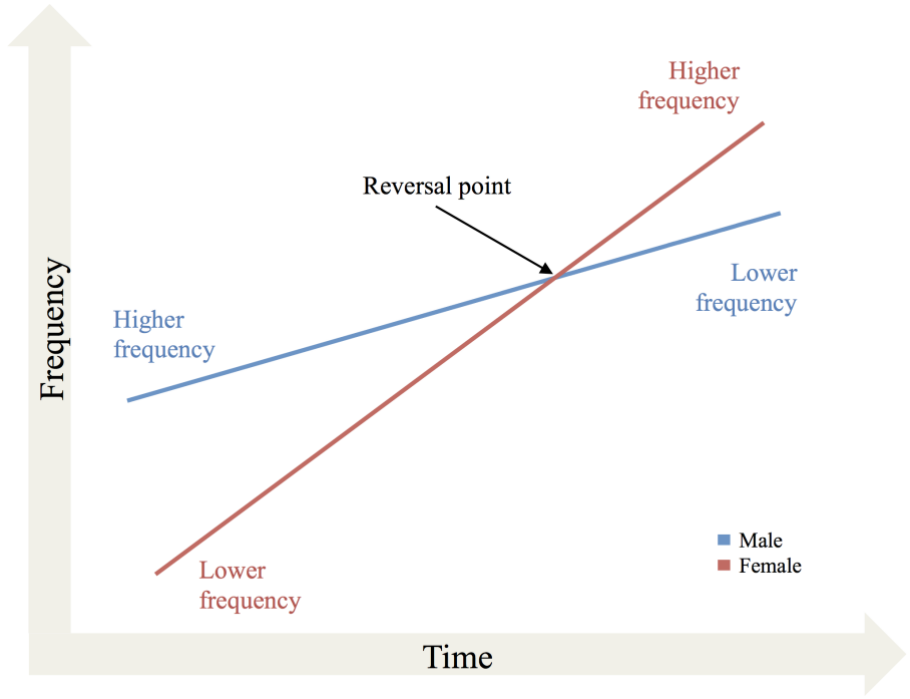


***Figure 4*. Reversal of gender pattern over time.**

Methodology

4.1. Research Question

The purpose of this study is to explore the language usage of men and women in CMD from a diachronic viewpoint, in order to identify any changes and the development of those changes. Thus, the research question that guides the study is:

RQ: How, if at all, have gender patterns changed over time in CMD?

This study is quantitative: The frequencies of gender markers are calculated separately for males and females and then compared at different time points over the time period studied in order to identify variations.

4.2. USENET newsgroups and HUNC

4.2.1. USENET Newsgroups

USENET newsgroups are repositories within the USENET system comprised of user messages that are transmitted using the Network News Transfer Protocol (NNTP) (Emerson, 1983; Wikipedia, 2017c). USENET newsgroups can be described as a hybrid between email and a web forum (Usenet Archive, 2017), and they could be considered a precursor to present-day forums and social networks. Even though their popularity has declined in the past few years, many newsgroups are still active today as Google Groups.

Most early USENET newsgroups originally appeared with the *net* prefix around 1981, but the system underwent a major renaming process in 1986-1987, resulting in the creation of seven major hierarchies: *comp*, *news*, *sci*, *rec*, *soc*, *talk*, and *misc* (Wikipedia, 2017a, 2017c).

The hierarchy *alt* (short for *alternative*) was not part of the so-called *Big 7*, but was created to allow groups more freedom and with fewer rules (Wikipedia, 2017a). In 1996, the *humanities* hierarchy was added, and the name was changed to the *Big 8* (Wikipedia, 2017a). These hierarchies are listed in Table 3:

Table 3

The 'Big 8' Hierarchies.

Hierarchy	Topic
comp.	Computer-related Discussions
humanities.	Humanities topics
misc.	Miscellaneous topics
news.	Newsgroup-related matters
rec.	Recreation and entertainment
sci.	Science-related discussions
soc.	Social discussions
talk.	Talk about various controversial topics and discussions with no obvious categorization

Note. Adapted from "Big 8 (Usenet)," by Wikipedia, 2017.

USENET newsgroups are considered by many to be the "original wild west" of the Internet. According to Dzhingarov (2018), "Usenet was the internet before the internet was the internet; it's a piece of digital history that still accessible and useful today" (para. 2). The USENET population was fairly homogeneous at first, with users coming mainly from academic and professional networks. However, the user demographics changed dramatically after September of 1993, when USENET access started becoming available to all subscribers of internet service providers (e.g., America Online). This resulted in a wave of new users, also known as the "Eternal September": Since USENET users largely came from academic networks before 1993, the term refers to the waves of Freshmen every September, who would either acclimate to the USENET culture or "slowly fade from the service" (Dzhingarov, 2018, Usenet History 101, para. 4). This change in user population had a major impact on USENET culture and "netiquette" (see guidelines in Kehoe, 1996). Over the years, USENET became

infamous for its "flame wars," intense arguments that erupted often in discussion threads. Thus aside from the wealth of information shared in its newsgroups, USENET is a valuable resource as an early example of online communication that foreshadows contemporary behaviors in CMD.

At the time of this writing, the biggest collections of USENET newsgroup archives that are publicly available are the *Usenet Archive* and *Google Groups*. The *Usenet Archive* provides downloadable archives in MBOX format for a number of newsgroups; however, upon careful examination, it was discovered that a substantial amount of data is missing for certain time periods. *Google Groups* currently have the full archives for newsgroups dating back to 1981; the data before 1995 were donated by individuals, whereas the data after 1995 were archived by *DejaNews*, which was purchased by Google in 2001 (Wikipedia, 2016b). Unfortunately, *Google Groups* underwent an interface change in 2013 that severely reduced their search functionality and created many obstacles for researchers wishing to collect messages using automated methods.

4.2.2. The Historical USENET Newsgroups corpus (HUNC)

The need for a diachronic corpus comprising USENET Newsgroups data for this study sparked the idea for the *Historical Usenet Newsgroups Corpus* (HUNC) project, a diachronic corpus that will be available to other researchers upon its completion in order to promote historical research in CMD (Bourlai & Gao, 2017). Even though the creation of the corpus was prompted by the data needs of this study, it was not created exclusively for this project; rather, it was designed as a data resource to address a variety of research questions. Thus, the HUNC is an annotated corpus for general use according to Beißwenger and Storrer's classification (2008).

The HUNC includes 312 English-language USENET newsgroups from nine hierarchies with different topics and user populations to ensure representativeness, and it covers a time period from 1981 until 2016.[6] All available original newsgroups with the *net* prefix were selected, along with newsgroups from the following hierarchies that are part of the Big 8: *alt*, *comp*, *misc*, *new*s, *rec*, *sci*, *soc*, *talk*. The process of selection of the latter newsgroups was based on one of the following two criteria:

a) The newsgroups have data available beginning before January 1991 (covering at least a 25-year period at the time of collection)
b) or they are the successors of previous *net* newsgroups after the renaming process of 1986-1987.

The data for HUNC were collected from the Google Groups website using a Java crawler between May 2016 and May 2017. The crawler collected the data from each group in reverse chronological order from the latest thread to the earliest thread; it also crawled in reverse chronological order from the latest post in each thread to the earliest. The server seems to limit the collection of data to approximately 100,000 threads for each group, in that it does not respond after reaching that number. Consequently, the earliest messages for a few groups with more threads than the aforementioned limit were not collected. The posts were collected in their original format with all their metadata, including their post and thread URL on Google Groups, their author name and email/account identifier, the date they were posted, and their subject, as well as the newsgroups in which they appeared.[7] Table 4 summarizes the distribution of the data:

Table 4

Distribution of Data in HUNC.

Hierarchy	Number of Newsgroups	Number of Posts

[6] Since the data collection started in 2016, a number of groups do not have complete data for the year 2016.
[7] Posts in USENET usually appeared in more than one newsgroup, a practice called 'cross-posting.'

alt.	23	6,377,128
comp.	28	5,673,276
misc.	13	3,152,852
net.	105	297,213
news.	8	496,795
rec.	75	24,338,845
sci.	29	5,696,486
soc.	16	3,694,601
talk.	15	4,054,764
Total	**312**	**53,781,960**

The corpus follows a hierarchical organization based on date: For each group, the data are categorized by year, then by month, and each post is an individual file (Bourlai & Gao, 2017). The files include both metadata about the post and the actual content of the post, using simple XML tags for annotation. In the future, the corpus will also be converted into multiple CSV (Comma Separated Values) files for ease of use and distribution: Each CSV file will contain the data and metadata for an individual newsgroup.

The metadata, "noise" cleansing, annotation, and gender classification processes described below have only been performed for the subset of data (7.7 million posts) that was used in this study (see section 4.3). Once the annotation is complete for the entire corpus, the CSV version of HUNC will be uploaded on the *Kaggle* website (Kaggle, 2018), where it will be publicly available for other researchers to use. Since the CSV version may result in a loss of format in the content of the posts, the individual text file version of the corpus will also be available upon request.

4.2.2.1. Metadata Extraction and "Noise" Cleansing

The metadata for each post in the corpus was extracted using Java and Python scripts from the original metadata included in the raw collected data. Among the metadata extracted were the Post URL, Thread URL, Date, Author, Subject, and Newsgroups in which the post appeared (cross-posting). The author metadata that was extracted was the user information in

the “From:” field of each post: This comprised a full or partial name or a username and/or a full or partial email/account identifier. Once the above metadata was extracted, the data underwent a “noise” cleansing process (i.e., removal of additional meaningless information) using Python scripts: The original metadata at the beginning of the post was removed in order to keep only the actual content of the post.

4.2.2.2. Annotation

After the extraction of metadata and the “noise” cleansing process, unique IDs for each post, thread, and author were created; these were added as annotations at the beginning of each post, along with the date, subject, and newsgroups metadata. The metadata for the full date had to go through a normalization process in order to have uniform format throughout the corpus; as explained in the limitations below, the raw data format was not consistent in the corpus. The software used for this process, as well as for the gender classification described below, is *PolyAnalyst 6.5*. This software was chosen for its feature robustness, ease, and speed of use, since it supports “all steps of a data analysis process from data loading and manipulation, to advanced text and data analysis, and to custom reporting” (Megaputer Intelligence, 2018). The *Normalization* node successfully converted the varied date formats into the selected unified date format.

Table 5 presents a list of the annotated metadata and features in the posts and their description. The purpose of the annotation of these features is to facilitate the creation of subcorpora and the analysis of the data. For example, the Thread ID allows researchers to reconstruct a thread for conversation analysis. Similarly, the Author ID may be used to retrieve all the posts written by the same author across groups, regardless of Author name inconsistencies, in order to evaluate the effect of the newsgroups’ context on the author’s discourse.

Table 5

Annotated Features in HUNC.

Feature	Description
Post ID	Unique ID # of post
Thread ID	Unique ID # of thread
Full Date	Full post date (timestamp)
Year	Year of post publication
Author ID	Unique author ID #
Author Name	Author name as it appears in the "From" field
Author Gender	Male, female, unknown
Newsgroups	Newsgroups in which the post appeared (cross-posting)
Subject	Post subject as provided in original metadata
Content	Content of post

4.2.2.3. Author Gender Classification

Since the majority of posts included the full or partial name or nickname of the author, it was decided to use an automated name-based approach for the gender classification due to the size of the data. The gender classification in the default People Entity Extraction node of the *PolyAnalyst 6.5* software (Megaputer Intelligence, 2018) uses a name-based approach similar to the U.S. Census-based method (Mislove et al., 2011; Nilizadeh et al., 2016). The *PolyAnalyst 6.5* software includes name dictionaries in different languages with assigned gender in its default installation; it also allows modification of the dictionaries or the addition of custom dictionaries. The node also has the post-processing option of normalizing and aggregating the extracted entities (in this case the authors), which helped identify the majority of unique authors in the dataset with name/nickname variations or misspellings. Since both the U.S. Census-based method and *PolyAnalyst 6.5* use a binary approach to gender, the corpus follows a binary gender classification, as well; authors whose gender could not be identified were labeled as 'Unknown.' The default People entity of *PolyAnalyst*

6.5 classified the gender of authors in 4.5 million posts out of the 7.7 million posts that comprise the subset of data selected for this study. In addition to the default People Entity Extraction of *PolyAnalyst 6.5*, a custom entity was created to identify nicknames that use gendered words (e.g., *lady, lord, girl, boy, mom, daughter*, etc.). The gender annotation was successful for cases where there was a non-ambiguous[8] first name existing in the dictionaries or a username including a gendered word. The Expanded Pattern Definition Language (XPDL) of *PolyAnalyst 6.5* allows for robust rules that significantly improve the accuracy of the results: For example, nicknames such as *my aunt's favorite* or *your mom is a bitch* were excluded, since the gendered words may not refer to the author of the post. A sample of 2,000 cases gender-annotated by *PolyAnalyst 6.5* was manually evaluated by the author for accuracy, which was 98.35%. The cases that were not classified by *PolyAnalyst 6.5* included names with unconventional spelling (e.g., $cott, Ernest0, Eveleen, Eriq, etc.), non-English names transliterated using the Latin alphabet, usernames with no gendered words, email/account identifiers with no name information, and full names with first names as initials. Out of those remaining cases, a random sample of 52,035 authors were manually classified as male or female by the author and a group of four undergraduate research assistants.

4.2.2.4. Limitations of HUNC

There were several challenges associated with the data collection and annotation of HUNC that may be extended to the collection and annotation of diachronic CMC data in general.

First, CMC data in earlier years were usually generated by a smaller number of users who had access to computers at that time. Consequently, there were fewer posts for those years; this was reflected in the HUNC corpus for the data before the early 1990s. The

[8] For example, the name 'Alex' is labelled as both male and female.

selection criteria during the design process of the corpus aimed at including newsgroups that would ideally cover at least a 25-year period; however, some were inactive during certain years, resulting in data imbalances across newsgroups. Moreover, due to limitations posed by the Google Groups server, we were unable to collect the earliest data of groups with more than 100,000 threads.

Second, because of technological changes throughout the years, the format of the data may not be uniform throughout a diachronic CMD corpus. This was especially true for the data of HUNC, and it made the metadata extraction and annotation of the corpus more challenging. For example, the dates of the posts appeared in several formats: *9 feb 92*, *9 feb 1992*, *9 February 1992*, *02/09/1992*, *09/02/92*, etc. This created issues during the automated extraction of date metadata and the organization of the posts in each group based on their date. Not only were there differences in the format of the metadata, but also in the format of the content of the post: Quoted text would be preceded by different symbols like > or # at the beginning of each line, or by tabs, or without any formatting.

Finally, the *"From:"* field in the raw posts sometimes contained limited, partial, or inaccurate author information. In some cases, the full name or nickname of the author was not included along with the email/account identifier; sometimes, the latter would be partial, as well. Moreover, due to the technological affordances of the USENET system, users could post using variations of their name/nickname and email/account identifiers. For example, both *1341Bil...@acid.com* and *1348Bil...@acid.com* belong to the same author (Billy Whizz) and also have part of the email/account identifier missing. There were also cases where the author used completely different email/account identifiers because their previous one(s) might have been banned from a newsgroup; in such cases, the authors usually included their name or nickname, in order for people to recognize them. In addition, since computer access was limited during the early years of USENET, people would post in newsgroups using the

accounts of other users who had access, or they would have shared accounts. Consequently, the name in the *"From:"* field of the post and the signature at the end of the post sometimes offered limited information. This was especially problematic for automated gender classification, as well as for tracking the participation of individual users in the newsgroups over time.

The author considered carefully the above limitations when selecting the final subset of the data for the corpus used in this study, as well as in interpreting the results of the analysis. For example, accounts that included both a male and female name in the *"From"* field were excluded from the data, since the current gender classification methodology could not identify who was the author of a specific post; such information could only be gained by examining the signature of the actual content.

4.3. Data

The data that were used in this study comprise a sub-corpus of the HUNC representing 47 USENET newsgroups and a total of approximately 3.8 million messages: That was the size of the final dataset, after an additional data cleansing step described below. This sub-corpus was designed following the same principles for the creation of diachronic corpora as were used in HUNC. The data had to be selected in a principled way, with the deliberate inclusion of particular genres of a particular sample size in order to be representative of a phenomenon and to ensure the validity of results (Kennedy, 1998). The criteria for the selection of data were the following: (a) All major hierarchies of the HUNC were represented in the subcorpus with at least one newsgroup, and (b) the data were continuous, which in this case meant that they had data for the years used as time points in the study. It should be noted that each newsgroup was concatenated with its *net* predecessor from which the 1982-1986 data were acquired. For example, *net.women* (1983-1986) was

renamed to *soc.women* (1986-present), thus they were considered the same group for the purposes of this study.

In order to account for data gaps in certain years due to the inactivity of newsgroups or data collection issues, it was decided that the time points would consist of data from two consecutive years; for example, the first time point comprised data from 1982 and 1983. The time points studied were at four-year intervals: A small pilot study that aimed to explore the USENET Newsgroups data suggested that a four-year interval allows for changes to be identified without missing important points in their progression (Bourlai, 2016). Originally, the study covered a period between 1982 and 2015; however, USENET newsgroups became less popular after the appearance of social media sites, resulting in a smaller number of posts and a significant number of spam messages in most newsgroups. During the analysis of the data, the author noticed discrepancies in the results for the time points after 2007. After a closer exploration of that data, it was discovered that it contained numerous spam posts that created "noise" and heavily affected the results due to the smaller number of posts for those years. The author thus decided to limit the period studied from 1982 to 2007, in order to ensure the validity of the results.

While a 25-year time period may seem very small in comparison to traditional diachronic studies of offline language, it has been suggested that language change is accelerated in CMC (Stein, 2006). In addition, Herring (1998) found evidence of change in language use online between 1975 and 1986, within only an 11-year period. Table 6 presents the list of the newsgroups selected by hierarchy (see Appendix B for the distribution of data by individual group). With the exception of the *alt* hierarchy, the distribution of newsgroups across the hierarchies in the sub-corpus was very similar to the distribution of newsgroups across the hierarchies in the HUNC. Moreover, the distribution of data by time point (see Appendix B) reflects the popularity and activity of USENET newsgroups throughout the time

period studied: The available data slowly increased in size after the 1990-1991 time point when internet access became more widespread among the general public and the newsgroups were at their peak of popularity, and decreased after the 2002-2003 time point with the appearance of social media, which diminished the popularity of newsgroups.

Table 6

List of USENET Newsgroups Included in the Study.

Hierarchy	Number of Newsgroups	Names of Newsgroups
alt.	1	alt.flame
comp.	10	comp.arch, comp.dcom.modems, comp.lang.ada, comp.lang.apl, comp.lang.forth, comp.lang.lisp, comp.lang.misc, comp.lang.prolog, comp.lsi, comp.periphs
misc.	5	misc.kids, misc.legal, misc.misc, misc.taxes, misc.wanted
news.	1	news.software.b
rec.	17	rec.arts.books, rec.arts.drwho, rec.arts.tv, rec.birds, rec.food.veg, rec.games.bridge, rec.games.empire, rec.games.misc, rec.games.pbm, rec.games.trivia, rec.puzzles, rec.pets, rec.scuba, rec.skydiving, rec.sports.baseball, rec.sports.hockey, rec.video
sci.	5	sci.astro, sci.lang, sci.math, sci.research, sci.space.shuttle
soc.	5	soc.college, soc.misc, soc.motss, soc.singles, soc.women
talk.	3	talk.bizarre, talk.religion.misc, talk.humor
Total		*47*

Table 7 shows the distribution of user gender in the data. There was a striking difference between the number of male and female posts, with female posts comprising only 12% of the dataset. Appendix B shows the very limited female data in earlier years, as well as the only newsgroup with a higher number of female than male posts: *net.kids/misc.kids*. Even though the gender distribution in the corpus is in accordance with previous literature, it should be noted that there was a large number of posts in the selected data of the newsgroups

where gender could not be identified and which were not included in the study. For this reason, the gender marker of participation (number of posts) is not included in the analysis of this study, since not all posts belonging to each newsgroup were included in the final dataset. In a hypothetical case where there would be enough posts by women in the data with unidentified gender that it would significantly affect the ratio of female to male posts, it might suggest an attempt by women to use names or nicknames to conceal or make ambiguous their gender identity, in order to avoid unwanted attention and possibly harassment – a practice that is supported by previous literature (Bruckman, 1993).

Table 7

Distribution of Author Gender in the Subcorpus Used in the Study.

Gender	Number of Posts	Percentage of Posts
Male	3,298,548	87%
Female	491,154	13%
Total	***3,789,702***	*100%*

Some additional data cleansing steps were taken in order to reduce noise in the data that may affect the validity and accuracy of the results in this study: The quoted text and signatures in the post were removed using Python scripts. For example, the quoted text of a male in a post of a female could skew the frequencies of gender markers toward a more "masculine" linguistic style. The "signatures" at the end of posts, which were especially popular in the earlier years studied, usually included ASCII art or quotes from other people than the author; those could also affect the frequencies of some of the gender markers studied. The author manually checked posts from different newsgroups and years to identify the different patterns of quoting and signatures, then went through an iterative process of refining patterns and removing the noise. While a check of random posts indicated that they

were successfully removed, there might still be patterns that were not identified; however, those would be very rare and would likely not affect the results, considering the size of the dataset.

4.4. Analytical Methods

This study was informed by the Computer-Mediated Discourse Analysis (CMDA) methodological framework. CMDA is a methodological toolkit adapted from language-focused disciplines, from which researchers may select the methods that are appropriate to their data and research questions (Herring, 2004). The methodological toolkit is organized around four domains (or levels) of language (structure, meaning, interaction, social behavior) and one non-linguistic level (participation). Each of these language levels includes different linguistic phenomena related to communication issues, which may be studied using specific methods.

More specifically, this study drew methods from Corpus Linguistics, since it is a corpus-based study that uses Natural Language Processing (NLP) methods for the automated analysis of large amounts of text (Baker, 2010). NLP methods originate in the fields of Computational Linguistics, Computer Science, and Artificial Intelligence, and are based on statistical models and algorithms of Machine Learning (ML) that enable computers to classify text in categories according to its features (Jurafsky & Martin, 2009). NLP methods identify patterns using the structure of the language, which are often associated with semantic or pragmatic functions; consequently, they belong to the structure domain or level in the CMDA framework.

The data were analyzed using *Linguistic Inquiry and Word Count 2015 (LIWC2015)* and *PolyAnalyst 6.5*. Both tools analyze textual data using Natural Language Processing methods. The *LIWC2015* software is a new and improved version of the popular *LIWC2007*

(Pennebaker et al., 2015) and included CMD data from Twitter datasets in its development. According to the developers, the *LIWC* software was developed in order to "provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech" (p. 1). *LIWC2015* uses an internal dictionary comprising almost 6,400 words, word stems, and select emoticons grouped in different categories (Pennebaker et al., 2015).[9] The software matches each word in a given text file to existing words in its dictionary and provides an output file with the scores in each category. Except for Word Count and Words Per Sentence, all other categories are presented as the total percentage of the words belonging in that category; Analytical Thinking, Clout, Authenticity, and Emotional Tone are summary language variables calculated by algorithms based on previous language research and are standardized scores that have been converted into percentiles (LIWC, 2017). The list of categories in the default dictionary of *LIWC2015* is presented in Appendix C.[10]

The *LIWC2015* software was chosen because several of the linguistic categories included in the software have been identified as gender markers according to previous research (see Table 9). However, the author used all the categories provided by the software as variables in the study, in order to allow for the possible emergence of new markers that have not been identified until now. The only categories not used were Time Orientations (Past Focus, Present Focus, Future focus) and Relativity (Motion, Space, Time), as well as Netspeak[11] from the Informal Language category, because they were not deemed relevant and/or appropriate.

[9] A word may belong in more than one category in the dictionary. Pennebaker et al. (2015) give the example of the word 'cried,' which belongs in five categories: sadness, negative emotion, overall affect, verbs, and past focus.

[10] *Words in category* refers to the number of different dictionary words and stems that make up the variable category. All alphas were computed on a sample of ~181,000 text files from several of the language corpora used to develop this software. Uncorrected internal consistency alphas are based on Cronbach estimates; corrected alphas are based on Spearman Brown (Pennebaker et al., 2015).

[11] The Netspeak category was deemed redundant in this study because certain items included in the category are studied individually (i.e., emoticons) and others are not relevant to the online environment studied. The category

An important gender marker that is not included as a separate category in the default dictionary of *LIWC2015* is emoticons; *LIWC2015* only includes four basic emoticons in the categories of Netspeak and some emotion-related categories. Even though *LIWC2015* allows for the creation of custom dictionaries by researchers, it was not possible to create a custom emoticon dictionary because of the special characters included in emoticons. Consequently, the author used the Taxonomy node in *PolyAnalyst 6.5* in order to identify a set of emoticons in the posts and classify them into positive and negative. Using the Score Taxonomy and Derivation nodes, the raw frequencies for positive, negative, and total emoticons in each post were extracted and converted to percentages per total word count, in order to match the results of the *LIWC2015* software. The emoticons included in the positive and negative categories, presented in Table 8, were based on the *List of Western Emoticons* provided by Wikipedia (2018) and classified into positive and negative by the author. The positive and negative emoticon categories were mutually exclusive.

Table 8

Positive and Negative Emoticon Categories.

Polarity	Emoticons
Positive	*\0/* *\o/* 8) 8-D 8D =-D =3 =D =] \0/ \o/ :) :-) :-)) :-D :3 :> :D :] :^) :c) :o) :} =) =-3 =-D =3 =D B^D X-D XD x-D xD :') :'-) :'-) :* :-* :-X :X :^* *) *-) 0:) 0:-) 0:-3 0:3 0;^) 3:) 3:-) :-, :-P :-b :-p :3 :> :P :X :b :p :} ;) ;-) ;-] ;D ;] ;^) =-3 =3 x-p xp =p X-P XP d: }:) }:-) >:) >:-) >:P >;) O:-) @>-->-- @}-;-'--- <3

is more relevant to synchronous online communication and social media, since it includes mainly abbreviations and/or acronyms, "misspellings", and lexical items such as "follow" and "retweet." Those features grouped together do not represent online communication in general and would be more informative as separate categories.

Negative	:# :-# :'(:'-(:(:-(:-< :-[:-c :< :[:c :{ ;(:-|| :@ >.< >:(>:[D-': D8 D: D:< D; D= DX v.v :-. :-/ :-| :/ :L :S :\\ :| =/ =L =\\ >:/ >:\\ :$:-$ </3 8-0 :-O :-o :O :o >:O O-O O_O O_o o-o o_O o_o

Table 9 presents all the variables analyzed in this study grouped into categories based on the categorization of *LIWC2015* (Pennebaker et al., 2015); except emoticons, the variables are all categories included in LIWC 2015 (see appendix C for further information). If a variable has been explored as a gender marker in previous CMD literature, it is indicated with an asterisk along with some examples of relevant studies.

Table 9

List of variables analyzed in the study grouped in categories.

Category	Variable	
Summary Language Variables	Word count*	e.g., Baron, 2004; Bucholtz, 2001; Herring, 1992a, 1992b; Selfe & Meyer, 1991
	Analytical Thinking*	e.g., Savicki, Lingenfeltr, & Kelley, 1996 (fact-oriented language)
	Clout	
	Authenticity	
	Emotional Tone*	e.g., Guiller & Durndell, 2007; Kapidzic & Herring, 2011
	Words per Sentence	
	Words Longer than 6 Letters	
	Dictionary Words*	e.g., Baron, 2004 (standard forms)
Pronouns	Total Pronouns	
	Personal Pronouns*	e.g., Argamon et al., 2007; Herring & Paolillo (2006)
	1st person Singular Pronoun*	e.g., Ottoni et al., 2013; Savicki et al., 1996; Schwartz et al., 2013
	1st Person Plural Pronoun*	e.g., Ottoni et al., 2013
	2nd Person Pronoun*	e.g., Ottoni et al., 2013
	3rd Person Singular Pronoun*	e.g., Ottoni et al., 2013
	3rd Person Plural Pronoun*	e.g., Ottoni et al., 2013
	Impersonal Pronouns	
Common Grammatical Features	Total Function Words	
	Articles*	e.g., Argamon et al., 2007; Herring & Paolillo (2006); Kapidzic & Herring, 2011; Schwartz et al., 2013
	Prepositions*	e.g., Argamon et al., 2007
	Auxiliary Verbs*	e.g., Argamon et al., 2007
	Common Adverbs	
	Conjunctions*	e.g., Argamon et al., 2007

	Negations	
	Common Verbs	
	Common Adjectives*	e.g., Schwartz et al., 2013
	Comparisons	
	Interrogatives*	e.g., Savicki et al., 1996 (questions)
	Numbers*	e.g., Argamon et al., 2007; Huffaker & Calvert, 2005
	Quantifiers	e.g., Argamon et al., 2007
Affective Processes	Total Affective Processes	
	Positive Emotion*	e.g., Ottoni et al., 2013
	Negative Emotion*	e.g., Ottoni et al., 2013
	Anxiety*	e.g., Ottoni et al., 2013
	Anger*	e.g., Ottoni et al., 2013; Schwartz et al., 2013
	Sadness*	e.g., Ottoni et al., 2013
Social Processes	Total Social Processes*	e.g., Baron, 2004 (social orientation); Ottoni et al., 2013, Schwartz et al., 2013
	Family*	e.g., Ottoni et al., 2013
	Friends*	e.g., Ottoni et al., 2013
	Female References*	e.g., Herring, 1993, 2010
	Male References*	e.g., Herring, 1993, 2010
Cognitive Processes	Total Cognitive Processes*	e.g., Huffaker & Calvert, 2005 (cognitive terms); Ottoni et al., 2013
	Insight	
	Causation	
	Discrepancy	
	Tentativeness*	e.g., Fullwood et al., 2001 (hedges); Herring, 1996d; Herring, Johnson, & DiBenedetto, 1992 (hedges)
	Certainty*	e.g., Herring, 1994; Huffaker & Calvert, 2005
	Differentiation	
Perceptual Processes	Total Perceptual Processes	
	See	
	Hear	
	Feel	
Biological Processes	Total Biological Processes	
	Body	
	Health*	e.g., Ottoni et al., 2013
	Sexual*	e.g., Kapidzic & Herring, 2011; Ottoni et al., 2013; Subrahmanyan, Smahel, & Greenfield, 2006
	Ingestion*	e.g., Ottoni et al., 2013
Drives	Affiliation	
	Achievement*	e.g., Huffaker & Calvert, 2005 (accomplishment); Ottoni et al., 2013
	Power*	e.g., Ottoni et al., 2013; Savicki et al., 1996 (status indicators)
	Reward	
	Risk	
Personal Concerns	Work*	e.g., Ottoni et al., 2013
	Leisure*	e.g., Ottoni et al., 2013
	Home*	e.g., Argamon et al., 2007; Ottoni et al., 2013
	Money*	e.g., Ottoni et al., 2013
	Religion*	e.g., Argamon et al., 2007; Ottoni et al., 2013
	Death	
Informal Language	Total Informal Language*	e.g., Rao et al., 2010
	Swear Words*	e.g., Argamon et al., 2007; Fullwood et al., 2001; Ottoni et al., 2013; Savicki et al., 1996 (coarse

		and abusive words); Schwartz et al., 2013; Thelwall, 2008; Witmer & Katzman, 1997
	Assent*	e.g., Guiller & Durndell, 2007
	Nonfluencies*	e.g., Rao et al., 2010 (disfluencies)
	Fillers	
Punctuation	Question Marks*	e.g., Savicki et al., 1996 (questions)
	Exclamation Marks*	e.g., Rao et al., 2010; Waseleski, 2006
	Quotation Marks	
Emoticons	Total Emoticons*	e.g., Baron, 2004; Burger et al., 2011; Fullwood et al., 2001; Huffaker & Calvert, 2005; Rao et al., 2010; Witmer and Katzman, 1997; Wolf, 2000
	Positive emoticons*	
	Negative Emoticons*	

Note. Variables with an asterisk (*) been identified as gender markers in previous CMD literature.

4.5. Statistical Methods

Excepting the Word Count and Words Per Sentence, as well as the four summary variables (Analytical Thinking, Clout, Authenticity, and Emotional Tone), all the other *LIWC2015* variables underwent a series of statistical transformations in order to normalize the data for calculating the mean frequencies and using a linear regression model. There were several reasons for the statistical transformations described in this section. On the one hand, in the case of the absence of a variable from a record, the *LIWC2015* software assigns a 0 value for that variable; as a result, many of the variables had a significant number of zero values. On the other hand, in the case of two- or three-word records (e.g., "Me too!" or "I completely agree"), certain records would have very high values for certain variables (for example, 50% for personal pronouns in "Me too!" or "I completely agree"); this resulted in extreme values on the other end of the distribution tail. Moreover, the data were not symmetrically distributed across genders, time points, or newsgroups, due to the participation differences of males and females in newsgroups and the popularity of newsgroups at different time points. Consequently, the data had to be normalized in order to calculate means that are representative of the data, as well as to meet the requirements of the linear regression model.

4.5.1. R and Packages

All statistical tests in the study were conducted using *The R Project for Statistical Computing* (R Core Team, 2013), versions 3.4.3 and 3.4.4. Specifically, the software used with the *R* statistical computing language was *RStudio* (R Studio team, 2016), version 1.1.442. The "linear model" *lm()* function, version 3.4.4, from the *stats* R package (The R Stats Package, 2018) was used for creating a model for each variable that evaluates the significance of identified gender changes over time. The *ggplot2* (Wickham, 2009) package was used to calculate and plot the means of each variable for men and women over time to identify any changes in gender patterns; it was also used for creating plots that present trend lines for each variable by gender.

4.5.2. Descriptive Statistics

A first step to normalize the data was calculating the means of the variables by newsgroup, gender, and time point. This helped resolve the imbalance in the data distribution and address the possible language variation among newsgroups. Moreover, the number of data points or observations was decreased: Instead of approximately 4 million observations in the dataset (individual posts), the observations were decreased to 639 (combined posts per gender, newsgroup, and time point), which helped avoid any false identification of changes in gender patterns as significant by the models described in the next section, due to the size of the data used in the study.

Some variables had a normal distribution after this process; consequently the mean frequencies per gender and time point were directly calculated from the newsgroup means and plotted into graphs. However, a number of the variables were still skewed after the first step of the transformation. Those variables underwent a log transformation to further normalize the data, after which the mean frequencies per gender and time point were

calculated from the newsgroup means. The values were then transformed back to the original scale using an exponential transformation and were plotted into graphs.

4.5.3. Simple Linear Regression

In order to evaluate the significance of identified changes in gender patterns, a linear regression model was used for 72 variables. The data for the variables that were normalized in the first transformation step were used directly; the data for the rest of the variables underwent a log transformation, similarly to the calculation of the mean frequencies. The following expression was used to model the difference in the female and male frequencies for variables over time:

$$Variable = Intercept + Gender \times \beta_1 + Time \times \beta_2 + Gender \times Time \times \beta_3 + error$$

The model accounted for both gender and time (expressed in time points at 4-year intervals). The residuals for all of the models created had a normal distribution; consequently, it was decided that the model was appropriate for the study.

```
Residuals:
Personal Pronouns %
      Min        1Q    Median        3Q       Max
-0.061646 -0.009412  0.000112  0.008901  0.061862

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.0612666  0.0020148  30.408  < 2e-16 ***
GenderM           -0.0080949  0.0027315  -2.964  0.00316 **
Timepoint          0.0000948  0.0004400   0.215  0.82950
GenderM:Timepoint  0.0002716  0.0006031   0.450  0.65265
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01496 on 635 degrees of freedom
Multiple R-squared:  0.05366,   Adjusted R-squared:  0.04919
F-statistic:    12 on 3 and 635 DF,  p-value: 1.189e-07
```

***Figure 5*. Simple Linear Regression model output for personal pronouns.**

In the model output in Figure 5 for the Personal Pronouns variable, the women (F) were used as the baseline for the model; consequently, *GenderM* was the main gender effect difference in Personal Pronoun use for males, as compared to the baseline of females. *Timepoint* refers to the impact of time on the use of personal Pronouns by females (the baseline). *GenderM:Timepoint* is the value of interest for evaluating the significance of any change in the use of Personal Pronouns by women and men over time: It represents the difference between the use of the variable by females (the baseline) and males as time increased. Thus, the coefficient for the male use over time would be calculated as *Timepoint* + *GenderM:Timepoint.* It should be noted that the results of the model for the variables that had to undergo a log transformation were not on the same scale as the variables that were used without any transformation; the former are noted with a cross (+) in the tables presented in the Results section (5.2.).

Finally, in order to facilitate the evaluation and visual interpretation of the directionality of change in gender patterns over time, as well as to identify possible longitudinal trends, trend lines for each gender were plotted for all the variables that underwent linear regression analysis. The trend lines were created by calculating the predicted values for each variable, using the associated model and the original dataset. It should be noted that the variables with the log transformation underwent an exponential transformation to convert them back to their original percentage scale before being plotted, in order to further facilitate the interpretation of the graphs.

Results

5.1. Change Over Time in Overall Usage

This section presents the overall change over time in the *LIWC2015* variables and the additional Emoticon categories analyzed in this study. The mean frequencies for each variable are presented per time point. The figures for the variables are grouped into the 13 categories listed in Table 9 in section 4.4.

Figure 6 presents the summary language variables analyzed in the study. The number of words per post decreased over time until approximately 2000 but shows a sharp increase after that time point. However, the length of sentences decreased earlier in the time period studied (1990-1991), with a less pronounced increase after 2000. A closer look at the findings in Figures 19 and 20 of section 5.2 uncovers salient details: It was male posts and sentences that caused this dramatic increase after 2000.

The number of words with 6 letters or more shows fluctuation over time, between 18% and 19.4% of the total words in each post. The percentage of words identified in the *LIWC2015* internal dictionaries increased to 75% around 1990, but starts decreasing over time to 69% of the total words per post. This may be an indication of more non-standard forms appearing in the language of users over time, as supported by the findings in Figure 16 showing that language use becomes more informal over time.

The Analytical Thinking variable is a "factor-analytically derived dimension based on eight function word dimensions" that measures the degree of "formal, logical, and hierarchical thinking" in text (LIWC, 2017). While there is fluctuation over time, the use of words indicating analytical thinking drops overall from 72% to 70%. The summary variable of Authenticity measures the degree of honesty in the text, based on a series of previous studies on honesty and deception (LIWC, 2017). Participants in USENET newsgroups did not score high in authenticity overall, and became even less authentic over time, with a

striking drop from 37% to 29%. Their score for the summary variable Clout, the relative social status, confidence, and leadership displayed in writing, was moderate but increased steadily over time. Finally, the summary variable Emotional Tone, where higher scores indicate more positive tone (LIWC, 2017), shows an interesting change: According to the scores, the language in USENET newsgroups was more negative at the beginning of the time period studied, but became more positive until the middle of the 1990s, when it started to become more negative again.

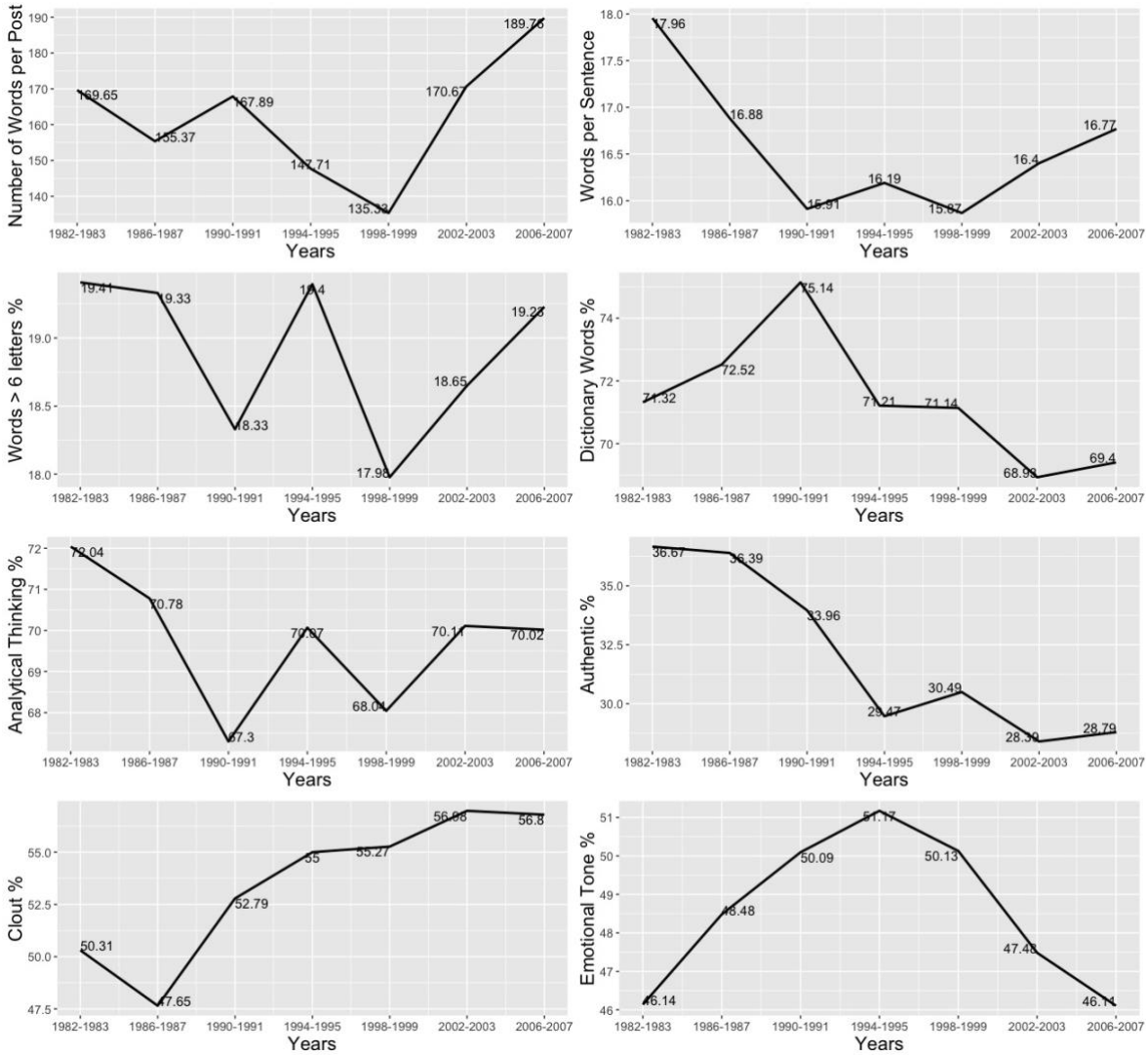


***Figure 6*. Overall change in Summary Language Variables over time.**

As seen in Figure 7, the overall use of function words decreased over time. Articles, prepositions, and common adverbs showed similar changes: a slight increase around 1990 and a decrease after that time point to lower frequencies of use. Adjectives had a sharp increase until 1990, when their use started fluctuating and slightly decreased. Common verbs, however, showed a steady increase in use over time.

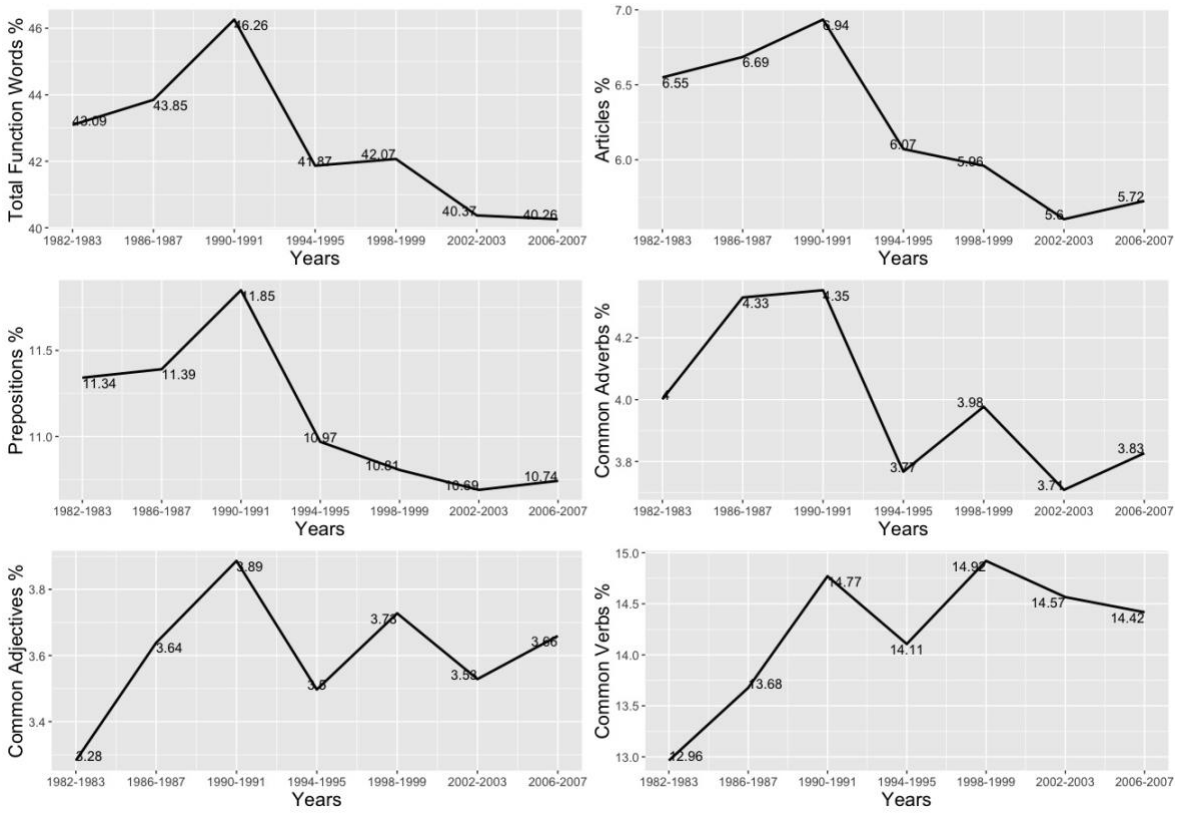


***Figure 7*. Overall change in Common Grammatical Features (part 1) over time.**

Conjunctions, quantifiers, and auxiliary verbs in Figure 8 presented similar patterns of evolution to dictionary words, articles, and prepositions: an increase around 1990 and a slow decrease after that time point. In contrast, the use of negation increased over time, with one steep drop in 1994-1995. Comparison words also showed an increase in use by the end of the time period studied, peaking around 1990 and decreasing slightly after that time point. Interrogatives had the highest frequency at the first time point and a steep decrease after that. Since most of the newsgroups were created at that time, understanding and coordinating the use of the newsgroups would have included more questions in interactions, which decreased

after the users became familiar with the new environment. However, the frequency of interrogatives increased again around 1990-1991, possibly due to the influx of new users as Internet access started becoming more widespread, slowly decreasing again when the new users became familiar with the platform. Finally, numbers showed an interesting change in frequency: Their use dropped dramatically in 1990-1991 and started rising again until the end of the time period studied.

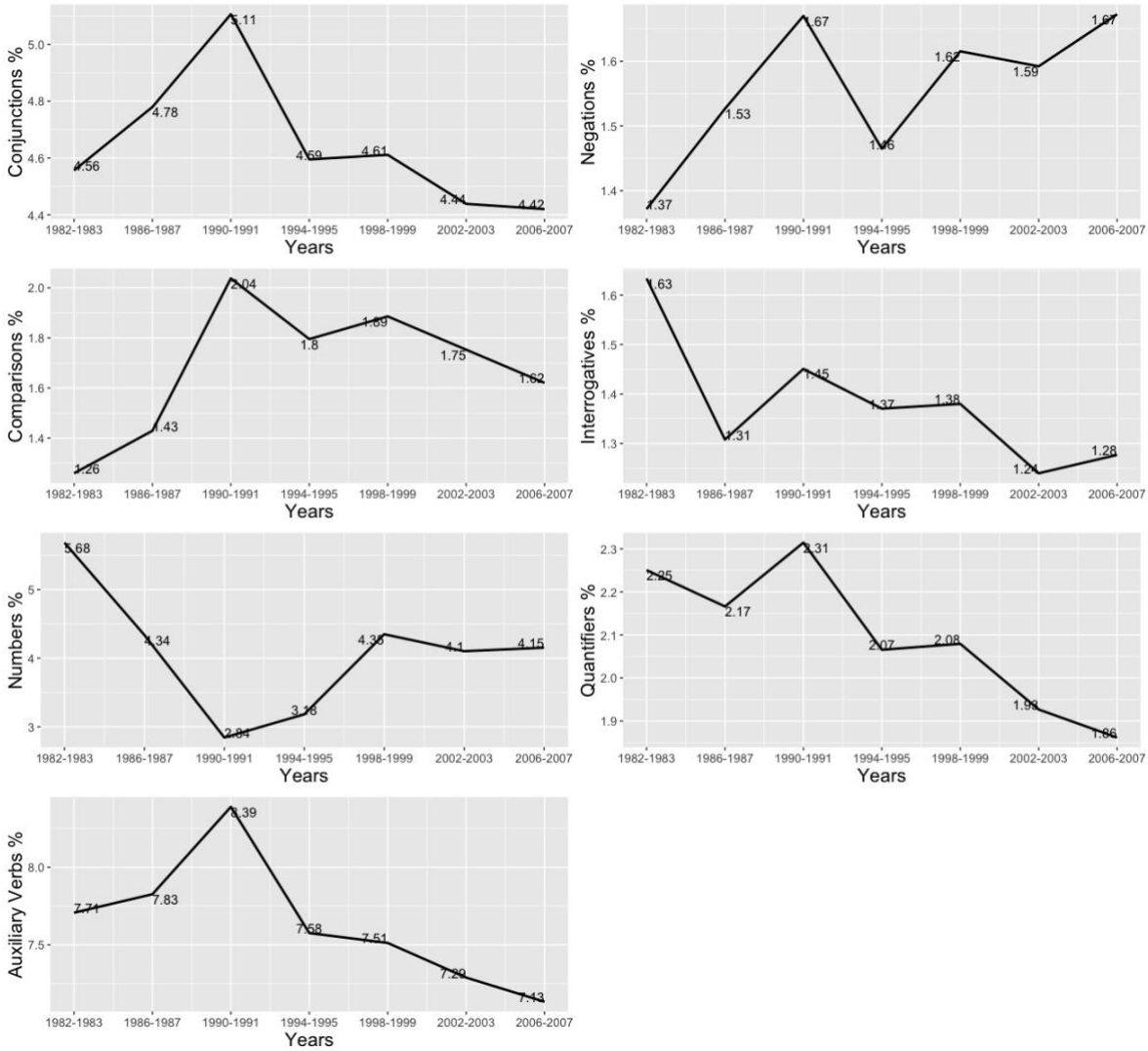


***Figure 8*. Overall change in Common Grammatical Features (part 2) over time.**

The evolution of pronoun use over time was not the same for all pronoun types (Figure 9). Overall pronoun frequency peaked around 1990-1991 but slowly decreased after that time point; a similar pattern is evident for impersonal pronouns. However, the use of

personal pronouns increased over time, with the exception of the 1st person singular pronoun and the 3rd person plural pronoun. The 2nd person pronoun and 1st person plural pronoun exhibited very similar patterns of change over time, while the 3rd person singular pronoun appears to peak in use later, around 2002-2003.

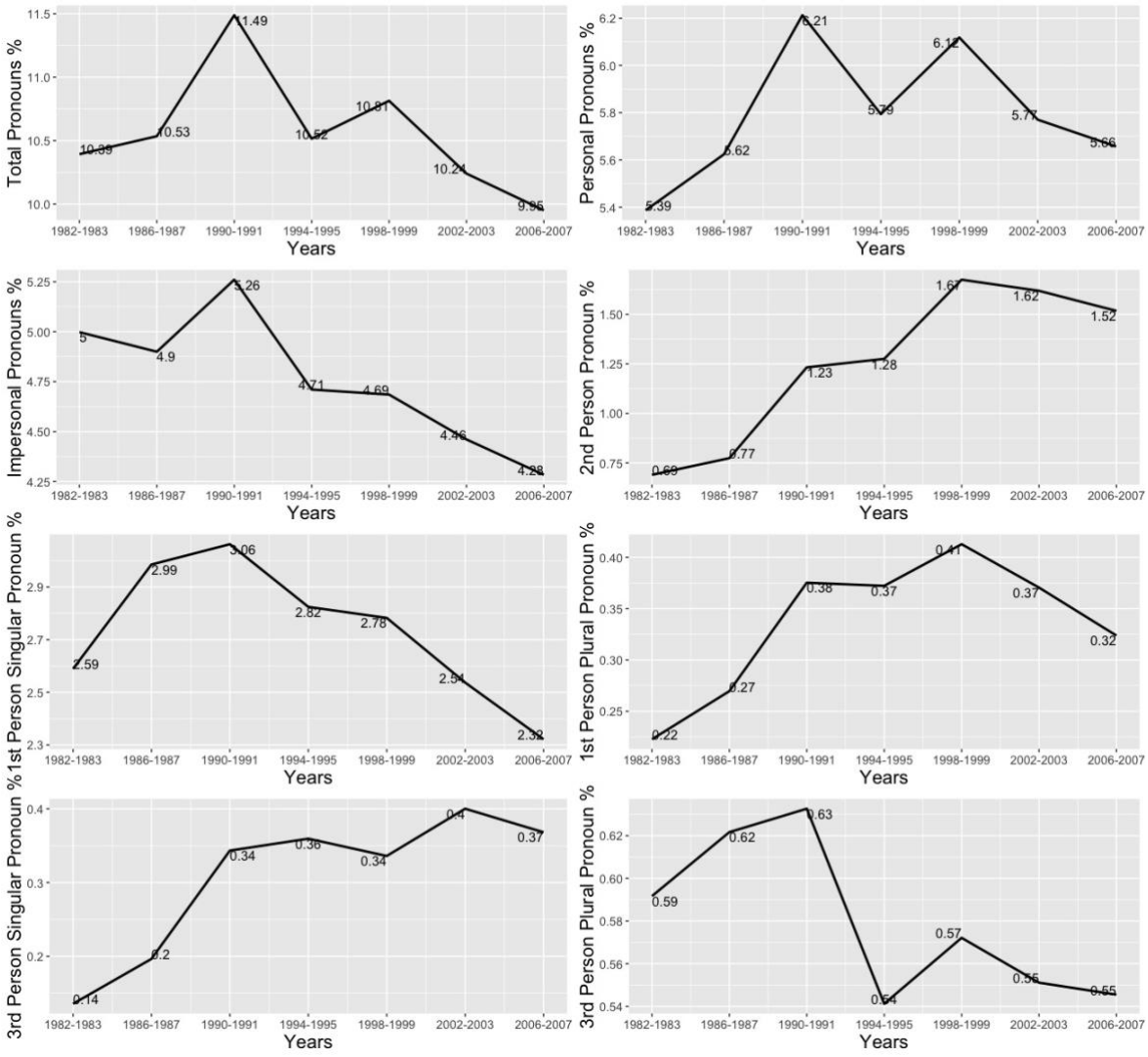


***Figure 9*. Overall change in Pronouns over time.**

The expression of emotion increased over time overall, as shown in Figure 10. Positive emotion had higher frequencies than the rest of the affective processes in the category. Anger, anxiety, sadness, and negative emotion overall exhibited frequency peaks at 1990-1991 and 2002-2003.

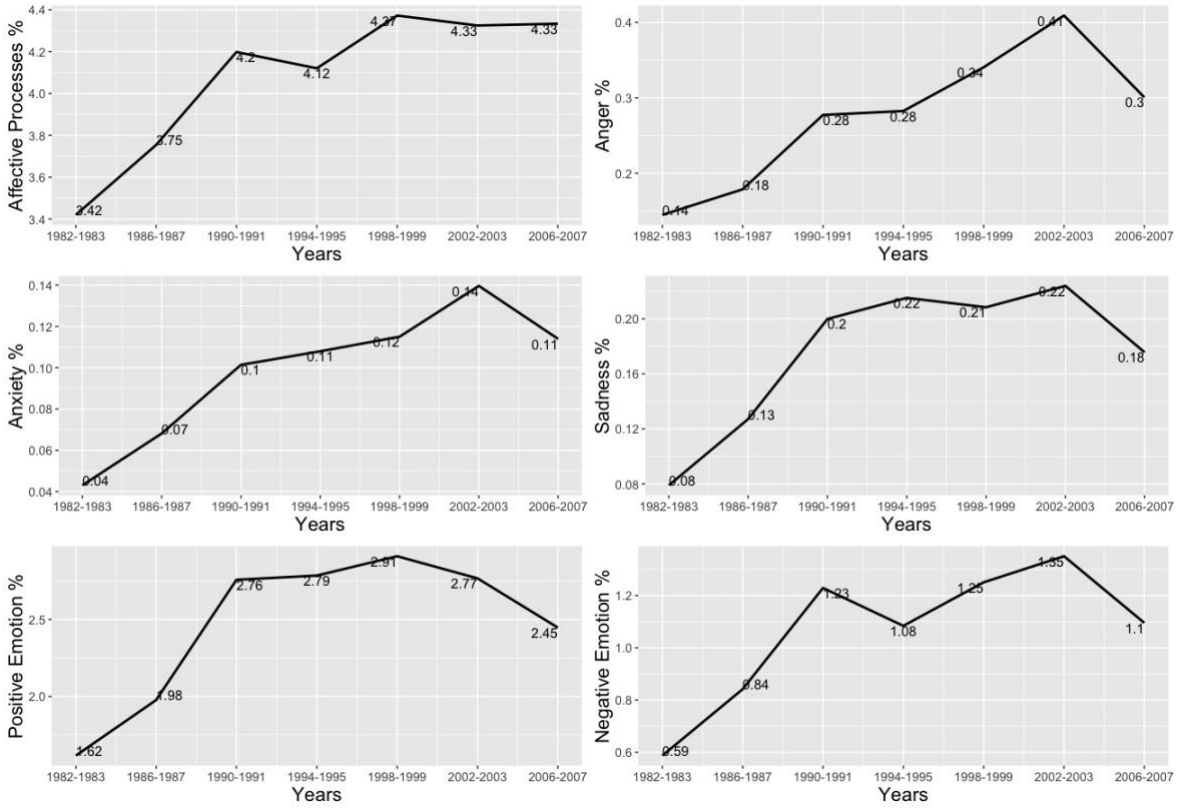


***Figure 10*. Overall change in Affective Processes over time.**

As can be seen in Figure 11, not all social processes followed the same evolution over time. There was a sharp increase in overall social processes around 1990, as well as specifically in the use of words referring to friends and the use of male references, both of which peaked around 1995 in frequency. However, the use of female references presented different changes over time: Similarly to other social processes in this category, the frequency of female references picked around 1990; however, it dropped dramatically after that time, until 2000 when it started increasing again with another peak by the end of the time period studied. The use of words related to family was the only variable in this category that showed an overall decrease in frequency over time.

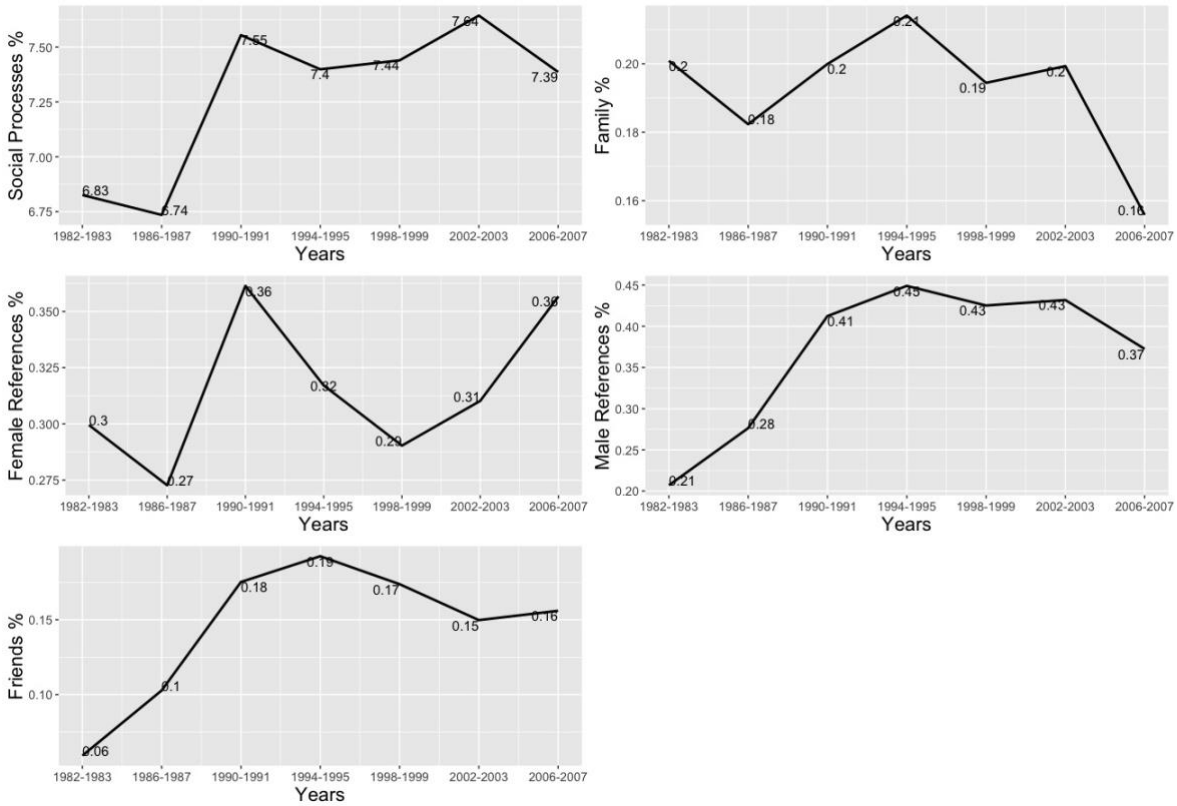


***Figure 11*. Overall change in Social Processes over time.**

Figure 12 presents the changes in words expressing cognitive processes over time. Overall, the expression of cognitive processes decreased over time. Insight, discrepancy, and tentativeness evolved in a similar way, peaking around 1990 and dropping after that time in frequency. Words expressing differentiation did not show any peaks but steadily decreased in use over time, whereas causation decreased over time with peaks around 1990, 2000, and 2005. In contrast to the rest of the cognitive processes, certainty is the only variable in the category that increased in frequency over time, with a dramatic peak around 1990; taking a closer look at the data, we can see that the peak is caused by the sudden rise of words expressing certainty in the posts of women, since the male posts had consistently higher frequency over time (Figure 20).

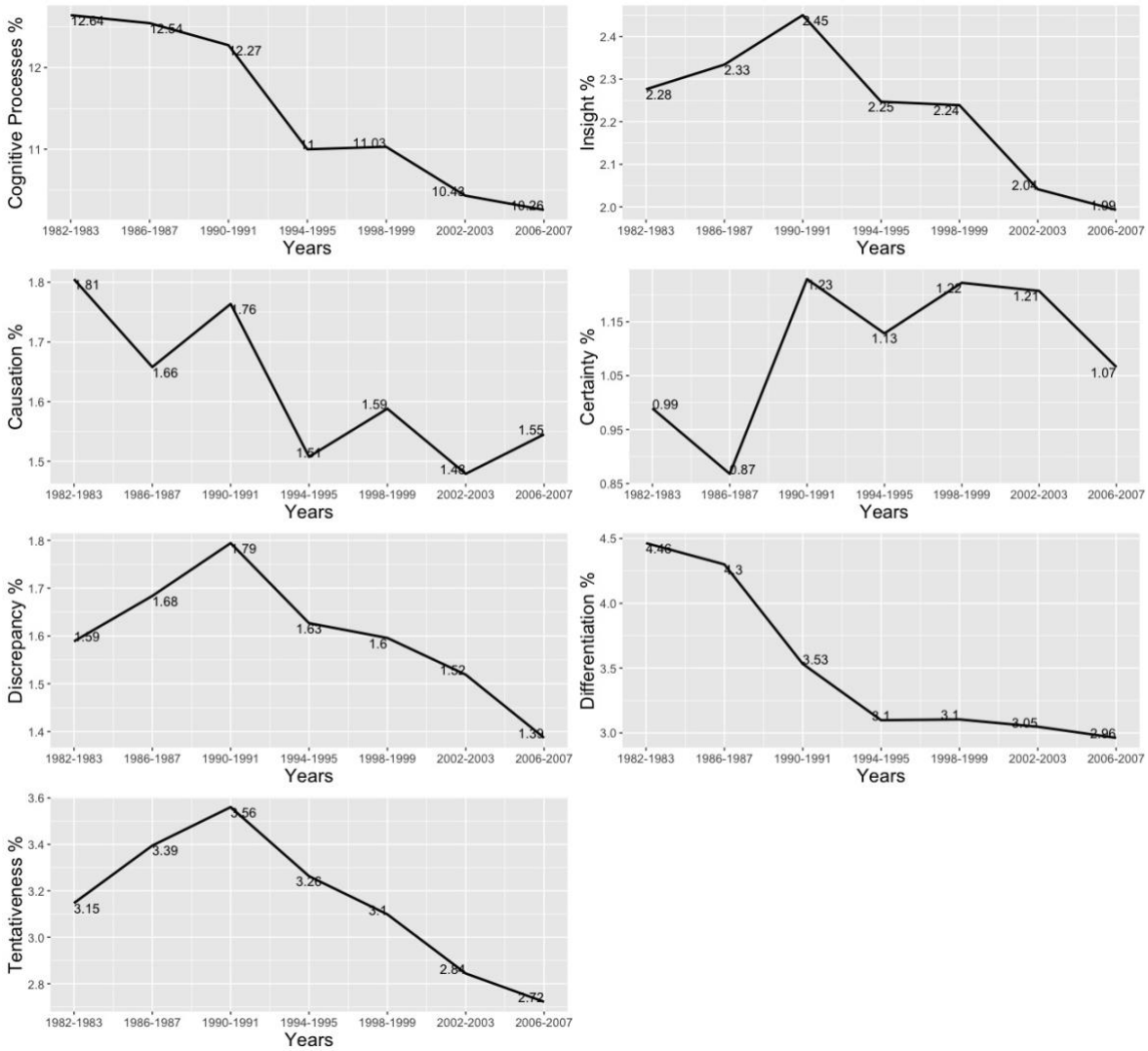


***Figure 12*. Overall change in Cognitive Processes over time.**

All perceptual processes except for the expression of touch (Feel) in Figure 13 showed a similar evolution over time: There was an abrupt peak around 1990 and fairly steady use after that for the remainder of the time period studied. However, although following the abrupt peak around 1990 like the rest of the variables in the category, words expressing the perceptual process of feeling had their highest peak after 2000, dropping in frequency after that time.

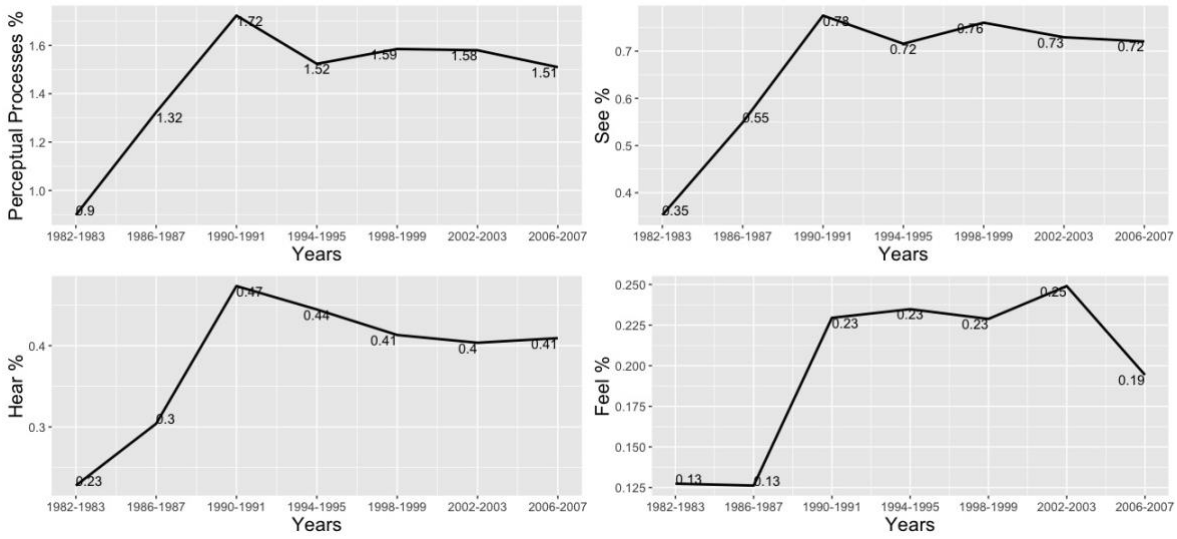


***Figure 13*. Overall change in Perceptual Processes over time.**

Most biological processes in Figure 14 exhibited similar changes. Overall, words for biological processes and health peaked in frequency around 1990 and maintained a fairly steady use after that time point. Words related to the body and ingestion increased in frequency over time and peaked around 2000, but ingestion dropped in frequency for the remainder of the time period studied. Sexual words were the only variable in the category that showed a different evolution: The frequency of sexual words increased steadily until 1994-1995 but dropped after that time point.

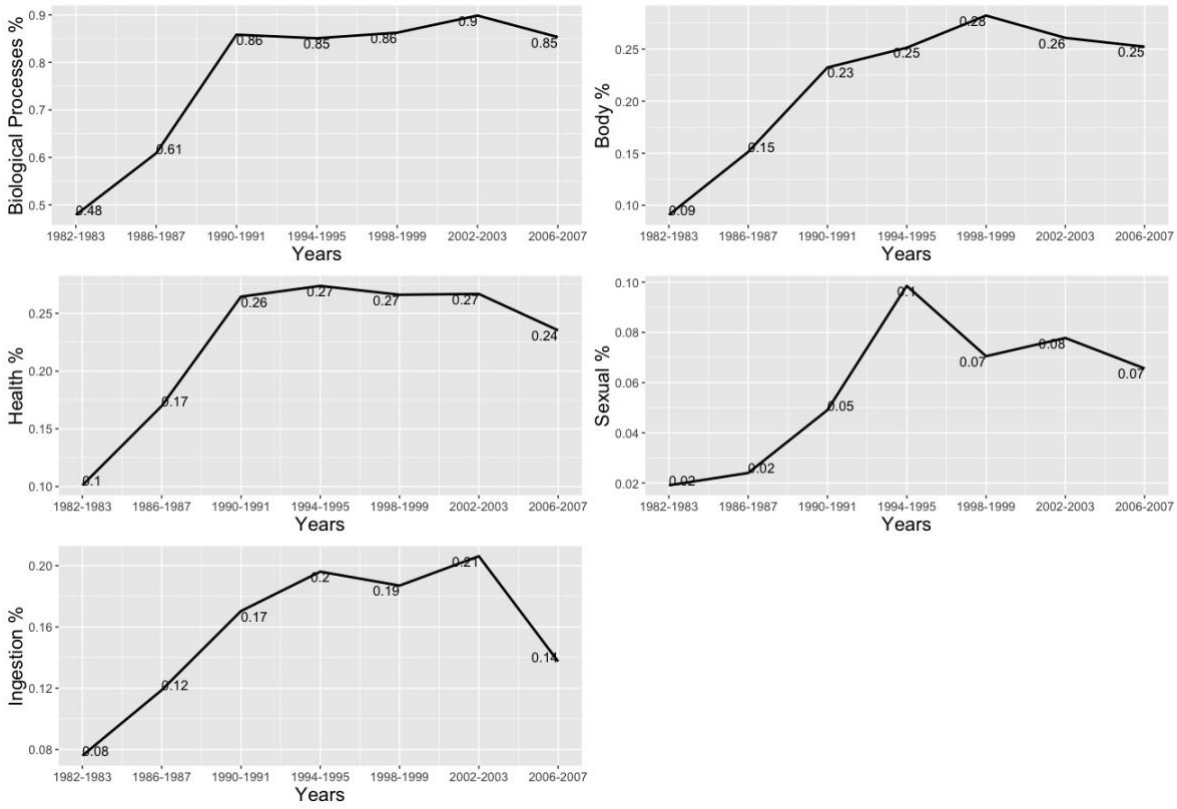


***Figure 14*. Overall change in Biological Processes over time.**

All drives except affiliation (Figure 15) exhibited an abrupt peak in frequency around 1990 in the dataset, like other variables analyzed in the study, and showed an overall increase in their expression through language. Achievement had a fairly steady expression as a drive in the users' posts after the 1990 peak mentioned above, while reward dropped after 2000 and risk dropped a little later around 2002. Power words increased steadily over time, with a slight drop around 2000. Affiliation was the only drive that peaked in frequency much later around 2002.

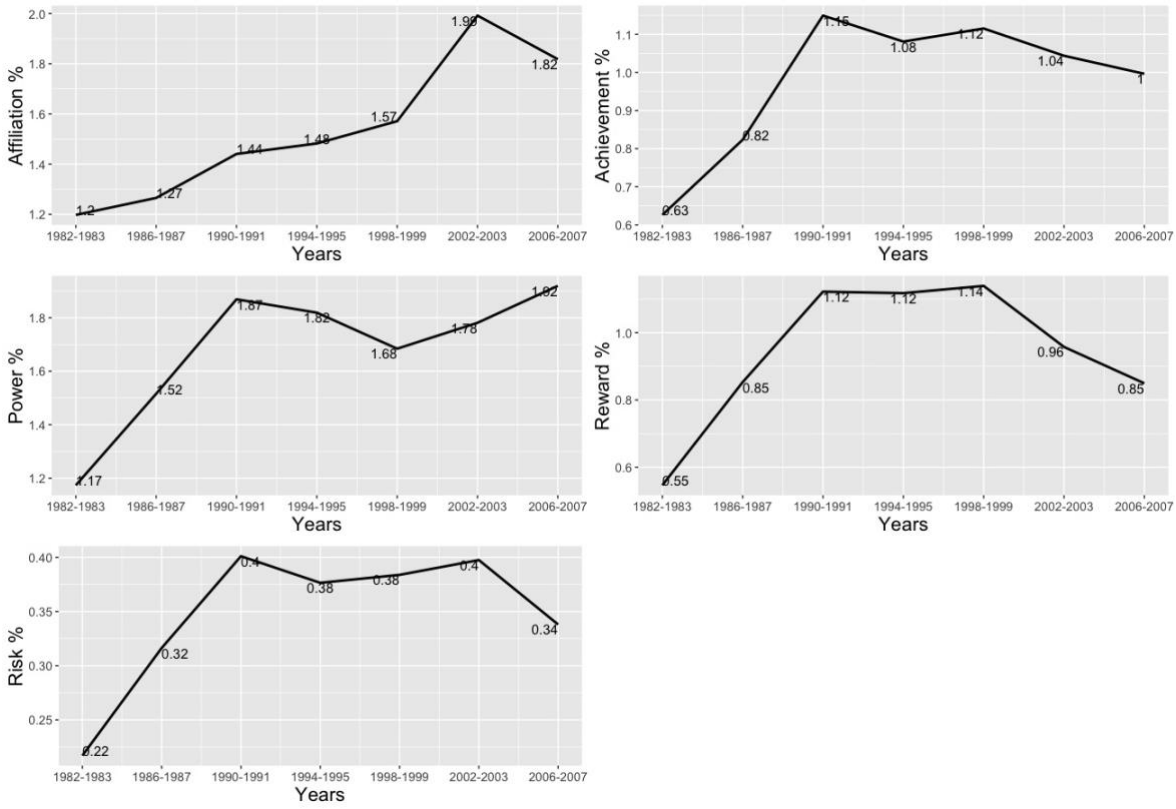


***Figure 15*. Overall change in Drives over time.**

Personal concerns in Figure 16 typically had peaks in frequency at a later time point in the dataset than most other variables. Work, leisure, home, and religion peaked in frequency around 1995, but exhibited different evolutions afterwards: Work kept increasing in frequency, leisure decreased, home dropped dramatically by the end of the time period studied, and religion had steady use until the end of the time period, when it decreased. The mention of money as a personal concern increased steadily over time until it peaked by the end of the time period studied, with an abrupt drop around 2002. Death showed a similar evolution to money until 2000, when it started dropping in frequency.

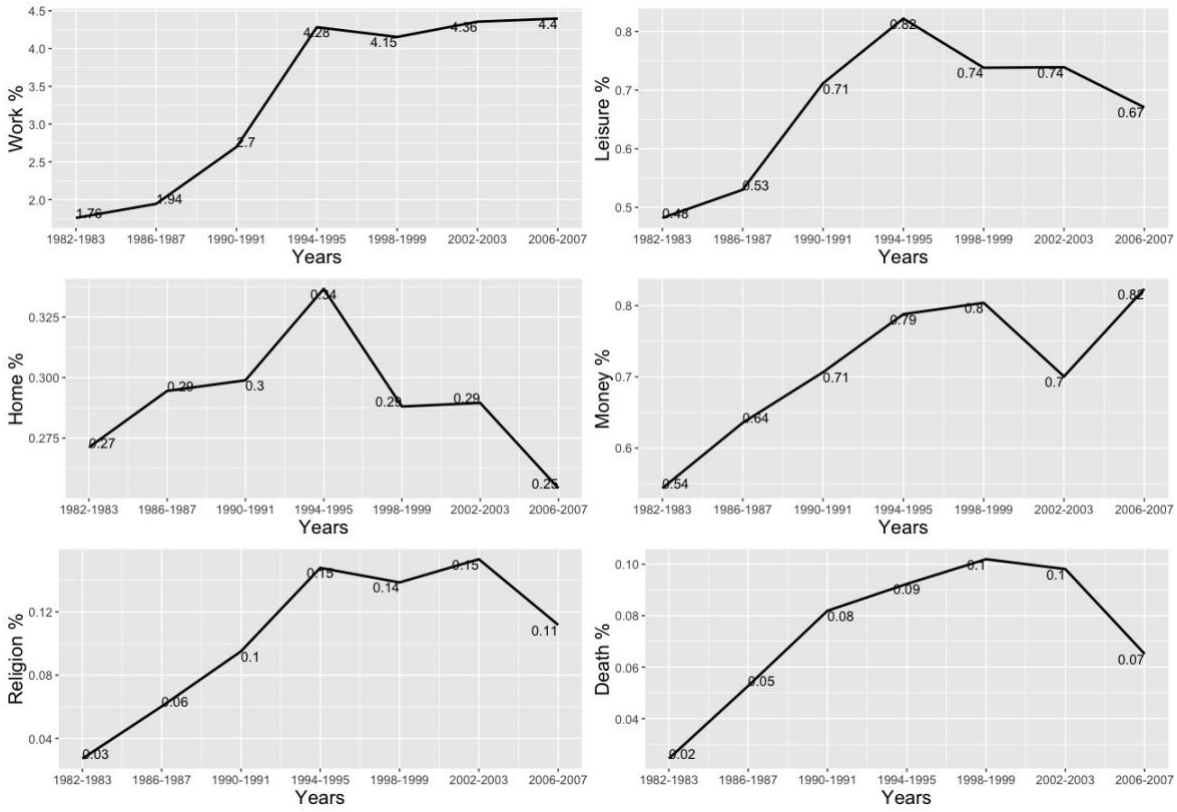


***Figure 16*. Overall change in Personal Concerns over time.**

Overall, language use in USENET newsgroups became more informal over time (see Figure 17), similarly to the findings of Herring in her MsgGroup study (1998). Interestingly, there is a drop in the use of informal language between the first time point in 1982-1983 and 1990, and a sudden rise after that. The expression of assent with words such as *woohoo* or *ok* or *yeah*, as well as the use of swear words, showed a steady increase in frequency over time. Nonfluencies such as *ahh* or *huh* or *mmm* also increased in use until 2000, when their frequency started dropping. Finally, fillers such as *anyway* or *blah* or *dunno* exhibited fluctuation over time.

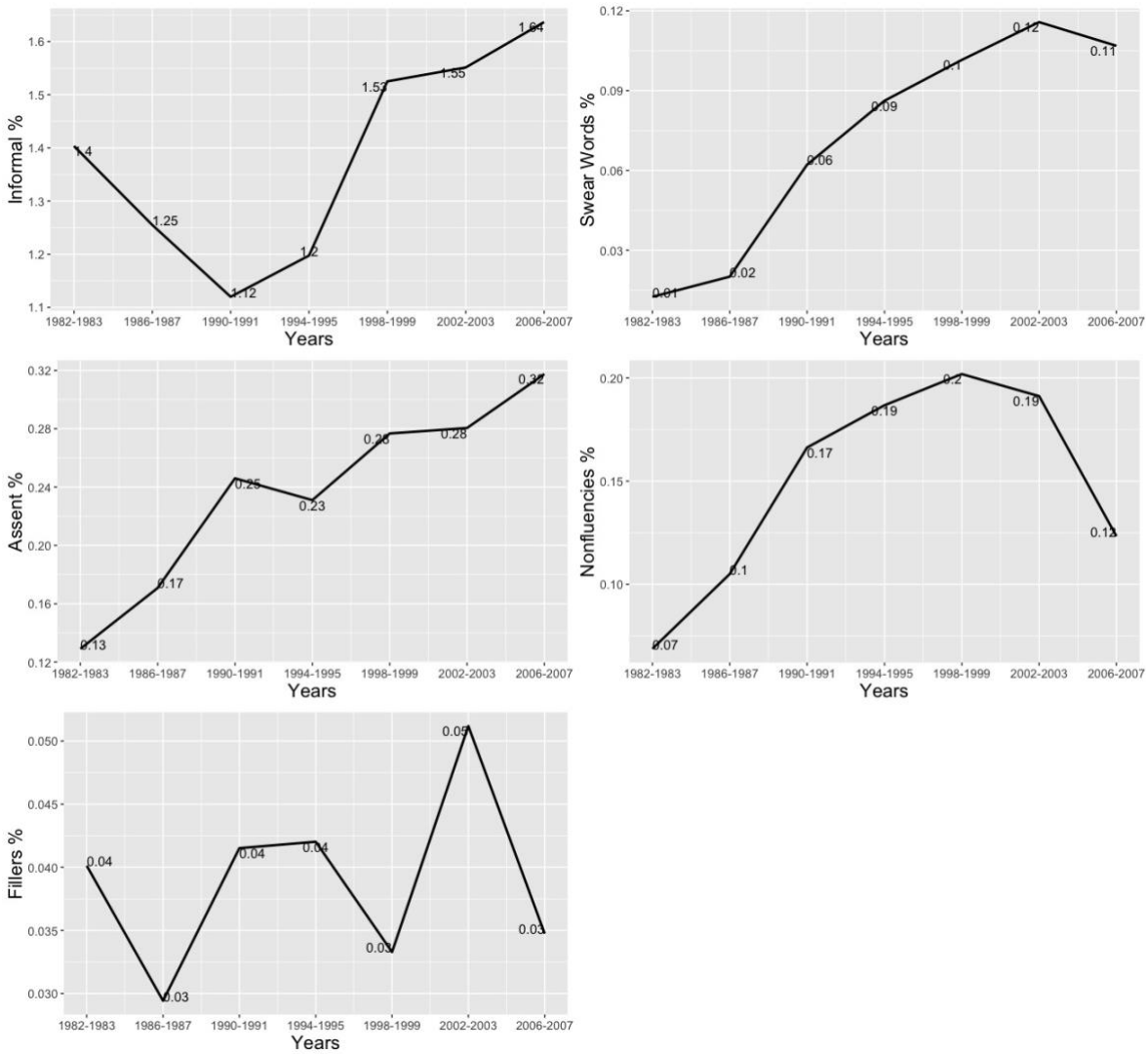


***Figure 17*. Overall change in Informal Language over time.**

The punctuation examined in this study showed different evolution patterns over time, as seen in Figure 18. There was an abrupt drop in the use of question marks around 1985, similar to the drop in interrogatives. However, the use of question marks increased dramatically until around 1995, when their use started decreasing again. The use of exclamation marks showed a dramatic drop until 1990, when their use increased again around 1995, but started dropping again afterward. Quotation marks were the only punctuation that exhibited an increase over time, peaking around 2002.

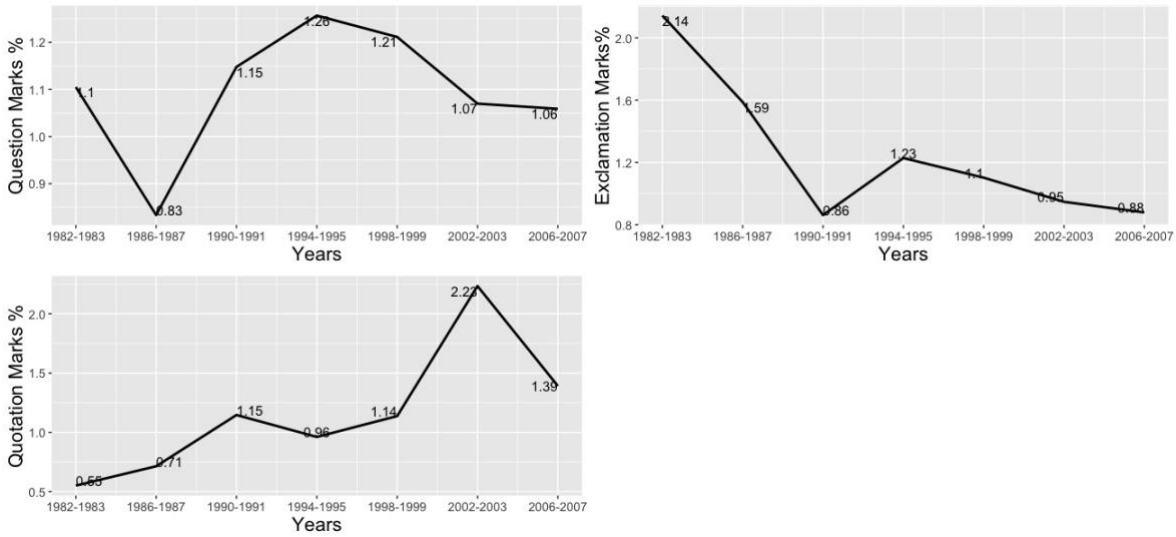


***Figure 18*. Overall change in Punctuation over time.**

Last, the use of emoticons overall – and positive emoticons, specifically – increased steadily over time, as seen in Figure 19. There were very few negative emoticons in the dataset; however, their use increased steadily until 2000, when they started decreasing in frequency.

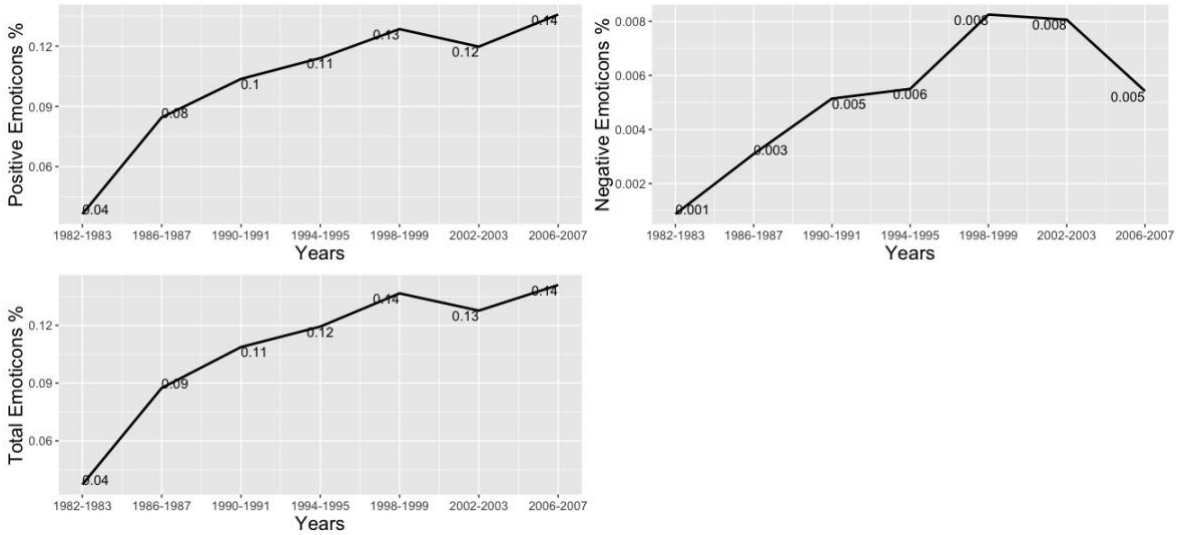


***Figure 19*. Overall change in Emoticons over time.**

5.2. Change Over Time in Gender Usage

This section presents the mean frequencies of the *LIWC2015* variables and the additional Emoticon categories analyzed in this study. Due to the number of variables, only

the language summary variables and variables identified as gender markers in previous research are included; the mean frequencies of the remaining variables are presented in Appendix D.

Men wrote longer posts than women during the time period studied, similarly to previous research (Herring, 1992a, 1992b). As seen in Figure 20, there was a decrease in post length for women over time, whereas the number of words in the posts of men increased rapidly after 2000, making the frequency difference between women and men even greater for this variable.

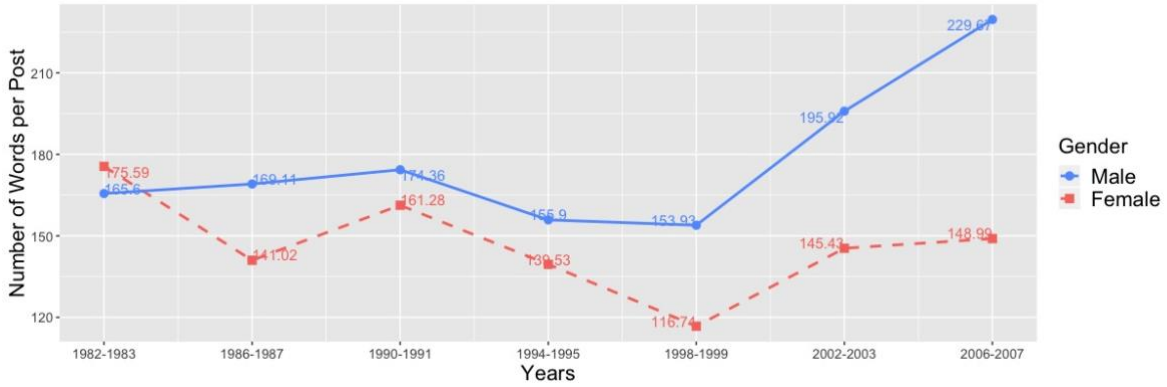


***Figure 20*. Average number of Words per Post by gender and time point.**

Regarding the average number of words per sentence, women used shorter sentences than men in the time period studied. Both genders showed a decrease in sentence length until 1990-1991, when the average numbers of words per sentence for each gender diverged further: Women's sentence length continued to decrease, while men started writing longer sentences (see Figure 21).

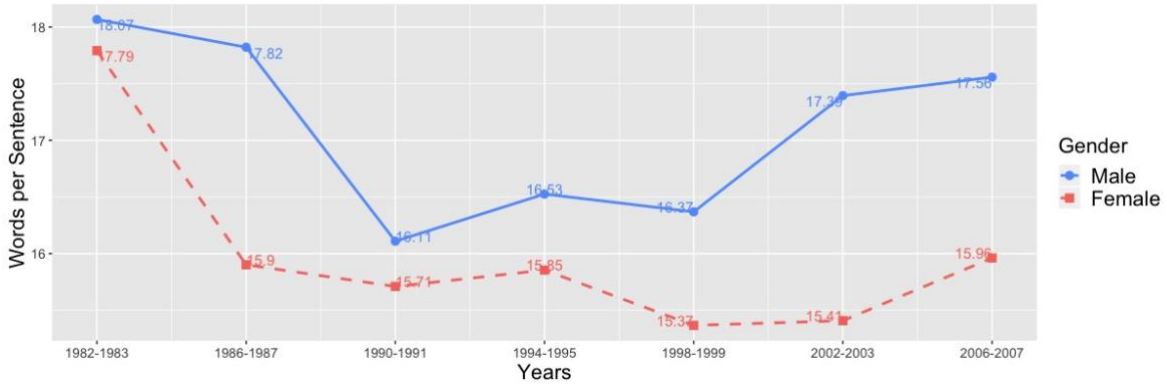


***Figure 21*. Average number of Words per Sentence by gender and time point.**

According to the LIWC website (2017), the Analytical Thinking variable is a "factor-analytically derived dimension based on eight function word dimensions" that indicates the use of words suggesting "formal, logical, and hierarchical thinking." As seen in Figure 22 below, the posts written by men used more words expressing analytical thinking, whereas women's posts used language that was more narrative and focused on personal experiences (LIWC, 2017).

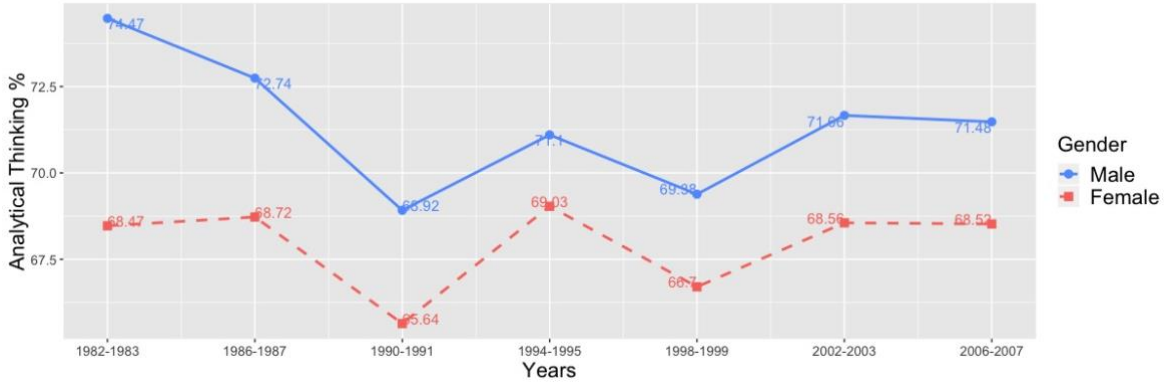


***Figure 22*. Percentage of Analytical Thinking by gender and time point.**

The variable Clout in Figure 23 refers to the "relative social status, confidence, or leadership that people display through their writing or talking" (LIWC, 2017). This summary variable is different from the variable Power presented below, which reflects a "need for power" (LIWC, 2017). While the scores for this variable increased over time for both

genders, women displayed slightly more confidence and leadership than men in the language they used in their posts.

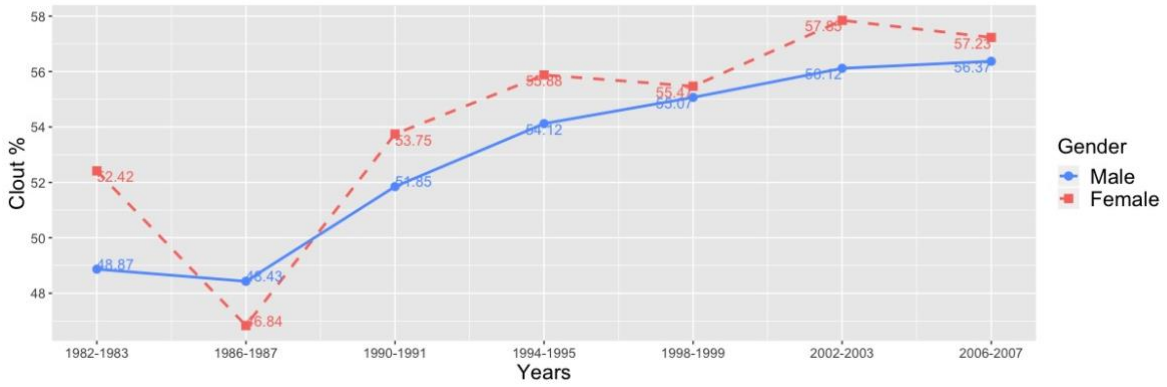


***Figure 23*. Percentage of Clout by gender and time point.**

However, even though the women's language in the data displayed leadership and confidence, it reflected less "attention to or awareness of relative status in a social setting" compared to the men's posts. As seen in the mean gender frequencies for the Power variable in Figure 24, men were more aware of their power status during the time period studied, whereas women's attention to power exhibited more fluctuation over time, peaking around 1990-1991 and after 2002.

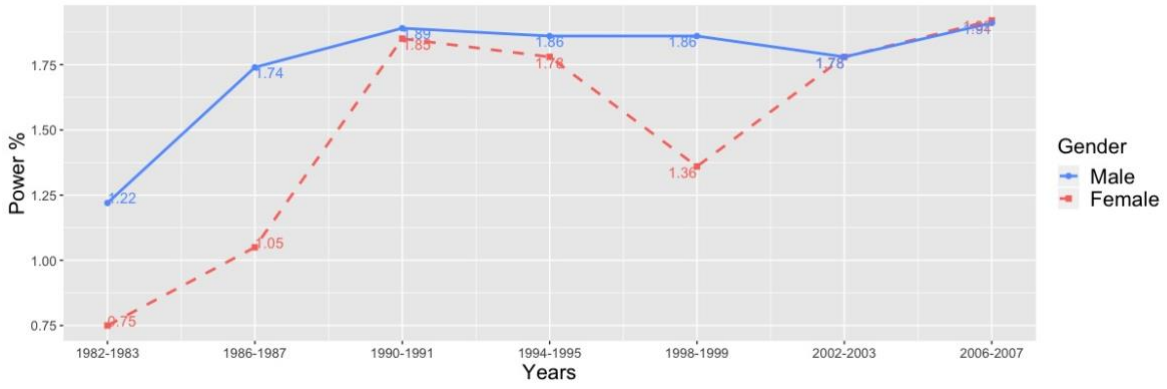


***Figure 24*. Percentage of Power by gender and time point.**

Even though neither women nor men scored high in authenticity, women used more language that is "personal, humble, and vulnerable" (LIWC, 2017). As seen in Figure 25 for the summary variable of Authenticity, the language of both genders could be interpreted as becoming less authentic over time.

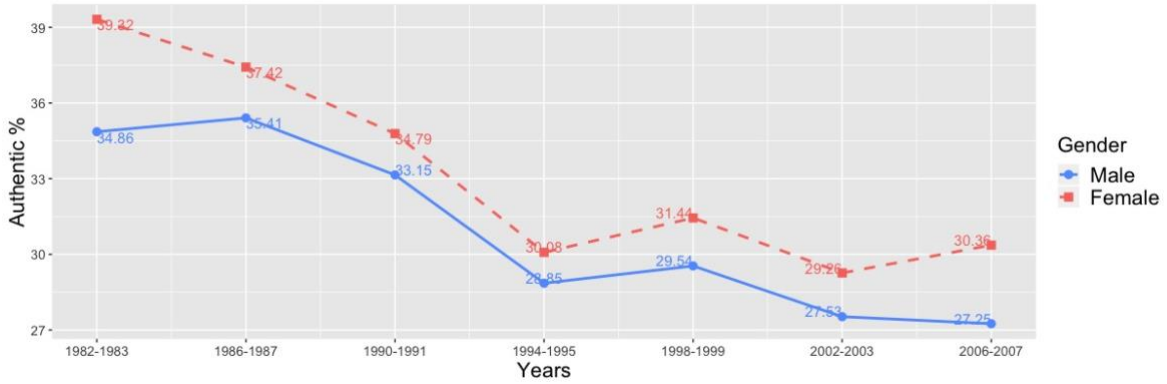


***Figure 25*. Percentage of Authenticity by gender and time point.**

The last summary variable, Emotional Tone, combines the Positive and Negative Emotion dimensions into a single summary variable; higher scores indicate more positive tone (LIWC, 2017). Figure 26 below shows that women's emotional tone was more positive than men's, especially between 1990 and 1999, slowly decreasing thereafter.

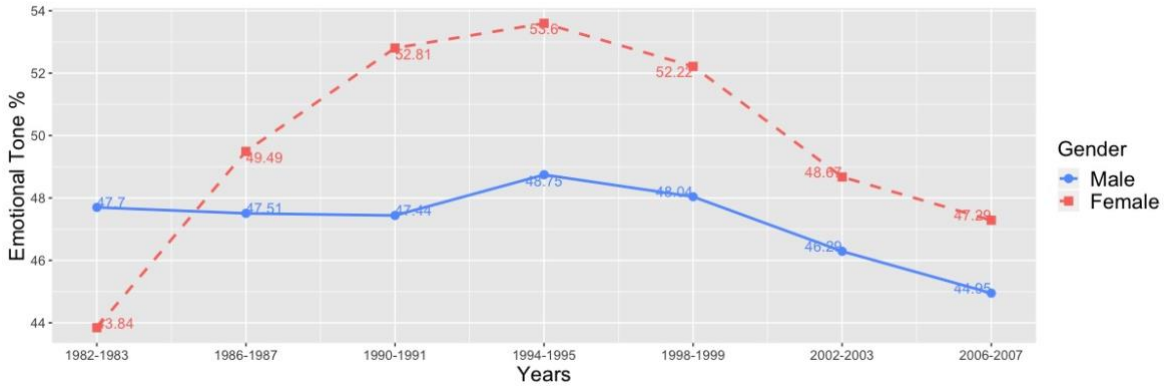


***Figure 26*. Percentage of Emotional Tone by gender and time point.**

Similarly, the variable for Positive Emotion shows a clear reversal of gender patterns over time. Even though women had lower frequencies of linguistic features indicating positive emotion at the beginning of the period studied, they expressed positive emotion with higher frequency after 1990, while men's frequencies slightly decreased after 1990 (see Figure 27).

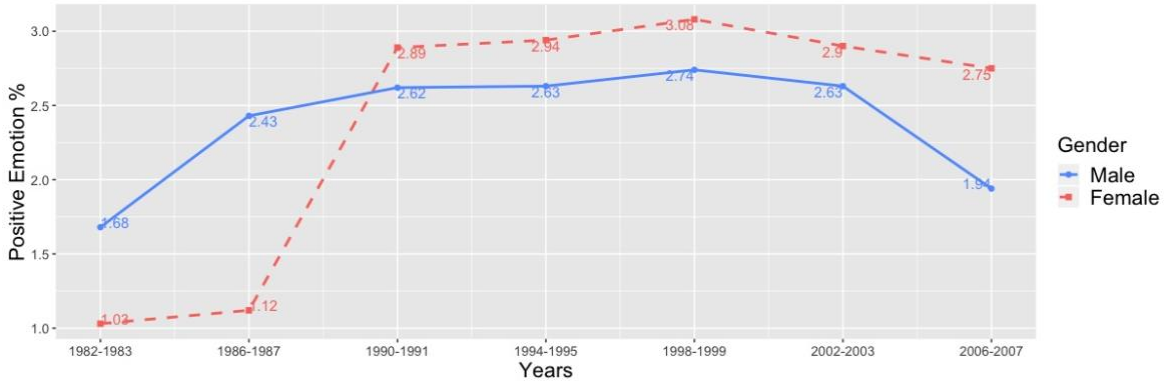


Figure 27. **Percentage of Positive Emotion by gender and time point.**

The variable for Negative Emotion in Figure 28 showed an interesting gender pattern over time, as well. Overall, women used language suggesting negative emotion in lower frequencies than men; however, there was an increase in frequency after 1987, until the women's negative emotion frequency surpassed that of men after 2002.

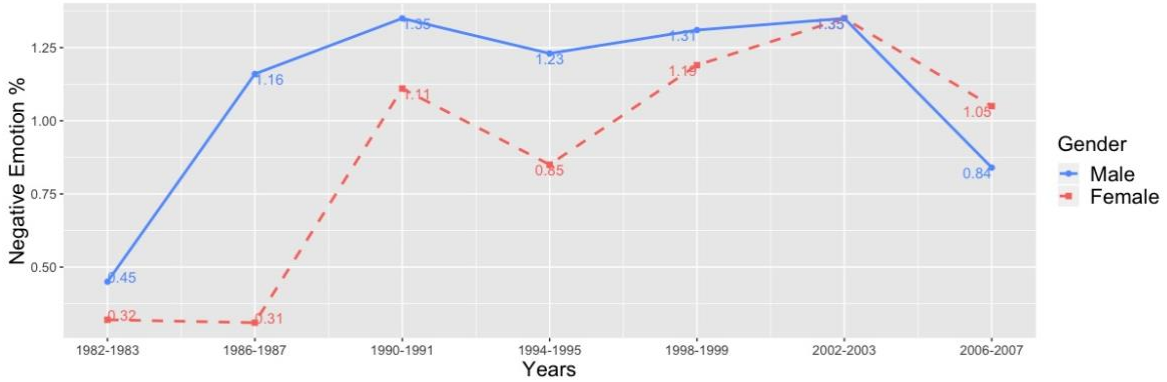


Figure 28. **Percentage of Negative Emotion by gender and time point.**

The variable Swear Words, a gender marker identified in previous studies (Argamon et al., 2007; Fullwood et al., 2001; Ottoni et al., 2013; Savicki et al., 1996; Schwartz et al., 2013; Thelwall, 2008; Witmer & Katzman, 1997), showed a particularly interesting pattern during the time period examined in this study. The frequency of swear words in posts increased over time for both genders. However, the increase was slightly higher for men, differentiating the use of this feature further between men and women (see Figure 29).

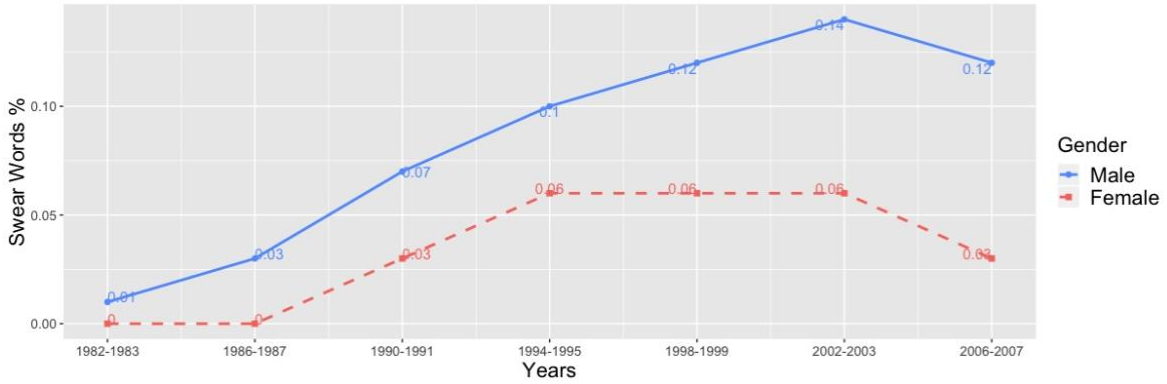


Figure 29. **Percentage of Swear Words by gender and time point.**

Another interesting variable studied in previous literature (e.g., Kapidzic & Herring, 2011; Ottoni et al., 2013; Subrahmanyan et al., 2006) was that of Sexual Words. While the use of sexual words showed a fairly steady increase for men over time, this feature had salient fluctuation after 1990 for women, as shown in Figure 30. There was an overall increase of sexual words in women's posts over time, which was similar to the frequency of men in 1994-1995 and even surpassed it in 2002-2003. However, the frequency of sexual words in women's posts was lower than that for men in the majority of the time points examined.

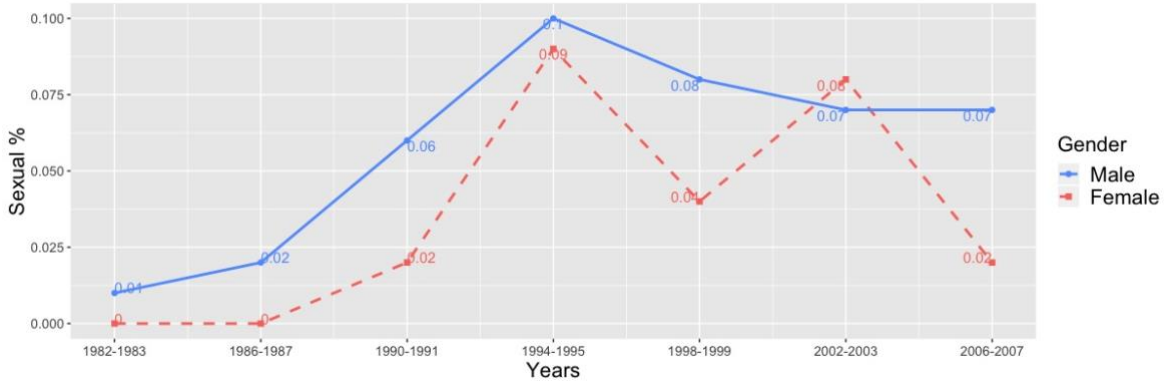


Figure 30. **Percentage of Sexual Words by gender and time point.**

Another feature examined by LIWC2015 that could be considered an important gender marker (Herring, 1993, 2010) is the frequency of male and female references in the posts analyzed. As seen in Figure 31, men made more references to other male entities than

women did; however, women showed a sharp increase after 1986-1987, until their use of male references leveled off and eventually became similar to that of men.

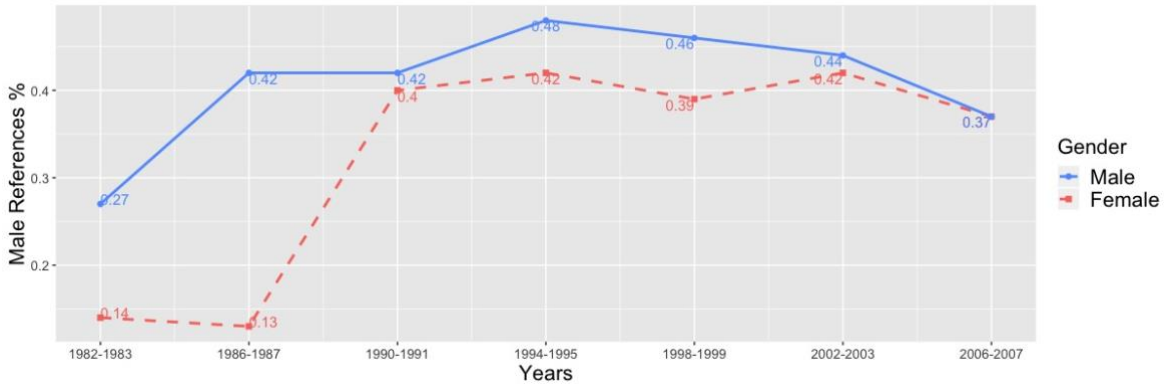


Figure 31. **Percentage of Male References by gender and time point.**

The use of female references, however, presented a very different pattern. Women made consistently more references to female entities over time, peaking around 1990-1991 and 2006-2007. In contrast, men's frequency of female references remained consistently low throughout the time period studied, without noticeable fluctuation (see Figure 32).

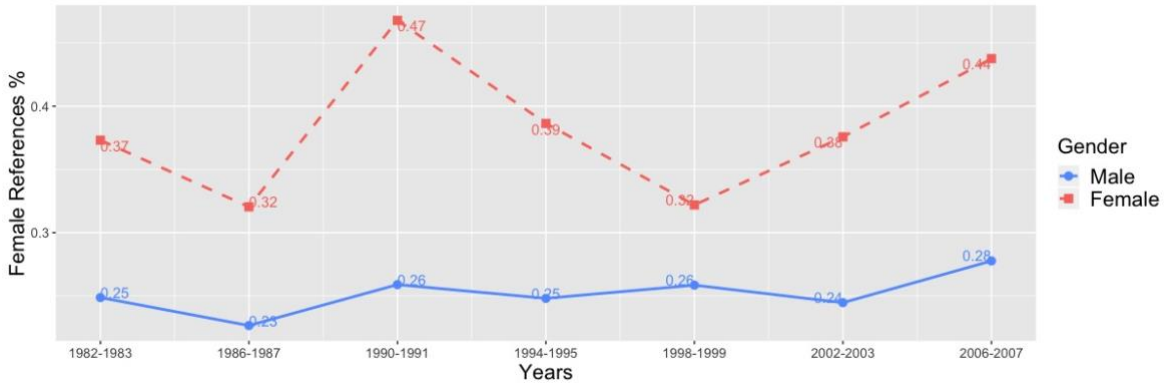


Figure 32. **Percentage of Female References by gender and time point.**

One of the gender markers identified in previous literature, Personal Pronouns (Argamon et al., 2007), exhibited a gender pattern that supports the findings of research in gender and language: Overall, women made much higher use of personal pronouns than men throughout the time period studied (see Figure 33).

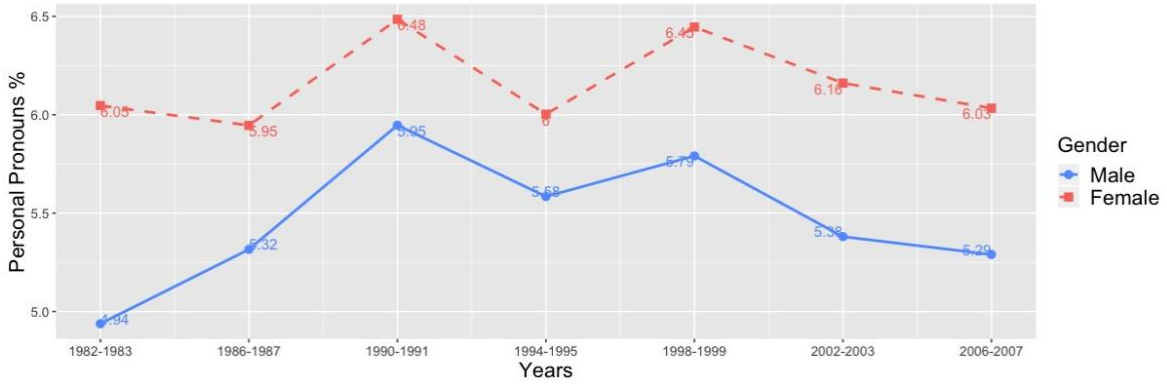


Figure 33. **Percentage of Personal Pronouns by gender and time point.**

A closer look at the different personal pronouns reveals that the patterns differ among them. While women made higher use of the first person singular pronoun, consistent with past literature (Ottoni et al., 2013; Schwartz et. al, 2013), men made higher use of the first person plural pronoun until women reached and surpassed the frequency of male use around 1994-1995. This comparison is shown in Figure 34.

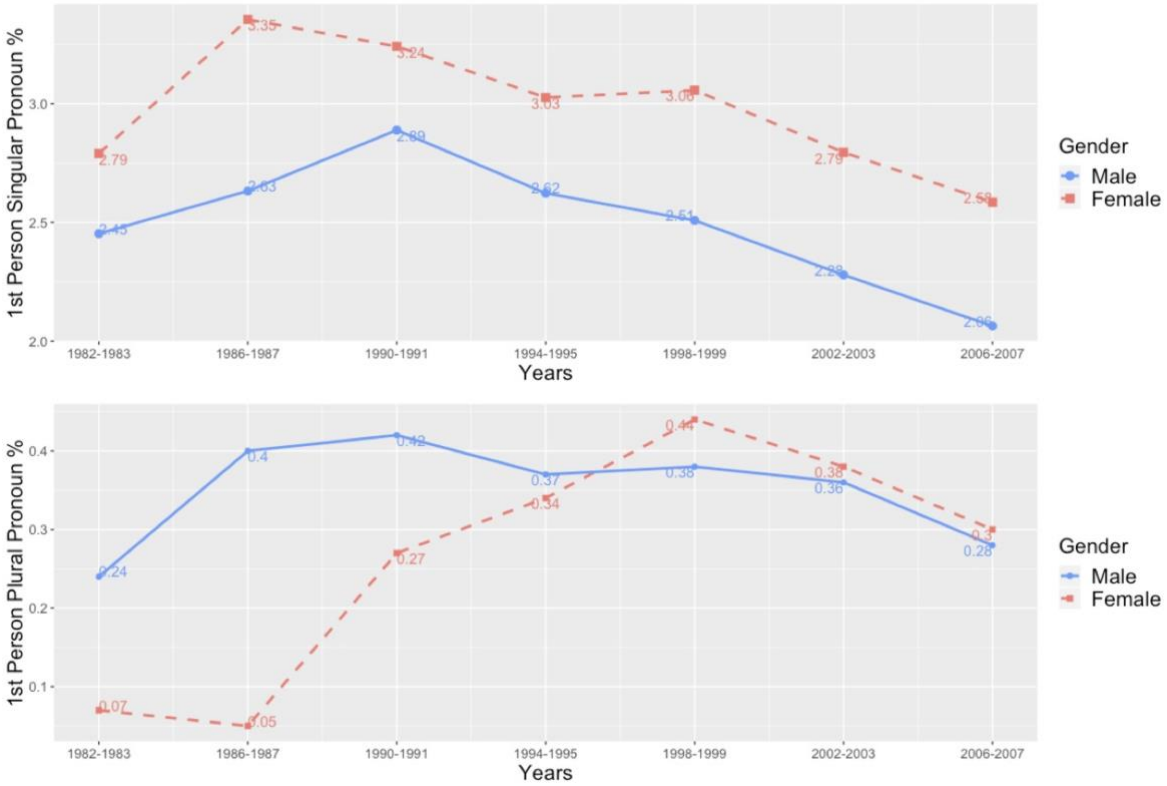


Figure 34. **Percentage of 1st Person Singular and 1st Person Plural Personal Pronouns by gender and time point.**

The second person pronoun showed some interesting changes over time, as well. Overall, the use of the pronoun increased; men used it with higher frequency than women until 1998-1999, when the use of this pronoun became similar for men and women (see Figure 35).

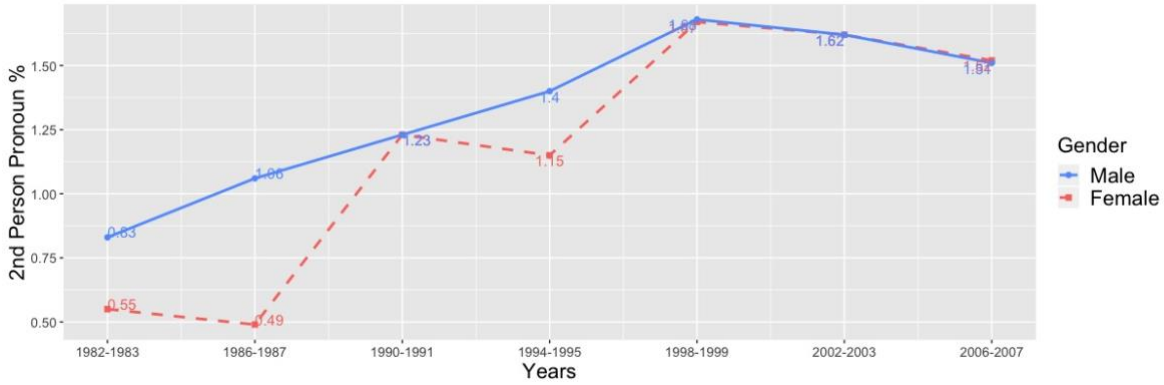


Figure 35. **Percentage of 2nd Personal Pronoun by gender and time point.**

The third person pronoun exhibited different patterns for the singular and plural over time: Overall, the use of the third singular pronoun increased, while the use of the third plural pronoun decreased. Men had consistently higher use for the singular pronoun until around 2000, when women surpassed men in use frequency. The third person plural pronoun showed a very different change pattern for men and women: Women's use of "they" decreased sharply before 1990, while men's use increased until 1994-1995, when it showed a slight decrease (see Figure 36).

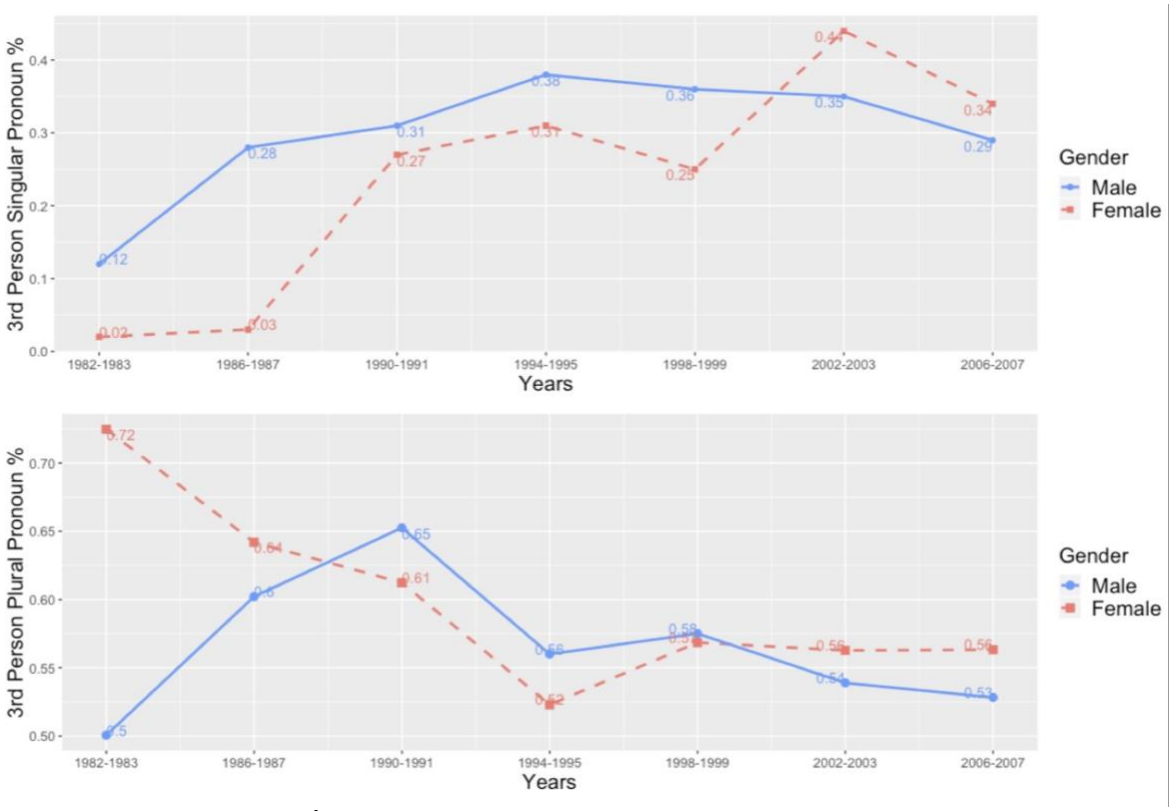


Figure 36. **Percentage of 3rd Person Singular and Plural Personal Pronouns by gender and time point.**

Finally, an important gender marker in online communication identified in previous literature is emoticons (Baron, 2004; Burger et al., 2011; Fullwood et al., 2001; Huffaker & Calvert, 2005; Rao et al., 2010; Waseleski, 2006, Witmer & Katzman, 1997; Wolf, 2000). Only 6.61% of the posts contained emoticons, and 93% of the total emoticons identified were positive. As seen in Figures 37 and 38 below, the frequency of emoticon use increased over time for both men and women. Men used more negative emoticons than women overall. Their use of positive emoticons increased over time, surpassing the frequency of women at certain points. Women's use of emoticons peaked around 1998-1999 but dropped after that.

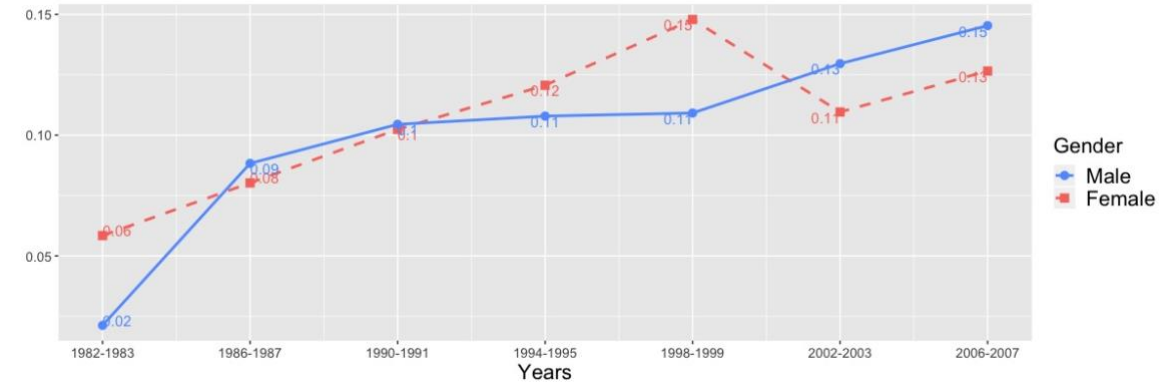


Figure 37. **Percentage of Positive Emoticon use by gender and time point.**

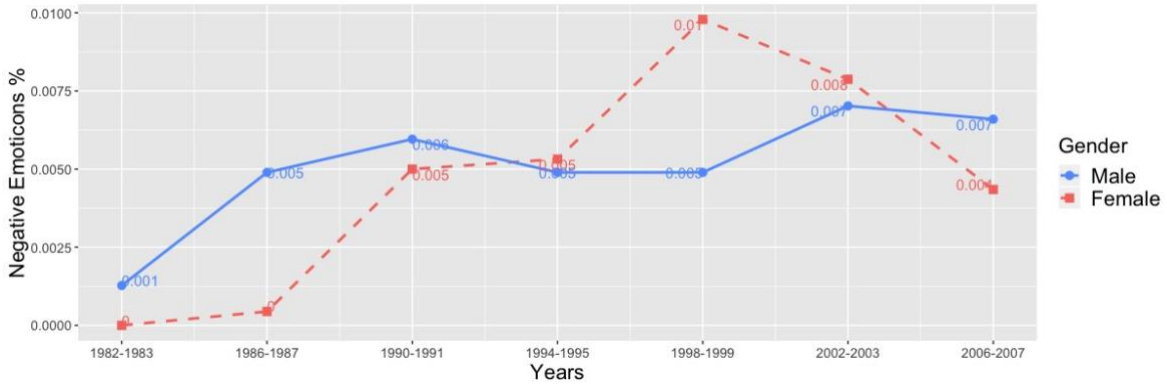


Figure 38. **Percentage of Negative Emoticon use by gender and time point.**

5.3. Significance of Change in Gender Patterns

The changes over time in the frequencies of the 72 variables were analyzed and evaluated for significance. A linear regression analysis was used to test if the interaction of gender and time had a significant effect on the use of each linguistic feature. In total, 24 variables exhibited significant change in their gender patterns during the 25-year period studied (indicated with bold font in Tables 10-21). The results of the analysis for each variable are presented in 13 categories (see Table 9 in section 4.4): Summary Language variables, Pronouns, Common Grammatical Features, Affective Processes., Social Processes, Cognitive Processes, Perceptual Processes, Biological Processes, Drives, Personal Concerns, Informal Language, Punctuation, and Emoticons. As noted in section 4.5, the variables that had to undergo a log transformation are indicated with a cross in the tables, since their calculated results are in a logarithmic scale and not in the original percentile scale.

As seen in Table 10, gender and time individually, as well as the interaction of gender and time, were not significant predictors of words longer than 6 letters in the model, and the model did not explain any significant proportion of the variance in their use. However, gender and time explained a significant proportion of the variance in dictionary words. There was significant variation in the overall use of dictionary words over time and between men and women in general, but not in the gender patterns over time.

Table 10

Gender and Time Models on Summary Language Variables.

	Words > 6 Letters Model+				Dictionary Words Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-1.6758	0.0194	-86.584	< .0001	0.7541	0.0071	106.227	< .0001
Gender	0.0221	0.0262	0.842	0.4000	-0.0262	0.0096	-2.717	0.0068
Time	-0.0027	0.0042	-0.628	0.5300	-0.0080	0.0016	-5.149	< .0001
Gender*Time	-0.0009	0.0058	-0.148	0.8830	0.0027	0.0021	1.281	0.2005
R2		0.002				0.070		
F		1.33				16.99**		

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

According to the results of the linear regression in the category of Common Grammatical Features, gender and time explained a significant proportion of the variance in the use of all variables except for adjectives. As time progressed, there was significant change in the gender patterns of function words, comparisons, and numbers. While gender and time individually were significant predictors of the use of function words and comparisons, time alone was not a significant predictor of the use of numbers. However, time was a significant predictor of the overall use of articles, prepositions, common adverbs, quantifiers, and auxiliary verbs. Gender alone was also a significant predictor of the use of adverbs and common verbs, and it approached very close to the significance threshold for adjectives, conjunctions, and auxiliary verbs. See Table 11.

Table 11

Gender and Time Models on Common Grammatical Features.

Total Function Words					Articles Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.4684	0.0054	86.368	<.0001	0.0699	0.0015	45.495	<.0001
Gender	-0.0247	0.0074	-3.358	<.0001	0.0010	0.0021	0.486	0.6270
Time	-0.0093	0.0012	-7.852	<.0001	-0.0022	0.0003	-6.663	<.0001
Gender*Time	**0.0036**	**0.0016**	**2.210**	**0.0275**	0.0004	0.0005	0.844	0.3990
R2		0.127				0.119		
F		32**				29.75**		
Prepositions Model					Common Adverbs Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.1195	0.0020	59.038	<.0001	0.0455	0.0012	36.888	<.0001
Gender	-0.0039	0.0027	-1.435	0.1517	-0.0043	0.0017	-2.543	0.0112
Time	-0.0021	0.0004	-4.767	<.0001	-0.0011	0.0003	-4.234	<.0001
Gender*Time	0.0010	0.0006	1.717	0.0865	0.0006	0.0004	1.629	0.1038
R2		0.054				0.037		
F		13.14**				9.27**		
Common Adjectives Model					Common Verbs Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0339	0.0011	31.534	<.0001	0.1395	0.0026	53.830	<.0001
Gender	0.0024	0.0015	1.653	0.0988	-0.0108	0.0026	-3.080	0.0022
Time	0.0004	0.0002	1.809	0.0709	0.0014	0.0006	2.488	0.0131
Gender*Time	-0.0004	0.0003	-1.098	0.2728	0.0013	0.0008	1.660	0.0974
R2		0.004				0.064		
F		1.85				15.65**		
Conjunctions Model					Negations Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0504	0.0012	43.786	<.0001	0.0153	0.0007	21.799	<.0001
Gender	-0.0027	0.0016	-1.751	0.0804	-0.0019	0.0010	-2.026	0.0432
Time	-0.0008	0.0003	-3.036	0.0025	0.0001	0.0002	0.656	0.5121
Gender*Time	0.0003	0.0003	0.741	0.4589	0.0004	0.0002	2.088	**0.0372**
R2		0.025				0.018		
F		6.48**				4.91**		
Comparisons Model+					Interrogatives Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-4.6396	0.1288	-36.015	<.0001	0.0162	0.0008	19.462	<.0001
Gender	0.6894	0.1747	3.948	<.0001	-0.0012	0.0011	-1.078	0.2814
Time	0.1031	0.0281	3.664	<.0001	-0.0005	0.0002	-2.691	0.0073
Gender*Time	**-0.1140**	**0.0386**	**-2.957**	**0.0032**	0.0001	0.0002	0.325	0.7452
R2		0.029				0.020		
F		7.29**				5.26**		
Numbers Model+					Quantifiers Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-3.5126	0.1169	-30.062	<.0001	0.0239	0.0007	35.002	<.0001
Gender	0.6405	0.1584	4.044	<.0001	-0.0003	0.0009	-0.370	0.7110
Time	0.0355	0.0255	1.392	0.1644	-0.0008	0.0001	-5.141	<.0001
Gender*Time	**-0.0932**	**0.0350**	**-2.664**	**0.0079**	0.0002	0.0002	0.816	0.4150
R2		0.029				0.063		
F		7.45**				15.35**		
Auxiliary Verbs Model								
Variable	*B*	*SE B*	*t*	*p*				
Intercept	0.0838	0.0015	54.999	<.0001				
Gender	-0.0034	0.0021	-1.670	0.0955				
Time	-0.0015	0.0003	-4.585	<.0001				
Gender*Time	0.0003	0.0005	0.550	0.5823				
R2		0.060						
F		14.61**						

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

Gender and time explained a significant proportion of the variance in the use of all pronouns (Table 12). The first person plural pronoun, second person pronoun, and third person singular pronoun presented significant variation in their gender patterns in general and over time, as well as in their overall use over time. The use of all pronouns, first person singular, and third person plural presented significant variation both overall across time and between the genders in general; however, their gender patterns did not change significantly over time.

Table 12

Gender and Time Models on Pronouns.

Total Pronouns Model					Personal Pronouns Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.1155	0.0027	43.489	<.0001	0.0613	0.0020	30.408	<.0001
Gender	-0.0105	0.0036	-2.905	0.0038	-0.0081	0.0027	-2.964	0.0032
Time	-0.0014	0.0006	-2.406	0.0164	0.0001	0.0004	0.215	0.8295
Gender*Time	0.0006	0.0008	0.739	0.4603	0.0003	0.0006	0.450	0.6527
R2		0.046				0.049		
F		11.2**				12**		
Impersonal Pronouns Model					2nd Person Pronoun Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0541	0.0013	40.706	<.0001	-5.3542	0.1418	-37.757	<.001
Gender	-0.0023	0.0018	-1.300	0.1940	0.5913	0.1923	3.075	0.0022
Time	-0.0015	0.0003	-5.110	<.0001	0.2042	0.0310	6.593	<.0001
Gender*Time	0.0003	0.0004	0.784	0.4340	**-0.0993**	**0.0425**	**-2.338**	**0.0197**
R2		0.063				0.083		
F		15.36**				20.13**		
1st Person Singular Pronoun Model					1st Person Plural Pronoun Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0333	0.0012	26.741	<.0001	-6.5179	0.1570	-41.508	<.0001
Gender	-0.0052	0.0017	-3.053	0.0024	0.9118	0.2129	4.283	<.0001
Time	-0.0008	0.0003	-3.053	0.0024	0.1649	0.0343	4.808	<.0001
Gender*Time	0.0000	0.0004	0.069	0.9451	**-0.1702**	**0.0470**	**-3.621**	**0.0003**
R2		0.088				0.038		
F		21.53**				9.38**		
3rd Person Singular Pronoun Model+					3rd Person Plural Pronoun Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-7.0920	0.2179	-32.547	<.0001	-5.9899	0.1738	-34.473	<.0001
Gender	1.0330	0.2954	3.497	0.0005	0.5567	0.2356	2.363	0.0184
Time	0.2705	0.0476	5.683	<.0001	0.0902	0.0380	2.376	0.0178
Gender*Time	**-0.1957**	**0.0652**	**-3.001**	**0.0028**	-0.0625	0.0520	-1.202	0.2298
R2		0.052				0.018		
F		12.59**				4.83**		

Note. Variables with a cross (+) have undergone a log transformation.
*$p<.05$. **$p<.01$.

Gender and time also explained the variance in the use of all the variables in the Affective Processes category (Table 13). Anger, anxiety, and positive emotion showed significant change in their gender patterns over time. However, gender and time individually were significant predictors in the expression of sadness and negative emotion, and their *p*-values for the interaction of gender and time were also very close to the significance threshold.

Table 13

Gender and Time Models on Affective Processes.

Total Affective Processes Model					Anger Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0355	0.0014	25.949	<.0001	-7.1098	0.1852	-38.401	<.0001
Gender	-0.0006	0.0019	-0.350	0.7270	0.8849	0.2510	3.526	<.0001
Time	0.0016	0.0003	5.483	<.0001	0.2233	0.0404	5.521	<.0001
Gender*Time	-0.0005	0.0004	-1.166	0.2440	**-0.1101**	**0.0554**	**-1.987**	**0.0473**
R2		0.081				0.074		
F		19.81**				17.88**		
Anxiety Model+					Sadness Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-8.2501	0.1831	-45.048	<.0001	-7.3155	0.1740	-42.051	<.0001
Gender	1.0613	0.2483	4.275	<.0001	0.7068	0.2359	2.997	0.0028
Time	0.2581	0.0400	6.453	<.0001	0.1861	0.0380	4.899	<.0001
Gender*Time	**-0.1649**	**0.0548**	**-3.008**	**0.0027**	-0.1002	0.0521	-1.924	0.0548
R2		0.081				0.051		
F		19.72**				12.41**		
Positive Emotion Model+					Negative Emotion Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-4.1850	0.1149	-36.418	<.0001	-5.2596	0.1593	-33.017	<.0001
Gender	0.3708	0.1558	2.380	0.0176	0.4978	0.2160	2.305	0.0215
Time	0.1151	0.0251	4.584	<.0001	0.1454	0.0348	4.18	<.0001
Gender*Time	**-0.0923**	**0.0344**	**-2.683**	**0.0075**	-0.0797	0.0477	-1.671	0.0952
R2		0.029				0.032		
F		7.34**				8.12**		

Note. Variables with a cross (+) have undergone a log transformation.
*$p<.05$. **$p<.01$.

With regard to Social Processes (Table 14), gender and time explained a significant proportion of the variance in all variables of the categories. Only male references and the use

of words related to friends showed significant variation in gender patterns over time. Gender alone was also a significant predictor in the use of words included in the total social processes category of LIWC2015 and related to family; it was also just above the significance threshold for female references.

Table 14

Gender and Time Models on Social Processes.

Total Social Processes Model					Family Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0729	0.0025	28.883	<.0001	0.0026	0.0003	7.994	<.0001
Gender	-0.0087	0.0034	-2.552	0.0110	-0.0009	0.0004	-2.109	0.0354
Time	0.0009	0.0006	1.661	0.0971	-0.0001	0.0001	-0.93	0.3528
Gender*Time	0.0004	0.0008	0.507	0.6120	<0.0001	0.0001	0.38	0.7041
R2		0.046				0.022		
F		11.18**				5.79**		
Female References Model					Male References Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0036	0.0005	7.017	<.0001	-6.5327	0.2090	-31.263	<.0001
Gender	-0.0012	0.0007	-1.792	0.0736	0.8619	0.2833	3.042	0.0024
Time	0.0001	0.0001	0.519	0.6042	0.1790	0.0456	3.923	<.0001
Gender*Time	0.0000	0.0002	-0.099	0.9215	**-0.1392**	**0.0626**	**-2.225**	**0.0265**
R2		0.054				0.028		
F		13.01**				7.13**		
Friends Model+								
Variable	*B*	*SE B*	*t*	*p*				
Intercept	-7.4735	0.1678	-44.532	<>0001				
Gender	0.6796	0.2275	2.987	0.0029				
Time	0.2026	0.0367	5.526	<.0001				
Gender*Time	**-0.1439**	**0.0502**	**-2.864**	**0.0043**				
R2		0.046						
F		11.34**						

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

A significant proportion of the variance in the use of all the variables belonging to the Cognitive Process category was explained by gender and time. The gender patterns in the use of words expressing tentative language showed significant change over time, but also in their overall use across time and between the genders. Gender was a significant predictor for words suggesting insight, causation, and certainty, and time was a significant predictor for

words suggesting cognitive processes in general, and specifically for words suggesting insight, certainty, discrepancy, and differentiation. See Table 15.

Table 15

Gender and Time Models on Cognitive Processes.

Total Cognitive Processes					Insight Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.1341	0.0023	58.329	<.0001	0.0259	0.0008	32.805	<.0001
Gender	-0.0027	0.0031	-0.855	0.3930	-0.0021	0.0011	-1.974	0.0489
Time	-0.0050	0.0005	-9.926	<.0001	-0.0008	0.0002	-4.665	<.0001
Gender*Time	0.0009	0.0007	1.247	0.2130	0.0004	0.0002	1.489	0.1370
R2		0.214				0.042		
F		58.89**				10.39**		
Causation Model					Certainty Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0168	0.0009	19.161	<.0001	-4.8549	0.1183	-41.056	<.0001
Gender	0.0024	0.0012	1.984	0.0476	0.3329	0.1603	2.076	0.0383
Time	-0.0003	0.0002	-1.457	0.1457	0.0623	0.0258	2.411	0.0162
Gender*Time	-0.0003	0.0003	-1.249	0.2121	-0.0453	0.0354	-1.280	0.2011
R2		0.023				0.012		
F		6.01**				3.50**		
Discrepancy Model					Differentiation Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0182	0.0007	26.577	<.0001	0.0456	0.0012	38.301	<.0001
Gender	-0.0009	0.0009	-1.006	0.3147	-0.0002	0.0016	-0.142	0.8870
Time	-0.0006	0.0001	-3.794	0.0002	-0.0028	0.0003	-10.632	<.0001
Gender*Time	0.0003	0.0002	1.304	0.1927	0.0003	0.0004	0.774	0.4390
R2		0.025				0.254		
F		6.43**				73.58**		
Tentativeness Model								
Variable	*B*	*SE B*	*t*	*p*				
Intercept	0.0375	0.0010	37.119	<.0001				
Gender	-0.0031	0.0014	-2.238	0.0256				
Time	-0.0014	0.0002	-6.342	<.0001				
Gender*Time	**0.0006**	**0.0003**	**2.032**	**0.0426**				
R2		0.075						
F		18.36**						

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

Gender and time explained a significant proportion of the variance in the use of the variables in the Perceptual Processes category, as well (Table 16). The gender patterns in the use of words related to sight and hearing showed significant variation over time, as well as

between genders in general and over time overall. Time was a significant predictor in the use of words related to perceptual processes in general, but also for touch (feel) individually. Gender was a significant predictor for touch (feel) and close to the significance threshold for perceptual processes in general.

Table 16

Gender and Time Models on Perceptual Processes.

Total Perceptual Processes+					See Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-4.6661	0.1319	-35.370	<.0001	-5.8557	0.1402	-41.756	<.0001
Gender	0.3348	0.1788	1.872	0.0617	0.7640	0.1901	4.019	<.0001
Time	0.0944	0.0288	3.278	0.0011	0.1704	0.0306	5.564	<.0001
Gender*Time	-0.0651	0.0395	-1.648	0.0998	**-0.1371**	**0.0420**	**-3.265**	**0.001**
R2		0.015				0.051		
F		4.18**				12.54**		
Hear Model+					Feel Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-6.2813	0.1606	-39.120	<.0001	-7.0187	0.1686	-41.642	<.0001
Gender	0.6636	0.2177	3.048	0.0024	0.5594	0.2285	2.448	0.0146
Time	0.1407	0.0351	4.011	<.0001	0.1555	0.0368	4.223	<.0001
Gender*Time	**-0.1123**	**0.0481**	**-2.336**	**0.0198**	-0.0865	0.0505	-1.714	0.0870
R2		0.027				0.034		
F		7.01**				8.50**		

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

Similarly, gender and time explained a significance proportion of the variance in the use of all variables in the Biological Processes category (Table 17). The results suggest that there was significant variation over time in the gender patterns of biological processes in general and specifically in words referring to the body, health, and ingestion. Gender and time individually were also significant predictors for all the variables in this category.

Table 17

Gender and Time Models on Biological Processes.

Total Biological Processes Model+					Body Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-5.5232	0.1662	-33.233	<.0001	-7.42207	0.19101	-38.856	<.0001
Gender	0.5123	0.2253	2.274	0.0233	1.0116	0.2590	3.906	0.0001
Time	0.1433	0.0363	3.947	<.0001	0.2495	0.0417	5.981	<.0001

Gender*Time	**-0.1000**	**0.0498**	**-2.011**	**0.0448**	**-0.1548**	**0.0572**	**-2.708**	**0.0070**
R2		0.023				0.071		
F		6.02**				17.20**		
	Health Model+				Sexual Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-6.9930	0.1807	-38.699	<.0001	-9.0910	0.2397	-37.932	<.0001
Gender	0.6442	0.2450	2.630	0.0088	0.9436	0.3249	2.904	0.0038
Time	0.1830	0.0395	4.636	<.0001	0.2967	0.0523	5.668	<.0001
Gender*Time	**-0.1092**	**0.0541**	**-2.019**	**0.0439**	-0.1079	0.0717	-1.504	0.1330
R2		0.038				0.079		
F		9.44**				19.31**		
	Ingestion Model+							
Variable	*B*	*SE B*	*t*	*p*				
Intercept	-7.3963	0.2130	-34.73	<.0001				
Gender	0.7702	0.2887	2.668	0.0078				
Time	0.1815	0.0465	3.902	0.0001				
Gender*Time	**-0.1260**	**0.0638**	**-1.977**	**0.0485**				
R2		0.027						
F		6.83**						

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

As far as Drives are concerned in Table 18, gender and time explained a significant proportion of the variance in the use of all variables in the category. Words indicating achievement, reward, and risk as drives in the language of men and women exhibited significant variation in their gender patterns over time. They also presented significant variation in their overall use over time, as well as in the use between men and women in general. Additionally, time was a significant predictor for the expression of affiliation and power as drives, and gender was very close to the significance threshold for power.

Table 18

Gender and Time Models on Drives.

	Affiliation Model				Achievement Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0105	0.0009	11.33	<.0001	-5.1619	0.1245	-41.475	<.0001
Gender	-0.0001	0.0013	-0.107	0.9150	0.4937	0.1687	2.926	0.0036
Time	0.0015	0.0002	7.385	<.0001	0.1147	0.0272	4.220	0.0162
Gender*Time	-0.0005	0.0003	-1.934	0.0536	**-0.0978**	**0.0373**	**-2.625**	**0.0089**
R2		0.133				0.025		
F		33.75**				6.54**		
	Power Model+				Reward Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*

Intercept	-4.4706	0.1052	-42.517	<.0001	-5.1265	0.1337	-38.338	<.0001
Gender	0.2766	0.1426	1.941	0.0527	0.4425	0.1813	2.441	0.0149
Time	0.0788	0.0230	3.432	0.0006	0.0966	0.0292	3.308	0.0010
Gender*Time	-0.0402	0.0315	-1.278	0.2016	**-0.0860**	**0.0400**	**-2.148**	**0.0321**
R2	0.023				0.014			
F	5.94**				4.08**			

Risk Model+				
Variable	*B*	*SE B*	*t*	*p*
Intercept	-6.3787	0.1640	-38.905	<.0001
Gender	0.7854	0.2223	3.534	0.0004
Time	0.1236	0.0358	3.452	0.0006
Gender*Time	**-0.1059**	**0.0491**	**-2.157**	**0.0314**
R2	0.033			
F	8.29**			

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

Regarding Personal Concerns (Table 19), a significant proportion of the variance in the use of all variables in this category was explained by gender and time. The use of words related to money presented significant variation in the frequency difference between men and women over time, as well as in their overall use over time, and between men and women in general. Time also had a significant effect on the use of words related to work, leisure, religion and death, whereas gender appeared to have a significant effect on the use of words related to home, religion, and death: Women wrote more about home, while men wrote more about religion and death.

Table 19

Gender and Time Models on Personal Concerns.

Work Model+					Leisure Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-4.2662	0.1168	-36.536	<.0001	-5.4526	0.1484	-36.743	<.0001
Gender	0.2426	0.1583	1.533	0.1260	0.3491	0.2012	1.735	0.0832
Time	0.1863	0.0255	7.304	<.0001	0.1006	0.0324	3.103	0.0020
Gender*Time	-0.0286	0.0350	-0.819	0.4130	-0.0757	0.0444	-1.703	0.0890
R2	0.131				0.012			
F	32.99**				3.48*			

Home Model					Money Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0037	0.0004	10.448	<.0001	-6.5504	0.1680	-38.998	<.0001
Gender	-0.0012	0.0005	-2.43	0.0154	0.9748	0.2277	4.281	<.0001
Time	-0.0001	0.0001	-1.181	0.2381	0.2415	0.0367	6.583	<.0001

Gender*Time	0.0001	0.0001	0.948	0.3436	**-0.1586**	**0.0503**	**-3.153**	**0.0017**
R2		0.020				0.080		
F		5.37**				19.55**		
	Religion Model+				Death Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-8.5497	0.2295	-37.261	<.0001	-8.4689	0.2026	-41.81	<.0001
Gender	1.0399	0.3111	3.343	0.0009	0.9027	0.2746	3.287	0.0011
Time	0.3101	0.0501	6.188	<.0001	0.2125	0.0442	4.802	<.0001
Gender*Time	-0.1251	0.0687	-1.822	0.0690	-0.0891	0.0606	-1.470	0.1420
R2		0.091				0.069		
F		22.38**				16.85**		

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

A significant proportion of the variance in the use of linguistic features indicating Informal Language was also explained by gender and time. While the individual variables examined in this category did not show significant variation in gender patterns over time, the overall use of informal language did. Moreover, the use of swear words, words indicating assent, nonfluencies, and fillers showed significant variation overall across time, and also between men and women. See Table 20.

Table 20

Gender and Time Models on Informal Language.

	Total Informal Language Model				Swear Words Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0088	0.0009	9.440	<.0001	-9.6823	0.2133	-45.401	<.0001
Gender	0.0044	0.0013	3.507	0.0005	1.0890	0.2891	3.767	0.0002
Time	0.0011	0.0002	5.589	<.0001	0.4372	0.0466	9.386	<.0001
Gender*Time	**-0.0009**	**0.0003**	**-3.301**	**0.0010**	-0.0901	0.0638	-1.412	0.1585
R2		0.046				0.216		
F		11.21**				59.74**		
	Assent Model+				Nonfluencies Model+			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0088	0.0009	9.440	<.0001	-7.3309	0.1856	-39.505	<.0001
Gender	0.0044	0.0013	3.507	0.0005	0.5909	0.2516	2.349	0.0191
Time	0.0011	0.0002	5.589	<.0001	0.1365	0.0405	3.368	0.0008
Gender*Time	-0.0067	0.0036	-1.865	0.0621	-0.0483	0.0556	-0.869	0.3854
R2		0.146				0.039		
F		37.27**				9.57**		
	Fillers Model+							
Variable	*B*	*SE B*	*t*	*p*				
Intercept	-9.5227	0.1786	-53.307	<.0001				
Gender	0.8492	0.2422	3.506	0.0005				

Time	0.1980	0.0390	5.074	<.0001
Gender*Time	-0.0772	0.0535	-1.444	0.1491
R2		0.082		
F		20.09**		

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

A significant proportion of the variance in the use of question marks, exclamation marks, and quotation marks in Table 21 was explained by gender and time. Even though the punctuation examined in this study did not present significant variation in gender patterns over time, its overall use had significant variation over time. The use of question marks and quotation marks also showed significant variation between genders and in their overall use across time. However, gender did not appear to have a significant effect on the use of exclamation marks.

Table 21

Gender and Time Models on Punctuation.

Question Marks Model+					Exclamation Marks Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	-5.4745	0.1679	-32.613	<.0001	0.0209	0.0025	8.376	<.0001
Gender	0.5793	0.2276	2.546	0.0111	-0.0034	0.0034	-1.008	0.3139
Time	0.1310	0.0367	3.572	0.0004	-0.0018	0.0005	-3.216	0.0014
Gender*Time	-0.0768	0.0503	-1.528	0.1269	0.0001	0.0007	0.169	0.8656
R2		0.029				0.031		
F		7.26**				7.91**		
Quotation Marks Model+								
Variable	*B*	*SE B*	*t*	*p*				
Intercept	-5.7575	0.1753	-32.846	<.0001				
Gender	0.8000	0.2376	3.366	0.0008				
Time	0.2333	0.0383	6.093	<.0001				
Gender*Time	-0.0801	0.0525	-1.527	0.1273				
R2		0.100						
F		24.64**						

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

Finally, gender and time explained a significant proportion of the variance in the use of emoticons in total, but also positive and negative emoticons individually. None of the three models evaluated as significant any changes in the gender patterns of the total, positive, and

negative emoticon use over time; gender was not considered a significant predictor in any of the models, whereas time appeared to have a significant effect on the overall use of emoticons in general, as well as on positive and negative emoticons. See Table 22.

Table 22

Gender and Time Models on Emoticons.

Total Emoticons Model					Positive Emoticons Model			
Variable	*B*	*SE B*	*t*	*p*	*B*	*SE B*	*t*	*p*
Intercept	0.0007	0.0002	3.893	0.0001	0.0007	0.0002	3.981	<.0001
Gender	-0.0003	0.0002	-1.223	0.2218	-0.0003	0.0002	-1.353	0.1765
Time	0.0001	<.0001	3.107	0.0020	0.0001	<.0001	2.876	0.0042
Gender*Time	0.0001	0.0001	1.092	0.2751	0.0001	<.0001	1.207	0.2280
R2		0.047				0.044		
F		11.46**				10.77**		
Negative Emoticons Model								
Variable	*B*	*SE B*	*t*	*p*				
Intercept	<.0001	<.0001	0.182	0.8557				
Gender	<.0001	<.0001	1.096	0.2736				
Time	<.0001	<.0001	3.490	0.0005				
Gender*Time	<.0001	<.0001	-0.943	0.3459				
R2		0.023						
F		5.95**						

Note. Variables with a cross (+) have undergone a log transformation.
*p<.05. **p<.01.

5.4. Trends in Gender Patterns

Under the language contact metaphor invoked in his study, based on the theoretical framework of language change as a result of language contact, we may use the concepts of convergence and divergence to interpret the patterns in which changes in the language use of men and women manifested in CMD. This section presents the results of evaluating the gender trend lines created for each variable in order to facilitate the identification of longitudinal trends that showed convergence or divergence in the linguistic styles of men and women.

A total of 62 variables out of the 72 analyzed in this study exhibited convergence, divergence, or convergence followed by divergence (reversal) in their gender patterns over

time. Moreover, the steepness of the trend line slopes was evaluated by comparing the absolute values of their slope angles, in order to identify which gender exhibited a faster rate of change and thus might be considered to have led the change in the use of that linguistic variable. The cases where there was difference in the steepness of the trend line slope that was less than one degree are indicated as *Not Applicable* (NA) in the tables. Among the features that did not present any gender pattern change in their trend lines were the following: words longer than 6 letters; the 1st person singular pronoun; words indicating social processes, death, and religion; sexual words; fillers and nonfluencies; exclamation marks; and quotation marks. The results of the trend line evaluations are summarized and presented separately below for linguistic features showing convergence, divergence, and reversal of gender styles. Only a few variables showcasing clear examples of the above longitudinal trends are presented in detail here, but figures with trend lines for all the variables can be found in Appendix E.

5.4.1. Convergence

About one third, or 32%, of the linguistic features analyzed in this study showed a convergence in the language use of men and women; these are summarized in Table 23. The rate of change in the use of most linguistic features was faster for women (higher absolute values of slope angles).

Table 23

Variables Showing Convergence and Gender with Faster Rate Change.

Category	Variable	Faster Change Rate		
		Female	Male	N/A
Summary Language Variables	Dictionary Words	X		
Common Grammatical Features	Total Function Words			X
	Common Adverbs	X		

	Common Adjectives	X		
	Common Verbs		X	
	Conjunctions	X		
	Interrogatives	X		
	Numbers		X	
	Auxiliary Verbs	X		
Pronouns	Total Pronouns	X		
	Personal Pronouns		X	
	Impersonal Pronouns	X		
	3rd Person Plural Pronouns	X		
Affective processes	Anger	X		
	Sadness	X		
Social Processes	Family	X		
Cognitive Processes	Causation			X
	Certainty	X		
Drives	Power	X		
	Risk	X		
Personal Concerns	Work	X		
	Home	X		
Punctuation	Question Marks	X		

One of the cases of convergence in the linguistic styles of women and men involved words indicating certainty. Even though their frequency increased in the posts of both genders, the incline of the slope was steeper for women, indicating a faster rate of change. The trend lines are shown in Figure 39.

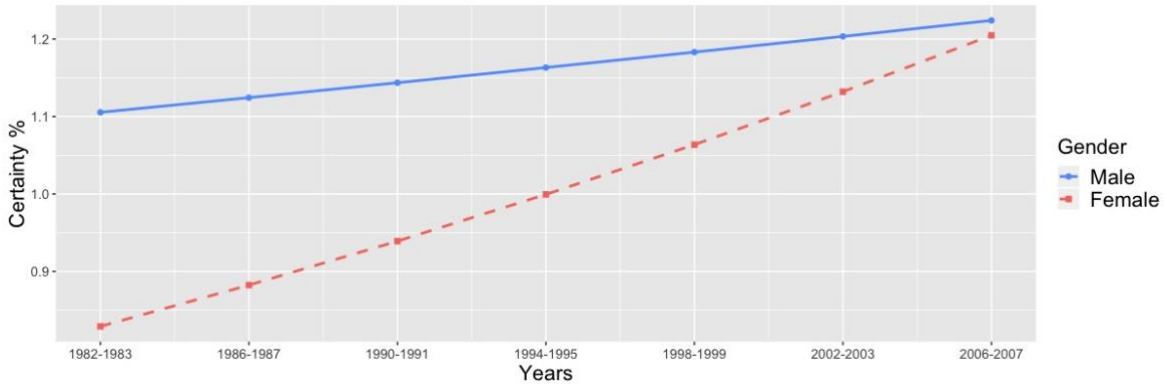


Figure 39. **Trend lines for Certainty by gender.**

The case of numbers in Figure 40 also presents an interesting case of convergence in longitudinal gender trends. Both genders showed change; however, it was in the opposite direction, rather than the same direction. Men started high and decreased their use of numbers, while women started low and increased their use, meeting almost in the middle in 2006-2007; this would indicate a move toward a more gender-neutral use of this variable, as defined in section 3.2.

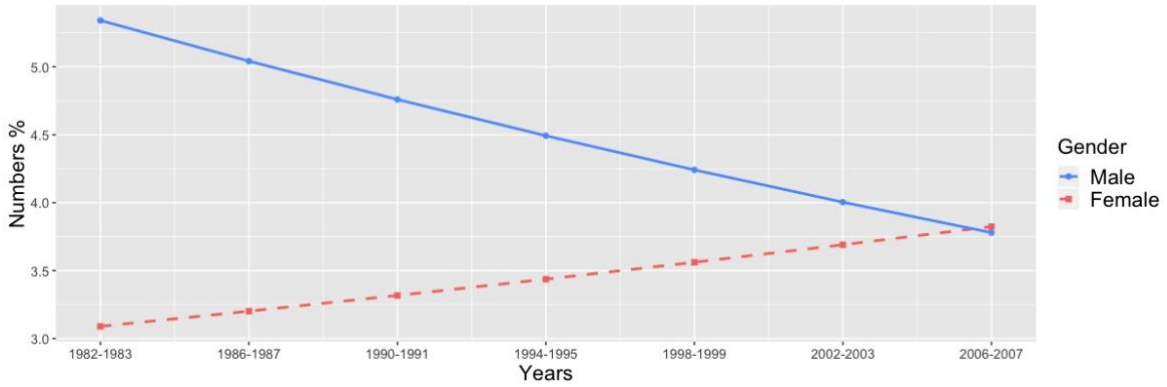


Figure 40. **Trend lines for Numbers by gender.**

5.4.2. Divergence

Only a few variables (8%) with different frequencies for women and men in their gender patterns were further differentiated, as shown in Table 24 below. As with the variables exhibiting convergence of gender usage over time, the changes in the majority of the features that diverge in their use between men and women appear to occur at a faster rate for women (higher absolute values of slope angles).

Table 24

Variables showing divergence and gender with faster rate change.

Category	Variable	Faster Change Rate		
		Females	Males	N/A
Common Grammatical Features	Articles	X		
Affective Processes	Total Affective Processes	X		
Social Processes	Female References	X		

Cognitive Processes	Differentiation			X
Drives	Affiliation	X		
Informal Language	Swear Words		X	

An example of a linguistic feature that decreased in frequency at a faster rate for one gender than for the other is articles. While their frequency in the posts of both men and women decreased, it decreased at a somewhat faster rate for women, making their use of this gender marker even more different from the use of men over time, as shown in Figure 41.

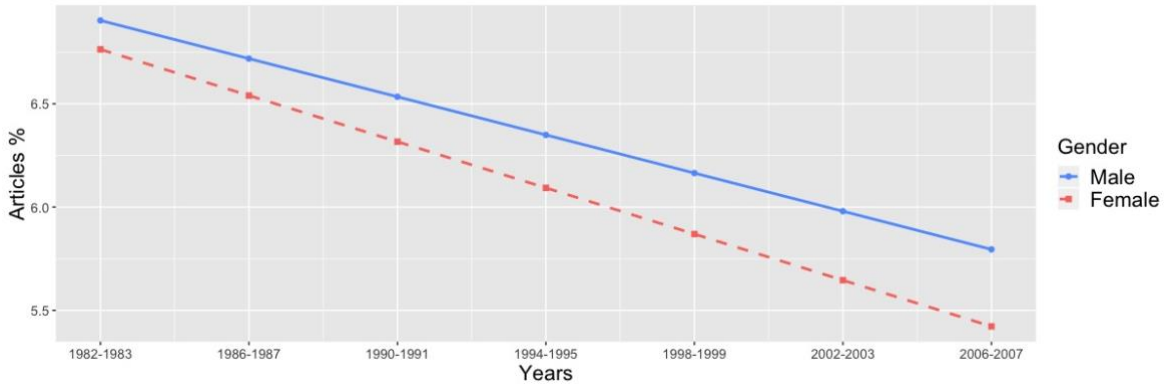


Figure 41. **Trend lines for Articles by gender.**

The only variable where the rate of change was faster for men was that of swear words. While their overall use increased, it did so faster for men, resulting in a greater difference in frequencies at the end of the time period studied (see Figure 42).

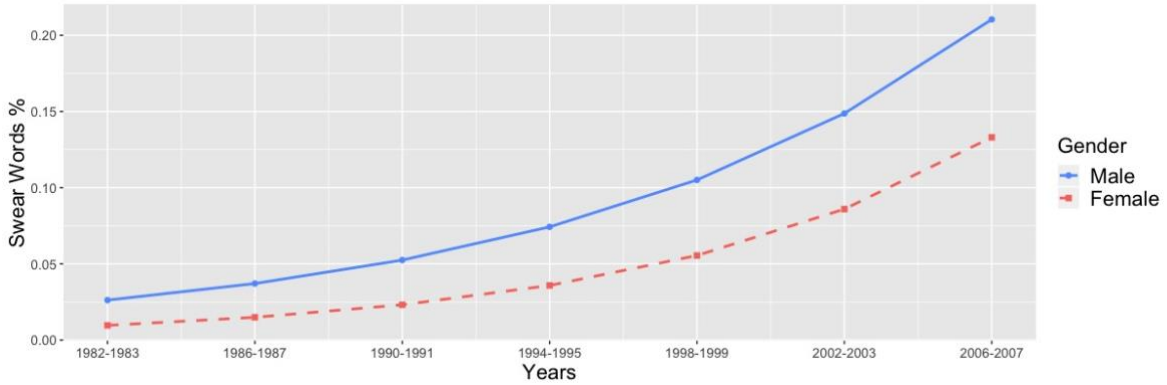


Figure 42. **Trend lines for Swear Words by gender.**

5.4.3. Reversal

The majority of the language variables analyzed (46%) showed a reversal of patterns. It should be noted that some variables exhibiting a reversal of patterns in this section had *p*-values just above the significance threshold with regard to the interaction of gender and time in their model, or else their changes were in a numerical scale too small to be considered significant by the models. For a few, the reversal happened earlier in the time period studied, showing non-traditional use of language by women and men that fell into more traditional gender patterns later. For most, though, the reversal happened later; a number of those changes clustered around or after 2000. Women's trend lines showed faster rates of change (higher absolute values of slope angles) for most of the variables presented in Table 25 below.

Table 25

Variables showing reversal of patterns and gender with faster rate change.

Category	Variable	Faster Change Rate		
		Females	Males	N/A
Common Grammatical Features	Prepositions	X		
	Negations		X	
	Comparisons	X		
	Quantifiers			X
Pronouns	2nd Person Pronoun	X		
	1st Person Plural Pronoun	X		
	3rd Person Singular Pronoun	X		
Affective Processes	Positive Emotion	X		
	Negative Emotion	X		
	Anxiety	X		
Social Processes	Friends	X		
	Male References	X		
Cognitive Processes	Total Cognitive Processes	X		
	Insight	X		
	Discrepancy	X		
	Tentativeness	X		
Perceptual Processes	Total Perceptual Processes	X		
	See	X		
	Hear	X		
	Feel	X		
Biological	Total Biological Processes	X		

Processes	Body	X	
	Health	X	
	Ingestion	X	
Drives	Achievement	X	
	Reward	X	
Personal Concerns	Leisure	X	
	Money	X	
Informal Language	Total Informal Language	X	
	Assent	X	
Emoticons	Total Emoticons		X
	Positive Emoticons		X
	Negative Emoticons	X	

As seen in Figure 43 below, women expressed less positive emotion compared to men at the beginning of the time period studied. However, the frequency of words expressing positive emotion increased in women's posts over time, while that of men remained fairly stable. Toward the middle of the time period, women's expression of positive emotion surpassed that of men's, thereby manifesting a more traditional gender pattern (e.g., Ottoni et al., 2013).

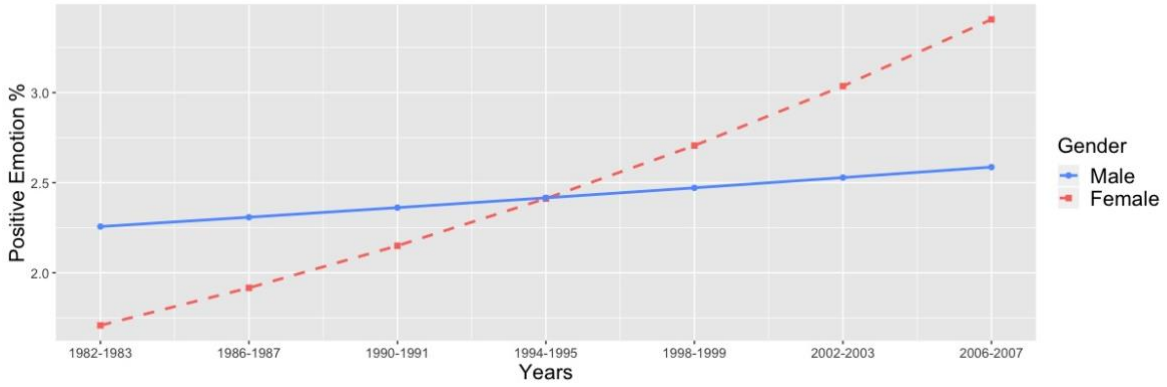


Figure 43. **Trend lines for Positive Emotion by gender.**

The use of words related to biological processes (Figure 44) and indicating tentativeness (Figure 45) showed different overall trends, but the reversal of their gender patterns happened around the same time, with women showing a faster rate of change in both. The mentions of biological processes increased over time in the posts of both women and

men; however, women showed a much faster rate of change, eventually surpassing the use of this variable by men around 2000. Similarly, the use of tentative language decreased over time for both genders, but it did so faster for women, resulting in a reversal of the traditional gender pattern (Fullwood et al., 2001; Herring et al., 1992).

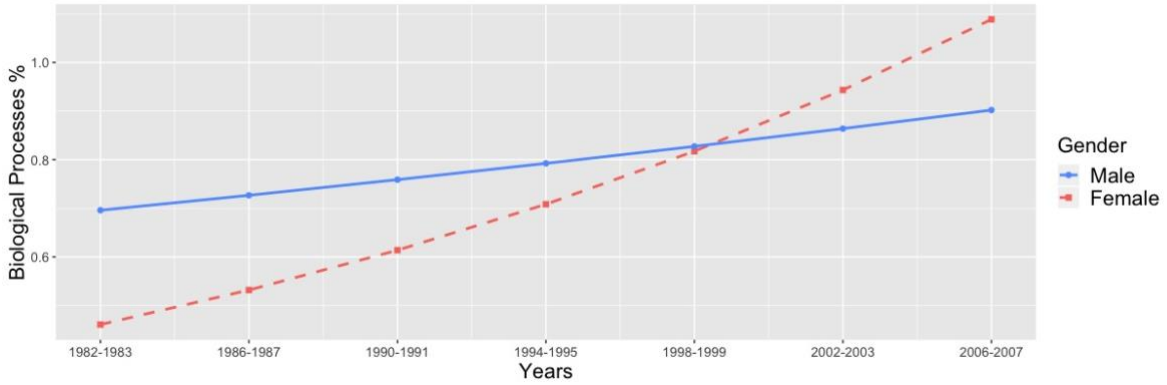


Figure 44. **Trend lines for Biological Processes by gender.**

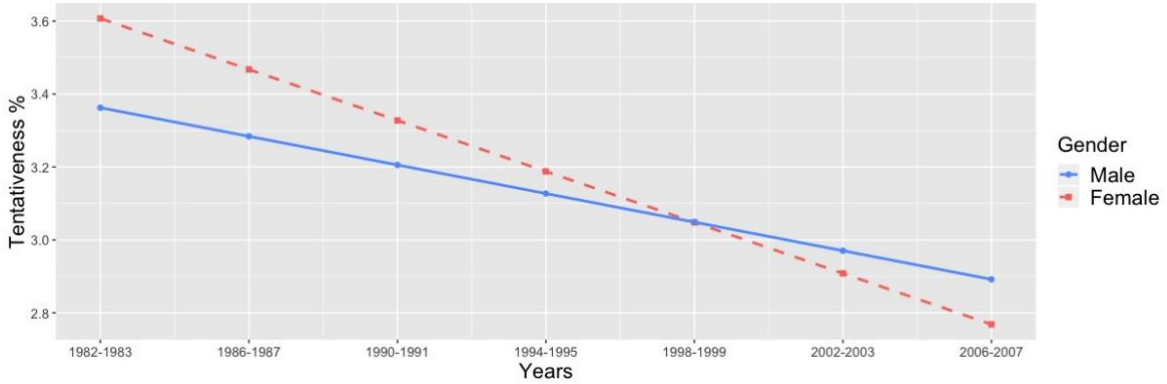


Figure 45. **Trend lines for Tentativeness by gender.**

5.5. Female-Predominant Newsgroups: the Case of *net.kids/misc.kids*

Differences in the linguistic styles of women and men often occur due to power and status (in)equalities; the language of men has been traditionally associated with higher status and power as opposed to feminine linguistic style (Coates, 1993). Similar dynamics have been identified in online environments, where the dominance of one gender may affect the linguistic styles of its members, as Herring et al. (1998) found in their study of the MegaByte University (MBU) mailing list.

Based on the ratio of posts between women and men, almost all the newsgroups in the dataset of this study were male-predominant. There was only one newsgroup where the number of female posts was higher: *net.kids/misc.kids*.[12] The focus of this newsgroup is children, their behavior, and their activities. Because the majority of caretakers of children are female, the main participants in the newsgroup discussions were women, posting more than twice as much in the group compared to men.

The comparison of one female-predominant newsgroup with 46 male-predominant newsgroups could not provide statistically valid grounds for any generalizable conclusions. Consequently, the newsgroup *net.kids/misc.kids* is presented as a short case study of a female-predominant group where the language styles of men and women exhibit differences compared to the male-predominant groups. In order to examine possible differences in the language used in the posts of male-predominant environments versus female-predominant environments, the means of all linguistic features were examined separately for *net.kids/misc.kids* and compared to the remaining dataset.

The use of some linguistic features in *net.kids/misc.kids* exhibited more traditional gender patterns than in the male-predominant dataset. One of the most striking differences is in the use of adjectives. Overall, the dataset shows higher use in the posts of men, a finding that is inconsistent with previous literature. However, men decrease sharply in their use of adjectives on *net.kids/misc.kids* over time, while women used more adjectives over time (see Figure 46).

[12] Even though net.women/soc.women started as a newsgroup for women, it was quickly overtaken by men who dominated the discussions, according to anecdotal accounts (Harter, 1998).

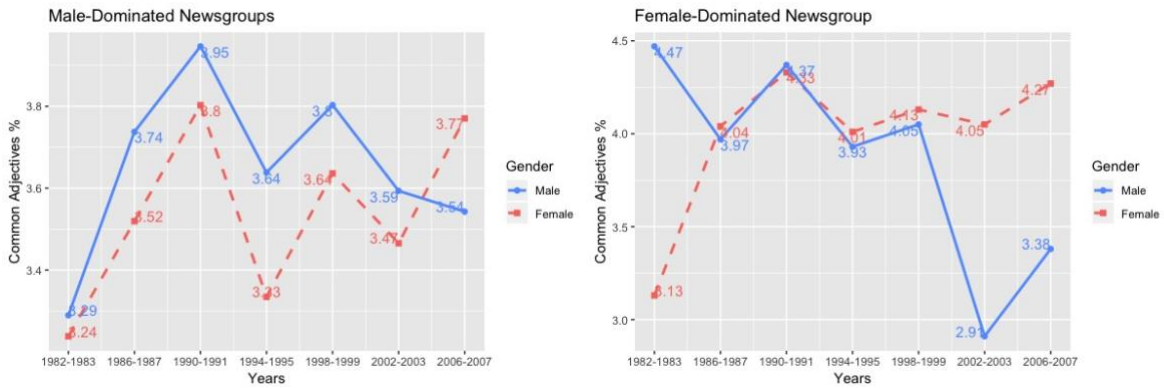


Figure 46. **Differences in the evolution of gender patterns for Adjectives in the male-predominant newsgroups and net.kis/misc.kids.**

Another gender marker showing different patterns in male environments versus the female-predominant environment is that of emotion (Figure 47). Women's language expressed more positive emotion than men throughout the entire time period studied in *net.kids/misc.kids*, as opposed to the higher frequency after 1990 in the general dataset, and men's expression of positive emotion decreased sharply after 2000 in the female-predominant group.

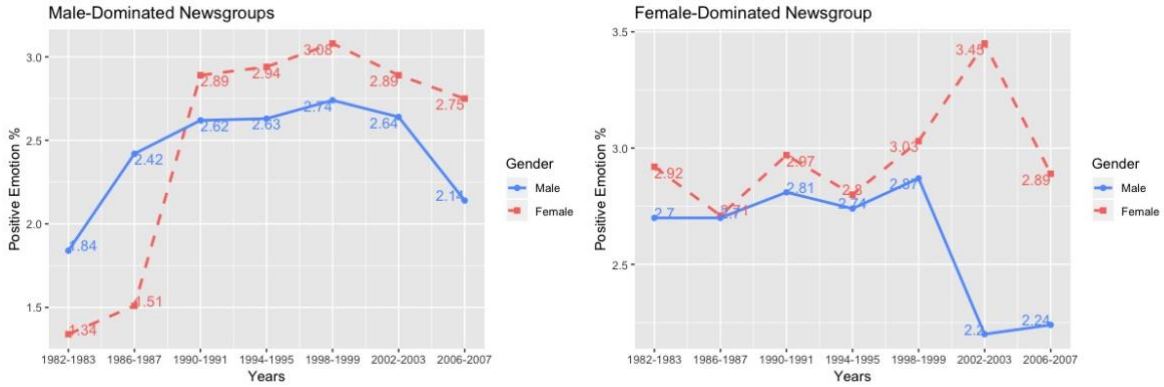


Figure 47. **Differences in the evolution of gender patterns for Positive Emotion in the male-predominant newsgroups and net.kis/misc.kids.**

Around the same time that the frequency of positive emotion in men's posts dropped, men started expressing more anger in net.kids/misc.kids, similarly to the general dataset. However, whereas women's expression of anger seemed to become more similar to men's

over time in the male-predominant environments, it remained much lower in the *net.kids/misc.kids* newsgroup (see Figure 48).

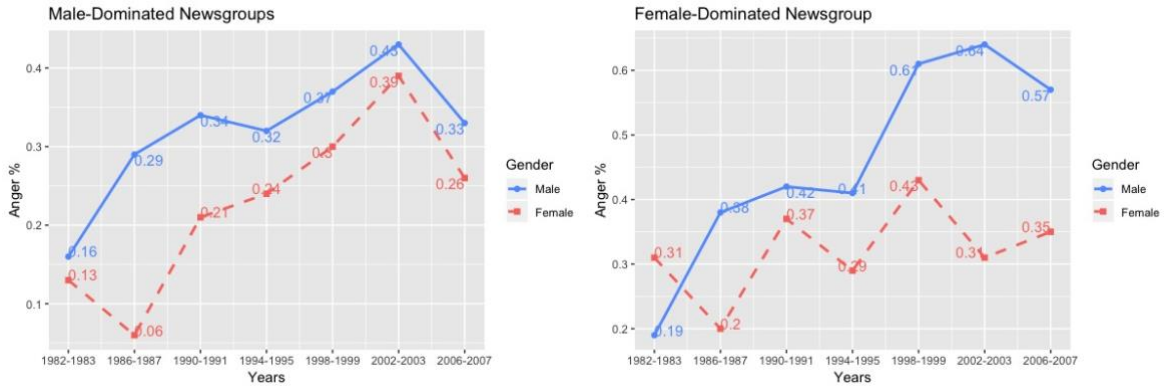


Figure 48. **Differences in the evolution of gender patterns for Anger in the male-predominant newsgroups and net.kis/misc.kids.**

The use of swear words showed a similar development in Figure 49. Men increased their use in both environments; however, their increase in net.kids/misc.kids jumped sharply around 2000, similar to anger, while women's use of swear words increased more steadily at a much lower rate.

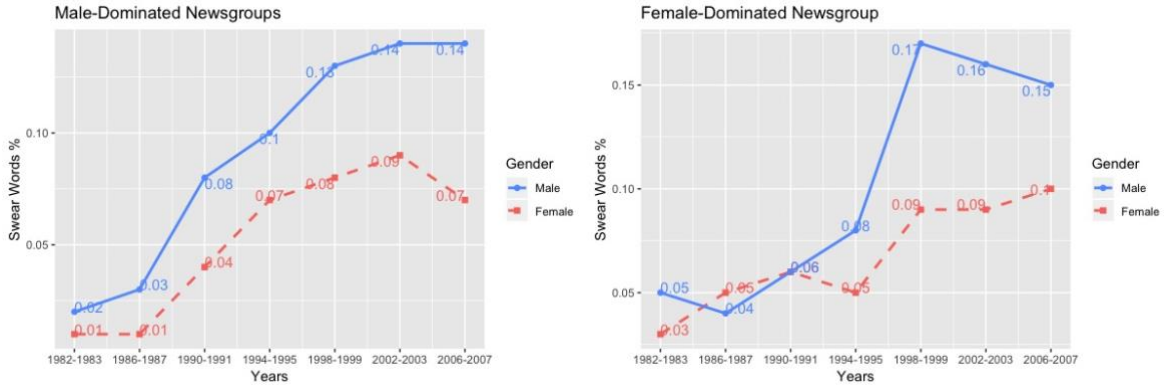


Figure 49. **Differences in the evolution of gender patterns for Swear Words in the male-predominant newsgroups and net.kis/misc.kids.**

The different use of emoticons in the male-predominant and female-predominant environments echoes the findings of Wolf (2000): Women made higher use of emoticons in *net.kids/misc.kids* compared to the male-predominant newsgroups (Figure 50).

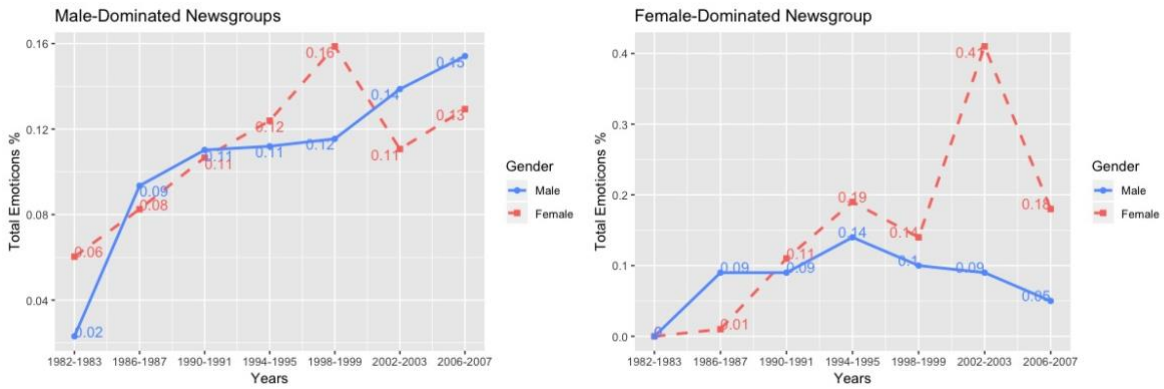


Figure 50. **Differences in the evolution of gender patterns for Emoticons in the male-predominant newsgroups and net.kis/misc.kids.**

One gender marker indicates a possible power shift in the female-predominant newsgroup: the use of words expressing tentativeness (Figure 51). According to traditional gender patterns, women use attenuated assertions, hedges, and qualifiers with higher frequency than men, something that has been associated with power and status inequality (Coates, 1993; Tannen, 1994). This traditional usage is supported overall in the male-predominant newsgroups, even though the use of tentative language decreased rapidly for both genders. However, men expressed more tentativeness than women did in the female-predominant newsgroup at several time points. Since this is a group discussing children with a larger number of women, men may have felt that they did not have the power or expertise to make strong assertions. Similarly, Herring et al. (1998) found that men used more tentative language (hedges) when women exhibited power in the thread they analyzed from MBU by participating more in the discussion than men.

Interestingly, however, women's use of tentative language exhibited an increase in *net.kids/misc.kids*, while men's tentativeness decreased after 2000. If we take into account men's rapid increase in the expression of anger and decrease in the expression of positive emotion around the same time, we might speculate that a power shift occurred in the newsgroup after that time. There is a precedent in *net.women/soc.women*: The newsgroup

was created by women to avoid the flaming in the male-predominant newsgroups and to create a safe environment for discussion among women, but men slowly introduced negative, aggressive behavior again into the newsgroup (Harter, 1998).

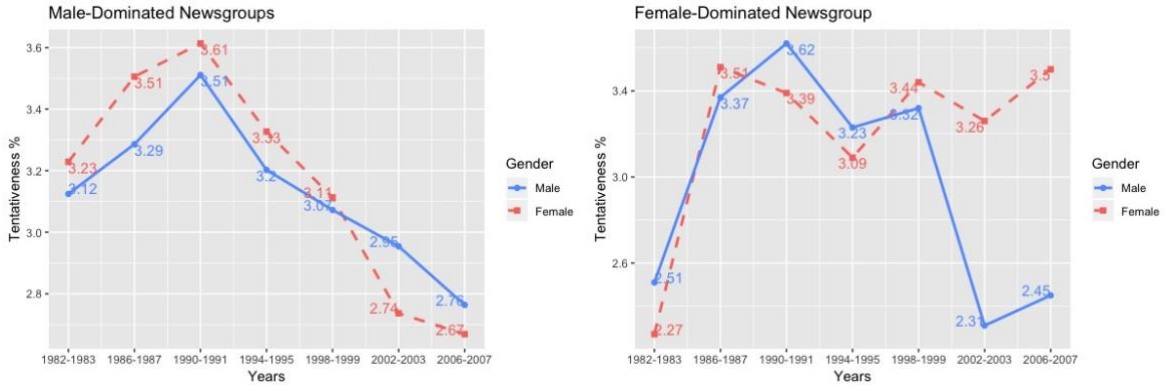


Figure 51. **Differences in the development of gender patterns for Tentativeness in the male-predominant newsgroups and net.kis/misc.kids.**

Furthermore, women's posts in *net.kids/misc.kids* include language expressing more confidence and leadership (< 60%) than in the male-predominant newsgroups (> 60%), as seen in Figure 52. However, in contrast to the male-predominant newsgroups, women's frequency of such language dropped in the female-predominant group after 1990.

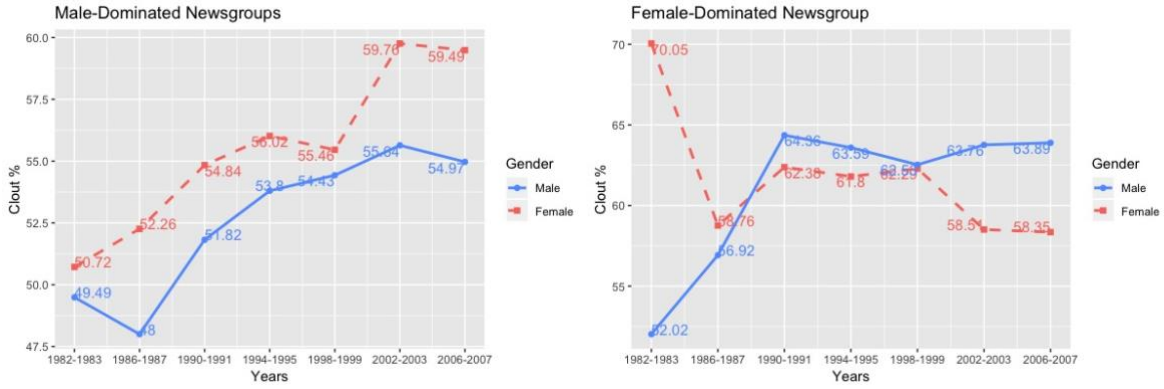


Figure 52. **Differences in the development of gender patterns for Clout in the male-predominant newsgroups and net.kis/misc.kids.**

We may understand how female-predominant online environments potentially provide a safer environment for women, where they feel comfortable stepping out of gender expectations and norms, by looking at the different gender patterns for sexual words compared to male-predominant environments. Men's posts include more explicit sexual

words in the male-predominant groups, which keeps increasing; it is only surpassed by women's posts at one time point throughout the period studied, although women's use decreased subsequently. In *net.kids/misc.kids*, women's use of explicit sexual words increased when their participation started to grow and they established their status in the newsgroup. However, right around the time when a power shift appears to have happened, the frequency of sexual words in their posts decreased. See Figure 53.

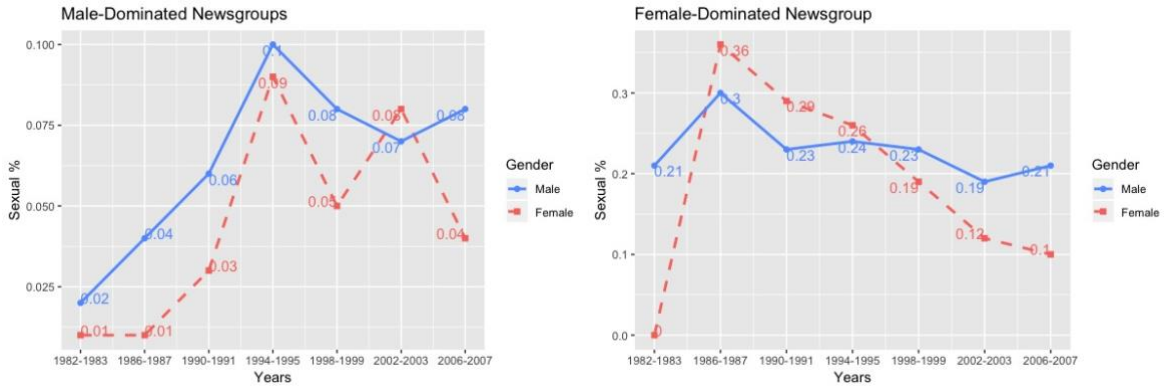


Figure 53. **Differences in the development of gender patterns for Sexual Words in the male-predominant newsgroups and net.kis/misc.kids.**

Overall, the language use in *net.kids/misc.kids* exhibited gender patterns that did not appear to follow the same evolution over time as the rest of the dataset with the male-predominant groups. Moreover, the change in gender patterns for linguistic features related to power and status at certain time points, combined with the negative linguistic behavior by men indicating possible conflict and gender polarization, are suggestive of a power and status shift within the newsgroup. According to Kehoe (1996), conflict in USENET newsgroups occurred often as "rabid arguments" or "flame wars" and, in cases such as the *net.women/soc.women* newsgroup, resulted in a hostile takeover by men (Harter, 1998). With men imposing their (more aggressive) interaction values in the newsgroup after 2000, women seemed to back away from linguistic behaviors expressing power and status and became more tentative and less comfortable stepping out of the expected language use based on the

social norms for their gender. Thus, even though women were still the majority of active participants in the newsgroup, their language usage suggests that they no longer held the same power and status in the newsgroup.

Discussion

This chapter discusses the changes identified with regard to the overall language usage in USENET newsgroups over time, the overall gender variation in USENET newsgroups, as well as the longitudinal changes and trends in the gender patterns for the linguistic features analyzed in the study. It is important to note that the design of this study and the nature of the data do not allow for firm conclusions as regards deep linguistic change. Given the history of USENET newsgroups (see sections 4.2.1 and 4.2.2.4) and the potential for linguistic variation introduced by the changing demographics of USENET users over the 25-year period studied, the trends identified in the previous chapter could be changes in the *discourse* or language usage of the groups but not changes in the *language* (English) itself or in the language styles of individual users. The interpretations suggested below are some broad speculations based on the context around the time and manner of the discussed changes in language use, offered as a starting point for discussion about language change in CMD.

6.1. Change in Overall Language Usage Over Time

The overall use of language in USENET newsgroups changed significantly for the majority of the features (87.5%) studied in the 25-year period examined. While there were differences among the evolution patterns for the variables, a number of them exhibited similarities concentrating around specific time points. The time and directionality of the common changes are suggestive of social factors at play, similarly to Herring's findings in her diachronic study of the MsgGroup and Linguist List (1999a.) Those social factors could be internal (social changes occurring within the newsgroups) or external (social changes occurring broadly in society).

To begin with, a number of variables showed abrupt peaks or drops in their frequencies around 1990, when there were several social changes occurring within the newsgroups. As seen by the distribution of data per time point (Appendix B), the number of messages quadrupled around 1990 compared to the previous time points, indicating a rise in the popularity of newsgroups. In addition to the increase in number and variation in users, this may have brought a shift in the purpose of the newsgroups. According to Wikipedia (2017c), their earlier intended use focused on the dissemination of news and information, which would imply the use of a more formal and structured language. As seen in section 5.1, the frequencies of features indicating more formal, complex language structure, such as function words, words per sentence, and words indicating analytical thinking, dropped after 1990. At the same time, the expression of social processes and the use of pronouns and informal language increased in frequency, suggesting a shift in the purpose of the USENET newsgroups to that of discussion groups, similarly to the findings of Singer et al. for Reddit (2014).

Another group of changes was concentrated around 1995, when internet access started to become more widespread. In the early 1990s, the Internet was introduced to a broader

public, and the dramatic expansion of users called for more networks to satisfy commercial uses (Press, 1994; Weis, 2010). A number of Internet Service Providers arose to satisfy that demand, allowing users to connect from home. Moving Internet access from educational or work environments to more private spaces could have encouraged users to engage in discussions of a more personal nature. Indeed, around 1995, we see an increase in use in words related to friends and family, as well as personal concerns such as leisure, home, money, and religion.

Finally, a cluster of changes in the patterns of certain features suggests that the communication in USENET newsgroups became more negative and polarized over time. The overall negative emotion and negative affective processes such as anger, anxiety, sadness, as well as the use of swear words, increased in frequency over time and peaked in the early 2000s. Interestingly, the 1st person plural and 2nd person pronouns seemed to follow a similar pattern, peaking around 2000, suggesting a possible polarization of *we* versus *you* in discussions.

6.2. Revisiting the Gender Patterns in Previous Literature

Overall, the majority of the linguistic features analyzed presented patterns that are consistent with the findings of previous research: Traditional (offline) gender patterns have transferred to online communication. According to the linear regression models, the majority of the features (69.4%) exhibited significant differences between females and males. The gender marker of participation was not examined in this study due to the fact that the gender of the authors could not be identified for all the posts in the dataset. However, the ratio of female to male posts in the data for which gender was determined (13% female versus 87% male) was similar to the findings of previous research (e.g., Savicki et al., 1996). The only exception was the female-predominant group net.kids/misc.kids, which had a greater number of messages by women.

With regard to post length, the results of this study are consistent with the findings of previous studies (Herring, 1992a, 1992b) which found that men made longer contributions; the men in the present study also wrote longer sentences. Women used personal pronouns with a higher frequency than men, especially the 1st person pronoun, as found in previous literature (Argamon et al., 2007; Ottoni et al., 2013; Schwartz et. al, 2013), whereas men used more articles (Argamon et al., 2007; Herring & Paollilo, 2006; Kapidzic & Herring, 2011; Schwartz et al., 2013). Women's language was overall more personal, emotional, tentative, and reflective of social processes, whereas men's language suggested more formal, logical, and hierarchical thinking (fact-oriented language, Savicki et al., 1996). Women also displayed more positive emotion, while men displayed more negative emotion. This pattern was also reflected in the use of emoticons: Women used more emoticons and more positive emoticons overall, while males used more negative emoticons overall. However, the difference in the overall use of emoticons between men and women was not evaluated as statistically significant by the regression model in the analysis.

As reported in previous literature, men made more explicit sexual references (Kapidzic & Herring, 2011; Subrahmanyan et al., 2006) and used a higher frequency of words suggesting power and status awareness (Ottoni et al., 2013). In addition, men used swear words with greater frequency, as found in previous research (e.g., Argamon et al., 2007; Fullwood et al., 2001; Ottoni et al., 2013; Savicki et al., 1996) and displayed more anger, consistent with the tendency reported in previous studies for men to "flame" (Herring, 1992, 1994). Men also referred to male entities more than women did, whereas women referred to female entities with a much higher frequency, as in previous studies (Herring, 1993, 2010). It is noteworthy, however, that women started talking more about males as time passed.

Women seemed to adhere to standard typography and orthography, based on the

higher percentage of words recognized as dictionary words by *LIWC2015*, although they had a higher frequency in the Informal Language category of *LIWC2015*, which includes popular abbreviations found in online communication. Women also used exclamation points with higher frequency than men, consistent with the findings of Waselesky (2006); however, the regression model did not evaluate the variation between men and women as significant.

Only one gender marker exhibited a gender pattern that differed saliently from the findings in previous literature: Adjectives were used with higher frequency by men in this dataset, in contrast to previous literature reporting higher use by women (Schwartz et al., 2013).

6.3. Changes and trends in gender patterns

When we revisit the research question posed in 4.1, it is clear that gender patterns have changed over time in USENET newsgroups. All linguistic features analyzed exhibit change over time, and this result is statistically significant for 24 of the 72 features modeled. Overall, women and men's use of common grammatical features converged over time, while affective processes, cognitive processes, biological processes, pronouns, and emoticons exhibited a reversal of patterns (convergence followed by divergence).

Examining the slopes of the trend lines allows for a more nuanced interpretation of the longitudinal trends in gender patterns. While the convergence of variables led to less differentiated or more "gender-neutral" language use, this usage is skewed toward one gender's linguistic style, rather than both genders converging toward the middle of the masculine-feminine spectrum. As seen in Table 22, the female trend lines had a steeper slope for most variables, which suggests that the frequencies in the use of those variables by females changed faster, in the process becoming more similar to the male frequencies. Even if the change in the use of a feature moves in the same direction for both genders (increases

or decreases over time), it can still shift toward the linguistic style of one gender, if the use by the other gender changes more rapidly toward the frequency of the former. For example, while the use of words expressing certainty increased for both genders in Figure 39, the slope of the trend line for females was more pronounced: Rather than meeting in the middle, as in the case of numbers in Figure 40, the female trend line moved toward the male trend line.

Many of the variables that exhibited a reversal of gender patterns had their reversal point (see Figure 4) toward the middle or the end of the time period. Some (e.g., tentativeness) reversed from a traditional to a non-traditional gender pattern over time. Additionally, the female trend lines in the majority of those variables had a steeper slope. If we take these pieces of information into account, it would seem that women's use of those variables changed faster until it became more similar to that of men, but then women moved *beyond* the masculine linguistic style along the gendered language spectrum. This *hypermasculinization* of features, as in the use of words related to money, achievements, and rewards as personal concerns and drives, is similar to the phenomenon of *hypercorrection*, the tendency to use "greater percentages of features ordinarily associated with a higher social class" (Mendoza-Denton, 2011, p. 183). Given that males generally use more words associated with money, achievement, and reward (Ottoni et al., 2013) and less tentative language (Fullwood et al., 2001; Herring et al., 1992), as well as the fact that men and their language have been traditionally associated with higher status and power (Coates, 1993), it may be that the women in the USENET newsgroups "over-use" such features to adopt the status and power associated with those features. Whatever the reason, the language in USENET newsgroups appears to have shifted toward a more masculine style over time, with hypermasculinization in the use of certain features by women.

This shift in the language use of the newsgroups could have its roots in external social factors related to changes in the balance of power and status between men and women in

society, such as the feminist movement that started in the 1970s. Internal social factors related to changes in the balance of power and status over time within the newsgroups could have played a role, as well. Understanding the reasons behind the changes identified here and the degree of influence, if any, of the computer medium, would require further research that would ideally include comparisons of online and offline data at similar time points. However, we can try to speculate about the reasons for the changes in certain features that show important changes in gender patterns based on the findings of previous literature, as well as the nature of the identified trends and the times when major changes occur.

One possible explanation for the shift toward a more masculine linguistic style over time in USENET newsgroups is the "List Effect" identified by Herring (1996): the principle "whereby the communicative practices of the majority of active participants become normative for the group as a whole" (p. 85). The Internet was strongly male-dominated until 2000, with participation in public discussion forums and newsgroups remaining mostly male even after that (Herring & Stoerger, 2014). The gender distribution in the data of this study reflect that: 87% of the posts were made by male users. With the exception of *net.kids/misc.kids*, which was presented as a case study of a female-predominant newsgroup in section 5.5, all the newsgroups in the dataset analyzed are male-predominant. Thus, a shift toward a more masculine linguistic style over time could be attributed to women adopting the communicative practices of the majority of active participants in the newsgroups: the men. This is further supported by the different evolution of language in the female-predominant newsgroup net.kids/misc.kids, where the use of gender markers diverged further over time. Men seemed to adopt the linguistic style of the majority of active participants in that newsgroup, the women, until they started engaging in more masculine linguistic behaviors suggestive of conflict (and possible harassment), while the women reverted toward a more traditional feminine linguistic style, resulting in what appeears to be a shift in power and

status within the newsgroup, as explained in section 5.5.

Another possible explanation for some of the changes in gender patterns occurring around 1995 is the popularization of the Internet, as discussed in section 6.1. Changes in the demographic make up of the newsgroups after the influx of new users – especially of younger users who did not have access to the Internet before that – could have brought with it different linguistic styles. While changes in the social structure within the newsgroups is an internal factor, the different linguistic styles of new users could be attributed to external social factors, i.e., larger changes in society. Other factors to consider include the effects of the feminist movement. For example, younger generations of women became more outspoken about their body and their health (e.g., Boston Women's Health Book Collective, 1976), a possible explanation for the reversal in patterns of words referring to biological processes. Additionally, the changes in the gender patterns for personal concerns and drives could be explained by the increasing number of women joining the workforce. The results of the analysis suggest that, over time, women started talking more about their achievements, were driven more by reward, power, and risk, and were more interested in and talked more about money – even more than men toward the end of the time period studied, resulting in a case of "hypermasculinization." Moreover, women's use of words related to home decreased over time, which could be interpreted in light of the same societal change, whereby working women spend less time at home.

In contrast to the early optimistic belief that online interactions would be more equal and democratic (Hauben & Hauben, 1997) and that CMC would be especially beneficial in leveling the linguistic inequalities women faced (Graddol & Swann, 1989), the findings of this study agree with Kling's suggestion that technology alone cannot be effective as an agent of social change without the requisite social conditions being met (1994). Rather, the changes that appear to level gender inequalities in the study, such as women's use of language

associated with power and status or their stepping out of expected language usage based on the social norms for their gender, can be attributed to internal social factors (female predominance in the newsgroup), as well as external social factors (change in the gender dynamics in society).

The majority of the findings in this study agree with the body of literature reporting that traditional gendered linguistic norms (Tannen, 1993, 1994, 2003) transferred from offline to online communication (Gregory, 1997; Hall, 1996; Herring, 1992a, 1992b, 1993, 1994, 1996b, 1996c, 2000, 2003a), and that power asymmetries have persisted (Gregory, 1997; Hall, 1996; Herring, 1996b). Males dominated the discussions in all the newsgroups except *net.kids/misc.kids*, consistent with the findings of Herring (1993, 2010) and Hert (1997). Moreover, the change in the gender patterns of linguistic features related to power and status at certain time points in *net.kids/misc.kids*, combined with the males' negative linguistic behavior, are suggestive of possible conflict and gender polarization, consistent with previous studies of online communication that associate aggressive linguistic behaviors with masculine linguistic styles (Gregory, 1997; Herring, 2003a; Guzzetti, 2008).

Conclusion

This exploratory study analyzed approximately four million posts from 47 newsgroups representing all major USENET hierarchies, in order to identify whether there were changes in the linguistic styles of women and men in CMD, as well as the manner in which they manifested, over a 25-year period. All the linguistic features examined exhibited change over time in their overall use, and the change was statistically significant for 63 out of the 72 features modeled. Similarly, there was change in the difference between the male and female frequencies for all the features over time that was statistically significant in 24 out of the 72 features. Using a language contact metaphor, the concepts of convergence and divergence were invoked to evaluate the directionality of change by examining the trend lines of the features. Although the language used in USENET newsgroups became less gender-differentiated over time, it shifted toward a more masculine linguistic style overall, with cases of "hypermasculinization" by women for linguistic features that are associated with status and power.

Since the newsgroups in the dataset were mainly male-predominant, the newsgroup net.kids/misc.kids was presented as a case study of the language used in a female-predominant newsgroup. The posts there exhibited different gender patterns compared to those of the male-predominant groups, as well as evidence of a possible shift in power and status that took place within the newsgroup, suggested by the use of language indicating conflict and polarization by men and the use of less powerful language by women. Even though identifying the reasons behind these gender pattern changes was not the focus or purpose of this study, an initial interpretation of the results suggests that the explanations lie in social factors, such as changes in the power and status of women in society (emancipation of women) and/or within the newsgroups (e.g., a "List Effect," Herring, 1996).

7.1. Implications

The results of this study are not only significant for research in Computer-Mediated Communication, but they have broader implications for a variety of areas, such as Social Informatics, Gender Studies, Historical Linguistics, Corpus Linguistics, and Natural Language Processing.

The study of gender in CMD is relevant to the fields of Social Informatics and Gender Studies, in that it provides insights into and a better understanding of the intricate workings of gender communication and expression in CMC. This enables researchers to identify issues such as inequality (including involving minority groups) and move towards solutions to rectify them. Through linguistic analysis, this study identified a possible shift in power and status in the female-predominant group of net.kids/misc.kids. As the case of the net.women/soc.women newsgroup suggests, women create online environments to avoid the behaviors associated with power and status inequalities they may experience in male-predominant groups (Hall, 1996), specifically flaming and harassment (Harter, 1998). The

posts in the newsgroup net.kids/misc.kids were more positive and friendly than the posts in the male-predominant newsgroups in the dataset. Moreover, women appeared to step out of gendered linguistic norms and expectations that are usual in a patriarchal society and talk about "taboo" topics (see use of sexual words). However, men appear to have eventually introduced masculine discourse behaviors that disrupted the power balance of the group: The sudden increase in the expression of anger could indicate conflict, trolling, flaming, or harassment. Such findings are not only important for policy-making and social movements such as feminism (Danet, 1998; Guzzeti, 2008; Hall, 1996; Herring et al., 1992, 1998; Marwick, 2013), but they can also be informative for Human-Computer Interaction (HCI) designers and web developers: Rather than excluding men completely from such online communities, a possible solution would be official mechanisms to report bad behaviors with appropriate repercussions, as well as algorithms that identify posts including flaming and harassment, something that a few online platforms have already started testing (e.g., Instagram, Holson, 2018).

Moreover, the findings of differences and similarities in masculine and feminine linguistic styles in this study can be used to improve the identification and prediction of user gender in social media using Natural Language Processing. There is increasing interest in author identification and author profiling (as seen in projects such as PAN[13]), in combination with the explosion of interest in Big Data for purposes of security and identifying deception, as well as for marketing via opinion/sentiment mining (Mukherjee & Liu, 2010; Sarawgi, Gajulapalli & Choi, 2011; Schwartz et al., 2013). This study analyzed a large, fairly representative dataset of online language, looking at over 72 different linguistic features. The results of these analyses could be used to improve the accuracy of gender predicting algorithms by adding new features or updating existing features over time or in different

[13] http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/

settings. Some of the variables offered by *LIWC2015* may have not been considered to be gender markers before, and variables whose gender patterns do not show changes in their longitudinal trends (for example, the 1st person singular pronoun) could be considered "persistent" gender markers and assigned higher weight in gender prediction algorithms.

The findings of this diachronic study have implications for Historical Linguistics and Corpus Linguistics, as well. This study has shown that CMD provides an excellent resource for studying language change, since there is a large amount of preserved, continuous, and contextualized data. Moreover, the results of the study support the suggestion that language change in online communication may happen faster or with greater intensity in a shorter period of time than in offline communication (Stein, 2006). Consequently, more initiatives are needed for collecting and including CMD in national corpora, especially since an increasingly large part of our communication happens through computer-mediated environments, including mobile devices. The methodology chapter of this study addressed several methodological challenges regarding diachronic research in CMC that may help anticipate or address issues in similar future endeavors. An additional benefit of the study to the areas of Historical Linguistics and Corpus linguistics is the creation of the *Historical USENET Newsgroups Corpus* (Bourlai & Gao, 2017), the first diachronic corpus of CMD that will become publicly available upon its completion. This corpus is expected to facilitate diachronic research in CMD and become a useful resource for research in a variety of areas.

7.2. Limitations

Several limitations of this study should be noted. First the small number of people having access to the Internet in earlier years, especially women, and the rise and fall in popularity of USENET newsgroups individually and as an online environment, resulted in a corpus that is not as balanced as might be desired across time, newsgroups, and genders. This may affect

not only the information that can be gained through analyzing the data, but also create challenges for the statistical analysis of the results. In addition to the proportion output of *LIWC2015* that created extreme values for certain variables in shorter posts (see section 4.5), this imbalance of data required a number of transformations in order for the author to be able to proceed with statistical testing.

The size of the dataset also poses some limitations in terms of the statistical interpretation of the results. While large datasets are preferred in order to reach valid conclusions, a dataset that is too large may result in false positives in terms of statistical significance. This was one of the reasons why the author decided to use the means per gender, newsgroup, and time point: This dramatically decreased the number of data points or observations from approximately four million to 639. However, it may have caused the opposite problem of false negatives: It is more difficult to find statistical significance in small datasets, which is why larger datasets are usually preferred, as mentioned above. Consequently, there may be more variables that show significant variation in their overall use over time, between men and women in general, or in their gender patterns over time, but the model failed to evaluate them as significant because there were fewer data points after the compression of the data. For example, the p-value in the difference in the use of adjectives between men and women ($p < .07$) is very close to the significance threshold of $p < .05$.

Moreover, there may still be some "noise" in the data after the cleansing process that could potentially affect certain variables. As explained in section 4.2.2, which describes the data cleansing and annotation process for HUNC, the original formatting of the posts was very heterogeneous due to technological changes over the time period covered. The size of the dataset allowed manual checking of only small samples in order to address the different formatting variations. Thus, there may be metadata about the post or possible attachments,

quoted text, or signatures at the end of the posts with quotes that may not have been included in the patterns identified to be removed with automated methods.

Another limitation of the study is the use of a binary, name-based approach to identify the gender of the authors in the data. Even though it was chosen as the most appropriate method due to the nature of this study and the size of the dataset, it oversimplifies the complex concept of gender identity. Moreover, the name-based approach relies on very little information to predict gender. If nicknames are used instead of actual names, it is very likely that they may not convey any gender association; there is also the case of anonymous authors. The gender of the authors in approximately 40% of the originally selected data was not identifiable using the automated method, which is the reason why participation was not analyzed as a gender marker in the study, since the dataset would not provide a complete picture of the phenomenon.

Finally, while this study does not explore the reasons behind the changes identified deeply, it provides some initial insights into gender communication and language change in online environments. Women and men communicate outside of cyberspace in the offline world, and there may be other factors not identifiable in this study that could lead to changes in their language use, and in language in general, over time. Addressing this would require further research, however, as explained in the next section.

7.3. Future Research

This study represents a first effort to systematically investigate gender and language in CMD from a diachronic perspective. Identifying changes in the linguistic styles of men and women and the manner in which these changes have manifested is the first step toward identifying and understanding the causes of those changes, and evaluating the effect of CMC on gendered communication and on language in general. The research would be further enriched

by adding data from more time points between the ones selected, in order to examine the identified changes in greater detail and to facilitate their interpretation.

A next step would be a comparison with a diachronic corpus comprising language from offline communication covering the same time period and sourcing data from the same time points. This would not only facilitate identifying differences between online and offline communication to help evaluate the effects of the medium on (gendered) language, but it would also help identify if the identified changes originated in offline or online communication, and whether they reflect deeper linguistic changes as opposed to changes in the discourse context.

Another potential line of research would be a diachronic study of CMD taking a non-binary gender approach, in order to address another gap in the literature of Historical Sociolinguistics and Gender Studies, as well as CMC. Since that would require manual identification of non-binary gender identities in the text of the posts, as well as a closer analysis of the post content, a smaller dataset focusing on newsgroups discussing related topics should be used.

Finally, a possible research project inspired by the findings of this study would be a diachronic analysis of female-predominant versus male-predominant newsgroups using a mixed quantitative and qualitative approach. If any power and status shifts are identified through the quantitative analysis of gendered language use within the groups, a qualitative approach would help uncover how and why they happen, and whether there is a way to ensure a democratic environment in online communication.

As the presence of CMD increases in everyday life and its role in interpersonal communications becomes greater, it is important to understand if and how it has changed the way we communicate as gendered selves and to use such insights to improve policy making, security, and user-oriented technology.

REFERENCES

Androutsopoulos, J. (2011). Language change and digital media: A review of conceptions and evidence. In T. Kristiansen & N. Coupland, N. (Eds.), *Standard languages and language standards in a changing Europe* (pp. 145-160). Olso: Novus Press.

Androutsopoulos, J., & Beißwenger, M. (2008). Introduction: Data and methods in computer-mediated discourse analysis. *Language@Internet, 5*, article 9. Retrieved from http://www.languageatinternet.org/articles/2008/1609

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text, 23*(3), 321-346.

Argamon, S., Koppel, M., Schier, J., & Pennebaker, J. W. (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday, 12*(9). Retrieved from http://pear.accc.uic.edu/ojs/index.php/fm/article/view/2003/1878

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

Baron, N. S. (1984). Computer mediated communication as a force in language change. *Visible Language*, *18*(2), 118-141.

Baron, N. S. (2004). See you online: Gender issues in college student use of Instant Messaging. *Journal of Language and Social Psychology*, *23*(4), 397-423. doi:10.1177/0261927X04269585

Baron, N. S. (2009). Are digital media changing language? *Educational Leadership*, *66*(6), 42-46.

Baumann, R. (2015). *Early Usenet history and archiving*. Retrieved March 11th, 2015 from https://ryanfb.github.io/etc/2015/02/23/early_usenet_history_and_archiving.html

Beißwenger, M., & Storrer, A. (2008). Corpora of computer-mediated communication. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 292-308). Berlin: Mouton de Gruyter.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. (2012). DeRiK: A German reference corpus of computer-mediated communication. In *Proceedings of Digital Humanities 2012*, University of Hamburg, July 16-22, Germany.

Beißwenger, M., Oostdijk, N., Storrer, A., & van den Heuvel, H. (2014). Building and annotating corpora of computer-mediated communication: Issues and challenges at the interface of corpus and computational linguistics. *Journal for Language Technology and Computational Linguistics*, *29*(2), iii-iv.

Berdicevskis, A. (2013). *Language change online: Linguistic innovations in Russian induced by computer-mediated communication*. Unpublished doctoral dissertation, University of Bergen, Norway.

Berdicevskis, A. (2014). The written turn: How CMC actuates linguistic change in Russian. In M. Gorham, I. Lunde, & M. Paulsen (Eds.), *Digital Russia: The culture, language and politics of new media communication* (pp. 107-122). New York: Routledge.

Boston Women's Health Book Collective (1976). *Our bodies, ourselves: A book by and for women.* New York : Simon and Schuster.

Bourlai, E. E. (2016, April). *Gender differences in soc.men and soc.women: A diachronic study.* Paper presented at Diachronic Corpora, Genre, and Language Change, April 8-9, University of Nottingham, UK.

Bourlai, E. E., & Gao, Z. (2017). *The Historical Usenet Newsgroups Corpus (HUNC): A diachronic CMC corpus.* Manuscript in preparation.

Braunmüller, K. H. J. (2009). *Convergence and divergence in language contact situations*. Amsterdam: John Benjamins Publishing Company.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Bruckman, A. S. 1993. Gender swapping on the Internet. *Proceedings of INET '93.* Reston, VA: The Internet Society.

Bucholtz, M. (2001). Geek feminism. In S. Benor, M. Rose, D. Sharma, J. Sweetland, & Q. Zhang (Eds.), *Gendered practices In language* (pp. 277-308). Stanford, CA: CSLI Publications.

Burger, J., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1301-1309). Stroudsburg, PA: ACL.

Bybee, J. L. (2015). *Language change*. Cambridge: Cambridge University Press, 2015.

Cameron, D. (1999). Performing gender identity: Young men's talk and the construction of heterosexual masculinity. In A. Jaworski & N. Coupland (Eds.), *The discourse reader* (pp. 442-458). London: Routledge.

Cameron, D. (2003). Gender issues in language change. *Annual Review of Applied Linguistics*, *23,* 187-201. doi:10.1017/S0267190503000266

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C., Hriba, L., Longhi, J., & Seddah, D. (2014). The CoMeRe corpus for French: Structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*, *29*(2), 1-30.

Cheshire, J., & Gardner-Chloros, P. (1998). Code-switching and the sociolinguistic gender pattern. *International Journal of the Sociology of Language,* 129-134. doi:10.1515/ijsl.1998.129.5

Coates, J. (1993). *Women, men, and language: A sociolinguistic account of gender differences in language*. London; New York: Longman.

Danet, B. (1998). Text as mask: Gender and identity on the Internet. In S. G. Jones (Ed.), *Cybersociety 2.0* (pp. 129–158). Thousand Oaks, CA: Sage.

Dzhingarov, B. (2018, May 1). A look into Usenet, the original wild west of the internet. *BestTechie.* Retrieved from https://www.besttechie.com/usenet-original-wild-west-of-the-internet/

de Oliveira, S. M. (2007). Breaking conversational norms on a Portuguese users' network: Men as adjudicators of politeness? In B. Danet & S. C. Herring (Eds.), *The multilingual Internet: Language, culture, and communication online* (pp. 256-277). Oxford: Oxford University Press.

D'Urso, S. C. (2009). The past, present, and future of human communication and technology research: An introduction. *Journal of Computer-Mediated Communication, 3*, 708-713. doi:10.1111/j.1083-6101.2009.01459.x

Eckert, P. (2011). Gender and sociolinguistic variation. In J. Coates & P. Pichler (Eds.), *Language and gender: A reader*, 2nd ed. (pp. 57-70). Oxford: Willy-Blackwell.

Edelsky, C. (1981). Who's got the floor? *Language in Society*, *3*, 383-421.

ELLO (English Language and Linguistics Online) (2018). *Gender Pattern: Language and Sex.* Retrieved September, 18, 2018 from http://www.ello.uos.de/field.php/Sociolinguistics/Genderpattern

Emerson, S. L. (October 1983). Usenet / A bulletin board for Unix users. *BYTE*, 219–236.

Facebook Diversity (2015). In *Facebook* [Group Page]. Retrieved September, 18, 2018, from https://www.facebook.com/facebookdiversity/posts/last-year-we-were-proud-to-add-a-custom-gender-option-to-help-people-better-expr/774221582674346/.

Freed, A. F. (1995). Language and gender. *Annual Review of Applied Linguistics*, *15*(1), 3-22. doi:10.1017/S0267190500002580

Fullwood, C., Morris, N., & Evans, L. (2011). Linguistic androgyny on MySpace. *Journal of Language and Social Psychology*, *30*(1), 114-124.

Gao, L. (2008). Language change in progress: Evidence from computer-mediated communication. *Proceedings of the 2oth North American Conference on Chinese Linguistics (NAACCL-20)* (pp. 361-377). Columbus, OH: The Ohio State University.

Graddol, D., & Swann, J. (1989). *Gender voices*. Oxford: Blackwell.

Gregory, M. Y. (1997). *Gender differences: An examination of computer-mediated communication*. Paper presented at the Annual Meeting of the Southern States Communication Association, April 2-6, 1997, Savannah, GA.

Guiller, J., & Durndell, A. (2007). Students' linguistic behavior in online discussion groups: Does gender matter? *Computers in Human Behavior*, *23*(5), 2240–2255.

Guzzetti, B. J. (2008). Identities in online communities: A young woman's critique of cyberculture. *E-Learning*, *5*(4), 457-474.

Hall, K. (1996). Cyberfeminism. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 147-170). Amsterdam: Benjamins. doi:10.1075/pbns.39.12hal

Harter, R. (1998). *The unofficial soc.women FAQ*. Retrieved from http://richardhartersworld.com/cri/1997/womenfaq.html

Hauben, M., & Hauben, R. (1997). *Netizens: On the history and impact of usenet and the internet*. Los Alamitos, CA: IEEE Computer Society Press.

Herring, S. C. (1992a). Gender and participation in computer-mediated linguistic discourse. Washington, D.C.: ERIC Clearinghouse on Languages and Linguistics, document ED345552.

Herring, S. C. (1992b). Men's language: A study of the discourse of the LINGUIST list. In A. Crochetière, J-C. Boulanger, & C. Ouellon (Eds.), *Les langues menacées: actes du XVe congrès international des linguistes*, Vol. 3 (pp. 347-350). Québec: Les Presses de l'Université Laval.

Herring, S. C. (1993). Gender and democracy in computer-mediated communication. *Electronic Journal of Communication*, *3*(2). Retrieved from http://www.cios.org/EJCPUBLIC/003/2/00328.HTML

Herring, S. C. (1994). Politeness in computer culture: Why women thank and men flame. In M. Bucholtz, A. Liang, L. Sutton, & C. Hines (Eds.), *Cultural performances: Proceedings of the Third Berkeley Women and Language Conference* (pp. 278–294). Berkeley, CA: Berkeley Women and Language Group.

Herring, S. C. (1996a). Introduction. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 1-10). Amsterdam: Benjamins.

Herring, S. C. (1996b). Two variants of an electronic message schema. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 81-108). Amsterdam: John Benjamins.

Herring, S. C. (1996c). Bringing familiar baggage to the new frontier: Gender differences in computer-mediated communication. In J. Selzer (Ed.), *Conversations* (pp. 1069-1082). Boston: Allyn & Bacon.

Herring, S. C. (1996d). Posting in a different voice: Gender and ethics in computer-mediated communication. In C. Ess (Ed.), *Philosophical perspectives on computer-mediated communication* (pp. 115-145). Albany: SUNY Press

Herring, S. C. (1998). Le style du courrier électronique: variabilité et changement. *Revue d'aménagement linguistique* (formerly *Terminogramme)*, *84-85* (March), 9-16.

Herring, S. C., Johnson, D. A., & DiBenedetto, T. (1998). Participation in electronic discourse in a "feminist" field. In J. Coates (Ed.), *Language and Gender: A Reader*. Oxford: Blackwell.

Herring, S. C. (1999). *Actualization of a counter-change: Contractions on the Internet.* Paper presented at the 14th International Conference on Historical Linguistics, August, 1999, Vancouver, Canada.

Herring, S. C. (2000). Gender differences in CMC: Findings and implications. *Computer Professionals for Social Responsibility Journal* (formerly *Computer Professionals for Social Responsibility Newsletter*), *18*(1). Retrieved from http://cpsr.org/issues/womenintech/herring/

Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 612-634). Oxford: Blackwell Publishers.

Herring, S. C. (2002). Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology, 36*, 109-168.

Herring, S. C. (2003a). Gender and power in online communication. In J. Holmes & M. Meyerhoff (Eds.), *The handbook of language and gender* (pp. 202–228). Oxford:

Blackwell.

Herring, S. C., Ed. (2003b). *Media and language change.* Special issue of the *Journal of Historical Pragmatics, 4*(1), 1-17.

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

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet, 4,* article 1. Retrieved from http://www.languageatinternet.org/articles/2007/761

Herring, S. C. (2010). Who's got the floor in computer-mediated conversation? Edelsky's gender patterns revisited. *Language@Internet, 7,* article 8. Retrieved from http://www.languageatinternet.org/articles/2010/2857

Herring, S. C., Johnson, D. A., & DiBenedetto, T. (1992). Participation in electronic discourse in a 'feminist' field. In *Locating Power: Proceedings of the 1992 Berkeley Women and Language Conference* (pp. 250-262). Berkeley: Berkeley Women and Language Group.

Herring, S. C., Johnson, D. A., & DiBenedetto, T. (1995). 'This discussion is going too far!' Male resistance to female participation on the Internet. In M. Bucholtz & K. Hall (Eds.), *Gender articulated: Language and the socially constructed self* (pp. 67-96). New York: Routledge.

Herring, S. C., & Martinson, A. (2004). Assessing gender authenticity in computer-mediated language use: Evidence from an identity game. *Journal of Language and Social Psychology, 23* (4), 424-446.

Herring, S. C., & Panyametheekul, S. (2003). Gender and turn-allocation in a Thai chat room. *Journal of Computer-Mediated Communication, 9*(1). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2003.tb00362.x/full

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439-459.

Herring, S. C., & Stoerger, S. (2014). Gender and (a)nonymity in computer-mediated communication. In S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The handbook of language, gender, and sexuality*, 2nd edition (pp. 567-586). Chichester, UK: John Wiley & Sons, Ltd.

Hert, P. (1997). Social dynamics of an on-line scholarly debate. *The Information Society, 13*, 329-360.

Holmes, J. (2007). Social constructionism, postmodernism and feminist sociolinguistics. *Gender & Language, 1*(1), 51. doi:10.1558/genl.2007.1.1.51

Holson, L. M. (2018, May 1). Instagram unveils a bully filter. *The New York Times*. Retrieved from https://www.nytimes.com/2018/05/01/technology/instagram-bully-filter.html

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00238.x/full

Huh, S., & Williams, D. (2009). Dude looks like a lady: Gender swapping in an online game. In W. S. Bainbridge (Ed.), *Online worlds: Convergence of the real and the virtual* (pp. 161-174). London, UK: Springer.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kaggle (2018). *Kaggle: Your Home For Data Science*. Retrieved from https://www.kaggle.com

Kapidzic, S., & Herring, S. C. (2011). Gender, communication, and self-presentation in teen chatrooms revisited: Have patterns changed? *Journal of Computer-Mediated Communication, 17*(1), 39-59.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53,* 59-68. doi:10.1016/j.bushor.2009.09.003

Kehoe, B. P. (1996). *Zen and the art of the Internet : a beginner's guide.* Upper Saddle River, N.J. : Prentice Hall PTR.

Kennedy, G. (1998). *An introduction to corpus linguistics.* London: Longman.

King, B. (2009). Building and analyzing corpora of computer-mediated communication. In P. Baker (Ed.), *Contemporary studies in linguistics: Contemporary corpus linguistics* (pp. 301-320). London: Continuum International Publishing.

King, B. (2015). Investigating digital sex talk practices: A reflection on corpus-assisted discourse analysis. In R. H. Jones, A. Chik, & C. A. Hafner (Eds.), *Discourse and digital practices: Doing discourse analysis in the digital age* (130-143). London: Routledge.

Kling, R. (1994). Reading "all about" computerization: How genre conventions shape nonfiction social analysis. *The Information Society, 10*(3), 147-172. doi:10.1080/01972243.1994.9960166

Labov, W. (1963). The social motivation of language change. *Word, 19,* 273–309.

Labov, W. (1966). Hypercorrection by the lower middle class as a factor in linguistic change. In W. Bright (Ed.), *Sociolinguistics: Proceedings of the UCLA Sociolinguistics Conference, 1964* (pp. 84–101). The Hague: Mouton.

Laniado, D., Kaltenbrunner, A., Castillo, C., & Morell, M. F. (2012). Emotions and dialogue in a peer-production community: The case of Wikipedia. *Wikisym 2012 Conference Proceedings – 8th Annual International Symposium on Wikis and Open Collaboration*, (WikiSym 2012 Conference Proceedings - 8th Annual International Symposium on Wikis and Open Collaboration), doi:10.1145/2462932.2462944

Las Casas, D. C. , Magno, G., Cunha, E., Gonçalves, M. A., Cambraia, C., & Almeida, V. (2014). Noticing the other gender on Google+. *Proceedings of the 2014 ACM conference on Web science (WebSci '14).* New York, NY: ACM. *DOI=http://dx.doi.org/10.1145/2615569.2615692*

Leeds-Hurwitz, W. (2009). Social construction of reality. In S. Littlejohn, & K. Foss (Eds.), *Encyclopedia of communication theory* (Vol. 1, pp. 892-894). Thousand Oaks, CA: SAGE Publications, Inc.

LIWC (2017). *Interpreting LIWC Output* page. Retrieved on March 16th, 2017 from https://liwc.wpengine.com/interpreting-liwc-output/

Margaretha, E., & Lüngen, H. (2014). Building linguistic corpora from Wikipedia articles and discussions. *Journal for Language Technology and Computational Linguistics*, *29*(2), 59-82.

Marwick, A. (2013). Donglegate: Why the tech community hates feminists. *Wired*. http://www.wired.com/opinion/2013/03/richards-affair-and-misogyny-in-tech/

Megaputer Intelligence Inc. (2018). *PolyAnalyst*. Retrieved from https://www.megaputer.com/polyanalyst/

Mendoza-Denton, N. (2011). Individuals and communities. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The SAGE handbook of sociolinguistics* (pp. 279-295). London: SAGE.

Mislove, A., Lehmann, S., Ahn, Y-Y., Onnela, J-K., & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. *5th International AAAI Conference on Web and Social Media*. Retrieved February 25, 2017 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816

Mukherjee, A., & Liu, B. (2010). Improving gender classification of blog authors. *EMNLP 2010 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (*EMNLP 2010*), 207-217.

Nevalainen, T. (2011). Historical sociolinguistics. In R. Wodak, B. Johnstone, & P. Kerswill (Eds.), *The SAGE handbook of sociolinguistics* (pp. 279-295). London: SAGE.

Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y-Y., Kapadia, A., & Rojas, F. (2016). Twitter's glass ceiling: The effect of perceived gender on online visibility. *The 10th International AAAI Conference On Web And Social Media (ICWSM-16*). Retrieved February 25, 2017 from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13003

Ottoni, R., Pesce, J. P., Las Casas, D., Franciscani Jr., G., Meira Jr., W., Kumaraguru, P., & Almeida, V. (2013). Ladies first: Analyzing gender roles and behaviors in Pinterest. *Proceedings of the 7Th International Conference on Weblogs and Social Media, ICWSM 2013*, (457-465). Cambridge, MA: AAAI Publications.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Press, L. (1994). Commercialization of the Internet. *Communications of the ACM, 37*(11), 17–21. https://doi-org.proxyiub.uits.iu.edu/10.1145/188280.188286

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: http://www.R-project.org/.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in Twitter. *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents* (37-44). doi:10.1145/1871985.1871993

Rodino, M. (1997). Breaking out of binaries: Reconceptualizing gender and its relationship to language in computer-mediated communication. *Journal of Computer-Mediated Communication, 3*(3). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00074.x/full

Rowe, C. (2011). Whatchanade? Rapid language change in a private email sibling code. *Language@internet*, 8, article 6. Retrieved from http://www.languageatinternet.org/articles/2011/Rowe

RStudio Team (2016). *RStudio: Integrated Development for R.* RStudio, Inc., Boston, MA. Retrieved from http://www.rstudio.com/

Sarawgi, R., Gajulapalli, K., & Choi, Y. (2011). Gender attribution: Tracing stylometric evidence beyond topic and genre. *Conll 2011 - Fifteenth Conference On Computational Natural Language Learning, Proceedings of the Conference* (*CoNLL 2011*), 78-86.

Savicki, V., Lingenfelter, D., & Kelley, M. (1996). Gender language style and group composition in internet discussion groups. *Journal of Computer-Mediated Communication*, *2*(3). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1996.tb00191.x/abstract

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *Plos ONE*, *8*(9), 1-16. doi:10.1371/journal.pone.0073791

Selfe, C. L., & Meyer, P. R. (1991). Testing claims for on-line conferences. *Written*

Communication, 8(2), 163–192.

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? In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 517-522). ACM.

Spender, D. (1998). *Man made language*. NY: New York University Press.

Stark, E., Ruef, B., & Ueberwasser, S. (2015). *Swiss SMS Corpus.* Zurich: University of Zurich. Retrieved April 3, 2017 from http://www.zora.uzh.ch/109703/

Stein, D. (2006). Language and Internet. Email, Internet, chatroom talk: Pragmatics. In E. K. Brown (Ed.), *Encyclopedia of language & linguistics,* 2nd ed. (pp. 116–124). Oxford, UK: Elsevier.

Subrahmanyam, K., Greenfield, P. M., & Tynes, B. (2004). Constructing sexuality and identity in an online teen chat room. *Journal of Applied Developmental Psychology, 25*(6), 651–666.

Subrahrnanyam, K., Smahel, D., & Greenfield, P. (2006). Connecting developmental constructions to the Internet: Identity presentation and sexual exploration in online teen chat rooms. *Developmental Psychology*, *42*(3), 395-406. doi:10.1037/0012-1349.42.3.395

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

Tannen, D. (Ed.) (1993). *Framing in discourse*. NY: Oxford University Press.

Tannen, D. (1994). *Gender and discourse*. New York: Oxford University Press.

Tannen, D. (2006). Language and culture. In R. W. Fasold & J. Connor Linton (Eds.), *An introduction to language and linguistics* (343-372). Cambridge: Cambridge University Press.

Thelwall, M. (2008). Fk yea I swear: Cursing and gender in MySpace. *Corpora: Corpus-Based Language Learning, Language Processing and Linguistics, 3*(1), 83-107. doi:10.3366/E1749503208000087

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science & Technology, 61*(1), 190-199. doi:10.1002/asi.21180

The R Stats Package (2018). *Fitting linear Models.* Retrieved July 10, 2018 from: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html

Thomson, R. (2006). The effect of topic of discussion on gendered language in computer-mediated communication discussion. *Journal of Language & Social Psychology, 25*(2), 167-178.

Usenet Archive (2017). *About* Page. Retrieved February 25, 2017 from https://archive.org/details/usenet&tab=about

van Gass, K. M. (2008). Language contact in computer-mediated communication: Afrikaans-English code switching on internet relay chat [IRC]. *Southern African Linguistics and Applied Language Studies*, *26*(4), 429-444.

Walton, S., C., & Rice, R. E. (2013). Mediated disclosure on Twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage. *Computers in Human Behavior, 29*, 1465-1474. doi:10.1016/j.chb.2013.01.033

Wang, Y. C., Burke, M., & Kraut, R. E. (2013). Gender, topic, and audience response: An analysis of user-generated content on Facebook. *Conference on Human Factors in Computing Systems - Proceedings*, 31-34. doi:10.1145/2470654.2470659

Waseleski, C. (2006). Gender and the use of exclamation points in computer-mediated communication: An analysis of exclamations posted to two electronic discussion lists. *Journal of Computer-Mediated Communication, 11*(4). Retrieved from

http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2006.00305.x/full

Weis, A. H. (2010). Commercialization of the Internet. *Internet Research, 20*(4), 420-435. https://doi-org.proxyiub.uits.iu.edu/10.1108/10662241011059453

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wikipedia (2017a). *Big 8 (Usenet).* Retrieved February 25, 2017 from https://en.wikipedia.org/wiki/Big_8_(Usenet)

Wikipedia(2018). *List of Emoticons.* Retrieved November 12, 2017 from https://en.wikipedia.org/wiki/List_of_emoticons

Wikipedia (2017b). *Usenet*. Retrieved February 25, 2017 from https://en.wikipedia.org/wiki/Usenet

Wikipedia (2017c). *Usenet newsgroups*. Retrieved February 25, 2017 from https://en.wikipedia.org/wiki/Usenet_newsgroup

Witmer, D. F., & 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). http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00192.x/abstract

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

Yates, S. J. (1997). Gender, identity and CMC. *Journal of Computer Assisted Learning*, *13*(4), 281-290.

APPENDICES

Appendix A: Indiana University IRB approval

Protocol #	1603328428
Protocol Type	Exempt
Protocol Status	Exempt
Title	Gender and Language in CMC: A Historical Analysis of Usenet Newgroups
Summary/Keywords	
Initial Submission Date	4/1/16
Approval Date	4/11/16
Expiration Date	
Last Approval Date	
Investigator	Herring, Susan Catherine
Lead Unit	BL-SLIS
Lead Unit Name	INFORMATION AND LIBRARY SCIENCE
Active	Yes

Appendix B: Detailed list of newsgroups used in the study, number of posts by time point and gender, and total number of posts.

USENET Newsgroups	Number of Posts per Time Point														Total (Group)
	1982 - 1983		1986 - 1987		1990 - 1991		1994 - 1995		1998 - 1999		2002-2003		2006-2007		
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	
net.flame / alt.flame	80	593	1	7	446	3,496	4,818	26,931	5,701	60,023	4,816	25,204	2,301	13,231	147,648
net.arch / comp.arch	-	50	17	2,329	67	9,728	192	10,695	249	16,239	846	25,598	863	10,288	77,161
net.dcom / comp.dcom.modems	10	96	22	860	68	5,905	1,107	44,493	1,485	42,288	202	5,580	13	563	102,692
net.lang.ada / comp.lang.ada	-	25	22	633	100	2,385	326	15,684	336	14,839	259	20,455	107	8,864	64,035
net.lang.apl / comp.lang.apl	-	18	5	54	5	502	48	2,626	23	3,326	71	2,521	51	1,216	10,466
net.lang.forth / comp.lang.forth	-	8	1	252	3	1,965	227	5,778	845	13,566	1,152	16,330	759	11,147	52,033
net.lang.lisp / comp.lang.lisp	1	33	28	511	70	2,228	133	5,437	272	15,571	440	37,900	406	34,587	97,617
net.lang / comp.lang.misc	5	153	33	803	32	3,880	112	5,783	103	2,729	53	2,133	59	2,084	17,962
net.lang.prolog / comp.lang.prolog	-	28	11	382	44	1,470	126	2,693	141	3,703	152	3,253	50	2,000	14,053
net.lsi / comp.lsi	-	10	-	145	9	436	52	1,518	9	575	4	143	-	2	2,903
net.periphs / comp.periphs	3	90	7	482	22	1,157	231	3,648	298	4,441	31	588	40	56	11,094
net.kids / misc.kids	9	20	87	179	5,712	4,283	55,205	28,801	28,026	19,946	43,412	15,106	13,366	6,105	220,257

USENET Newsgroups	Number of Posts per Time Point														Total (Group)
	1982 - 1983		1986 - 1987		1990 - 1991		1994 - 1995		1998 - 1999		2002-2003		2006-2007		
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	
net.legal / misc.legal	12	75	29	511	3	39	1,874	28,100	3,666	34,456	4,073	36,683	2,729	33,363	145,613
net.misc / misc.misc	106	931	43	691	169	2,455	791	7,030	795	4,323	490	9,230	296	8,050	35,400
net.taxes / misc.taxes	4	54	4	185	63	933	1,866	14,903	1,315	19,567	1,417	25,328	454	14,066	80,159
net.wanted / misc.wanted	39	549	35	958	221	3,967	1,280	8,915	258	1,510	91	287	5	11	18,126
net.news.b / news.software.b	32	142	6	667	120	4,573	50	2,709	18	147	10	26	4	5	8,509
net.books / rec.arts.books	13	80	23	322	3,001	12,350	12,245	38,474	10,951	44,731	2,647	22,679	1,318	15,561	164,395
net.tv.drwho / rec.arts.drwho	1	8	16	131	13	91	5,592	36,710	9,105	124,728	8,157	49,463	3,528	37,358	274,901
net.tv / rec.arts.tv	50	267	19	228	2	24	18	53	53	82	35	146	25	503	1,505
net.rec.birds / rec.birds	1	22	12	55	432	1,683	2,568	9,896	8,318	14,869	3,176	12,258	3,195	7,983	64,468
net.veg / rec.food.veg	-	1	57	148	1,636	3,078	3,192	16,235	3,240	9,618	559	2,020	200	448	40,432
net.rec.bridge / rec.games.bridge	5	38	1	25	184	3,046	1,108	17,417	1,295	27,699	1,875	48,451	1,164	33,995	136,303
net.games.emp / rec.games.empire	18	60	1	46	57	2,458	61	2,947	130	1,670	175	1,555	41	215	9,434

USENET Newsgroups	Number of Posts per Time Point														Total (Group)
	1982 - 1983		1986 - 1987		1990 - 1991		1994 - 1995		1998 - 1999		2002-2003		2006-2007		
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	
net.games / rec.games.misc	33	115	18	236	352	7,253	475	4,953	139	1,627	33	405	6	52	15,697
net.games.pbm / rec.games.pbm	-	13	1	101	75	1,067	418	5,845	934	4,825	189	1,493	51	428	15,440
net.games.trivia / rec.games.trivia	9	173	8	121	243	1,922	1,362	7,032	645	3,995	191	2,619	48	3,227	21,595
net.puzzle / rec.puzzles	1	26	2	188	146	3,954	955	14,445	1,775	24,249	916	24,944	579	12,807	84,987
net.pets / rec.pets	16	35	53	104	3,264	3,410	5,291	6,402	2,397	2,935	611	744	66	394	25,722
net.rec.scuba / rec.scuba	-	12	1	21	220	4,081	3,410	32,480	3,165	48,839	5,472	76,716	489	25,311	200,217
net.rec.skydive / rec.skydiving	-	6	-	24	95	1,117	1,495	15,453	9,201	40,492	3,267	31,500	1,609	3,898	108,157
net.sport.baseball / rec.sports.baseball	-	65	15	87	1,039	18,328	2,126	57,250	2,116	56,298	2,471	51,713	319	11,015	202,842
net.sport.hockey / rec.sports.hockey	-	21	1	88	614	8,503	3,552	50,934	879	18,746	125	4,600	1,429	4,259	93,751
net.video / rec.video	6	34	9	351	117	8,385	806	20,993	907	26,471	480	11,726	74	1,729	72,088
net.astro / sci.astro	-	10	44	157	121	7,532	1,105	38,287	1,784	44,370	1,825	44,894	1,118	30,859	172,106
net.nlang / sci.lang	31	325	31	375	339	4,799	684	11,116	1,252	37,783	1,105	40,168	2,986	42,054	143,048

USENET Newsgroups	Number of Posts per Time Point														Total (Group)
	1982 - 1983		1986 - 1987		1990 - 1991		1994 - 1995		1998 - 1999		2002-2003		2006-2007		
	F	M	F	M	F	M	F	M	F	M	F	M	F	M	
net.math / sci.math	28	234	2	231	-	6	1	113	4	79	47	642	4,955	98,309	104,651
net.research / sci.research	-	2	10	190	32	672	290	2,697	130	1,395	61	970	6	381	6,836
net.columbia / sci.space.shuttle	251	65	18	691	126	3,429	568	13,119	693	11,928	3,092	47,222	848	13,823	95,873
net.college / soc.college	1	31	33	242	192	2,038	699	6,381	571	2,118	383	949	4	80	13,722
net.social / soc.misc	10	31	48	136	30	503	70	551	87	702	163	625	6	33	2,995
net.motss / soc.motss	10	112	29	263	2,995	22,070	5,921	33,470	9,722	48,839	8,397	45,244	3,491	18,655	199,218
net.singles / soc.singles	92	350	314	898	2,632	12,711	6,245	14,926	9,971	35,828	17,236	38,757	2,504	9,137	151,601
net.women / soc.women	141	378	223	613	2,284	5,906	4,893	16,586	6,602	17,350	8,080	20,258	1,494	7,921	92,729
net.bizarre / talk.bizarre	-	3	15	254	197	3,163	1,582	16,971	2,708	13,055	1,575	16,767	1,643	14,483	72,416
net.religion / talk.religion.misc	43	432	39	505	136	2,366	1,342	17,019	4,248	39,070	1,944	16,300	406	3,967	87,817
net.rumor / talk.rumors	2	36	22	532	6	189	365	3,261	227	1,614	348	2,385	16	25	9,028
Total (Time Point)	1,063	5,860	1,416	17,012	27,734	195,536	136,877	727,760	136,789	963,254	132,154	844,578	55,121	544,548	3,789,702

Appendix C: List of categories/variables in LIWC2015

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Word count	WC	-	-	-	-
Summary Language Variables					
Analytical thinking	Analytic	-	-	-	-
Clout	Clout	-	-	-	-
Authentic	Authentic	-	-	-	-
Emotional tone	Tone	-	-	-	-
Words/sentence	WPS	-	-	-	-
Words > 6 letters	Sixltr	-	-	-	-
Dictionary words	Dic	-	-	-	-
Linguistic Dimensions					
Total function words	funct	it, to, no, very	491	.05	.24
Total pronouns	pronoun	I, them, itself	153	.25	.67
Personal pronouns	ppron	I, them, her	93	.20	.61
1st pers singular	i	I, me, mine	24	.41	.81
1st pers plural	we	we, us, our	12	.43	.82
2nd person	you	you, your, thou	30	.28	.70
3rd pers singular	shehe	she, her, him	17	.49	.85
3rd pers plural	they	they, their, they'd	11	.37	.78
Impersonal pronouns	ipron	it, it's, those	59	.28	.71
Articles	article	a, an, the	3	.05	.23
Prepositions	prep	to, with, above	74	.04	.18
Auxiliary verbs	auxverb	am, will, have	141	.16	.54
Common Adverbs	adverb	very, really	140	.43	.82
Conjunctions	conj	and, but, whereas	43	.14	.50
Negations	negate	no, not, never	62	.29	.71
Other Grammar					
Common verbs	verb	eat, come, carry	1000	.05	.23
Common adjectives	adj	free, happy, long	764	.04	.19
Comparisons	compare	greater, best, after	317	.08	.35
Interrogatives	interrog	how, when, what	48	.18	.57
Numbers	number	second, thousand	36	.45	.83
Quantifiers	quant	few, many, much	77	.23	.64
Psychological Processes					
Affective processes	affect	happy, cried	1393	.18	.57
Positive emotion	posemo	love, nice, sweet	620	.23	.64
Negative emotion	negemo	hurt, ugly, nasty	744	.17	.55

Anxiety	anx	worried, fearful	116	.31	.73
Anger	anger	hate, kill, annoyed	230	.16	.53
Sadness	sad	crying, grief, sad	136	.28	.70
Social processes	social	mate, talk, they	756	.51	.86
Family	family	daughter, dad, aunt	118	.55	.88
Friends	friend	buddy, neighbor	95	.20	.60
Female references	female	girl, her, mom	124	.53	.87
Male references	male	boy, his, dad	116	.52	.87
Cognitive processes	cogproc	cause, know, ought	797	.65	.92
Insight	insight	think, know	259	.47	.84
Causation	cause	because, effect	135	.26	.67
Discrepancy	discrep	should, would	83	.34	.76
Tentative	tentat	maybe, perhaps	178	.44	.83
Certainty	certain	always, never	113	.31	.73
Differentiation	differ	hasn't, but, else	81	.38	.78
Perceptual processes	percept	look, heard, feeling	436	.17	.55
See	see	view, saw, seen	126	.46	.84
Hear	hear	listen, hearing	93	.27	.69
Feel	feel	feels, touch	128	.24	.65
Biological processes	bio	eat, blood, pain	748	.29	.71
Body	body	cheek, hands, spit	215	.52	.87
Health	health	clinic, flu, pill	294	.09	.37
Sexual	sexual	horny, love, incest	131	.37	.78
Ingestion	ingest	dish, eat, pizza	184	.67	.92
Drives	drives		1103	.39	.80
Affiliation	affiliation	ally, friend, social	248	.40	.80
Achievement	achieve	win, success, better	213	.41	.81
Power	power	superior, bully	518	.35	.76
Reward	reward	take, prize, benefit	120	.27	.69
Risk	risk	danger, doubt	103	.26	.68
Time orientations	TimeOrie				
Past focus	focuspast	ago, did, talked	341	.23	.64
Present focus	focuspres	today, is, now	424	.24	.66
Future focus	focusfutur	may, will, soon	97	.26	.68
Relativity	relativ	area, bend, exit	974	.50	.86
Motion	motion	arrive, car, go	325	.36	.77
Space	space	down, in, thin	360	.45	.83
Time	time	end, until, season	310	.39	.79
Personal concerns					
Work	work	job, majors, xerox	444	.69	.93
Leisure	leisure	cook, chat, movie	296	.50	.86
Home	home	kitchen, landlord	100	.46	.83
Money	money	audit, cash, owe	226	.60	.90
Religion	relig	altar, church	174	.64	.91

Death	death	bury, coffin, kill	74	.39	.79
Informal language	informal		380	.46	.84
Swear words	swear	fuck, damn, shit	131	.45	.83
Netspeak	netspeak	btw, lol, thx	209	.42	.82
Assent	assent	agree, OK, yes	36	.10	.39
Nonfluencies	nonflu	er, hm, umm	19	.27	.69
Fillers	filler	I mean, you know	14	.06	.27
Punctuation					
All Punctuation	Allpunc	-	-	-	-
Periods	Period	-	-	-	-
Commas	Comma	-	-	-	-
Colons	Colon	-	-	-	-
Semicolons	SemiC	-	-	-	-
Question marks	QMark	-	-	-	-
Exclamation marks	Exclam	-	-	-	-
Dashes	Dash	-	-	-	-
Quotation marks	Quote	-	-	-	-
Apostrophes	Apostro	-	-	-	-
Parentheses (pairs)	Parenth	-	-	-	-
Other punctuation	OtherP	-	-	-	-

Note. Modified from "*The development and psychometric properties of LIWC2015"* by J.W. Pennebaker et al. (2015).

Appendix D: Figures showing the mean frequencies and trend lines of the studied variables by gender.

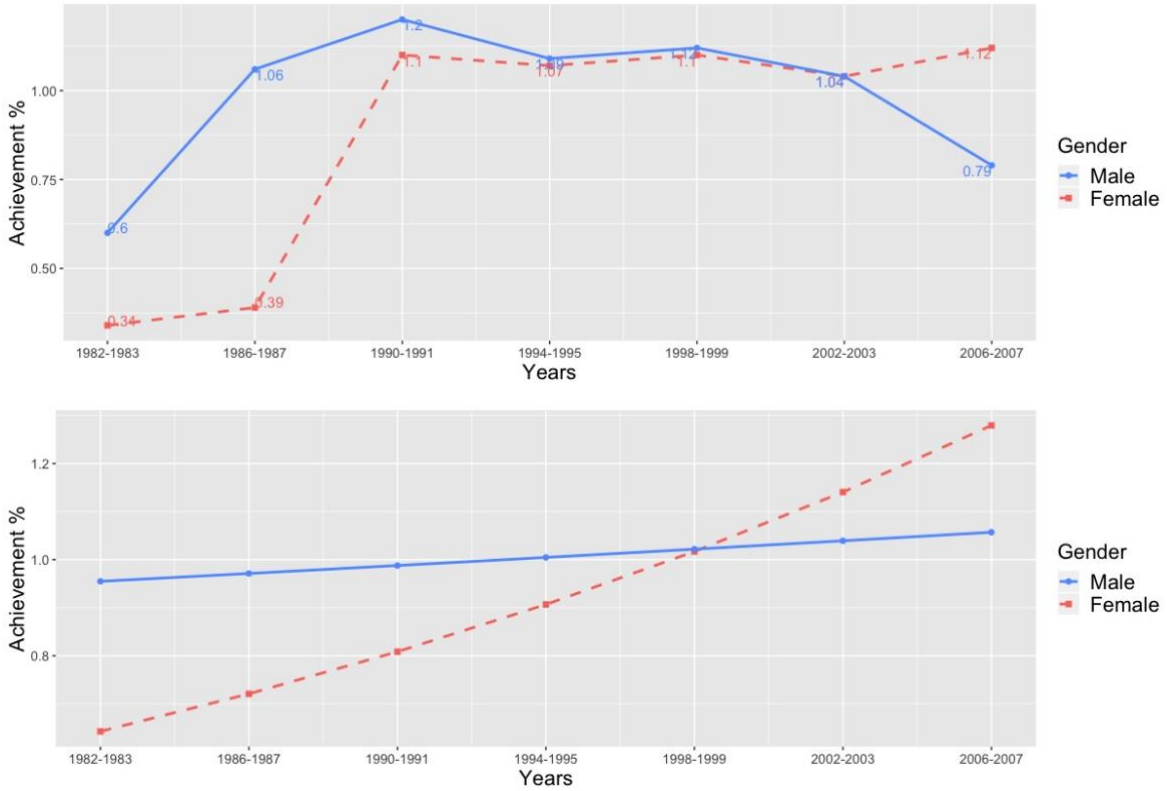
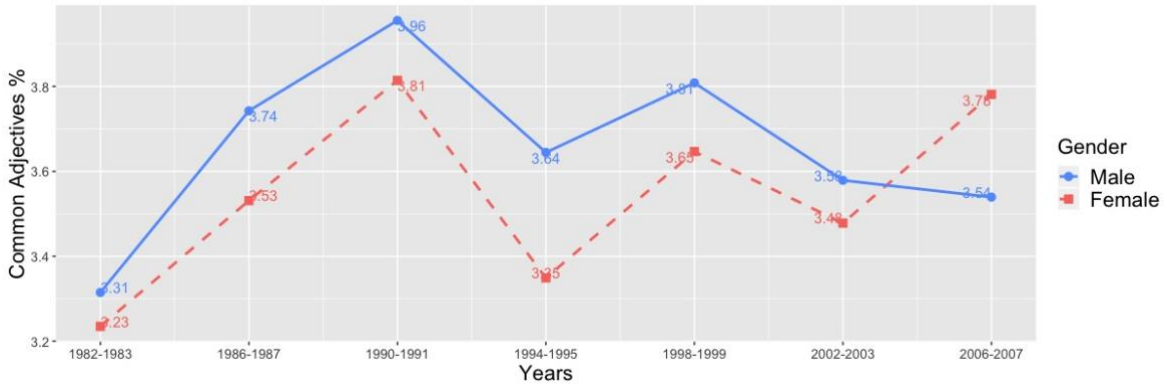


Figure D1. **Frequencies and trend lines of Achievement by gender over time.**



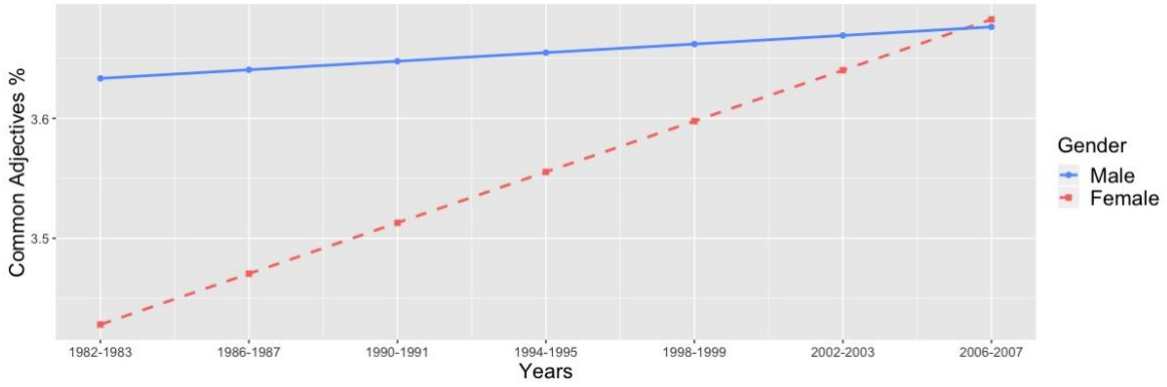


Figure D2. **Frequencies and trend lines of Adjectives by gender over time.**

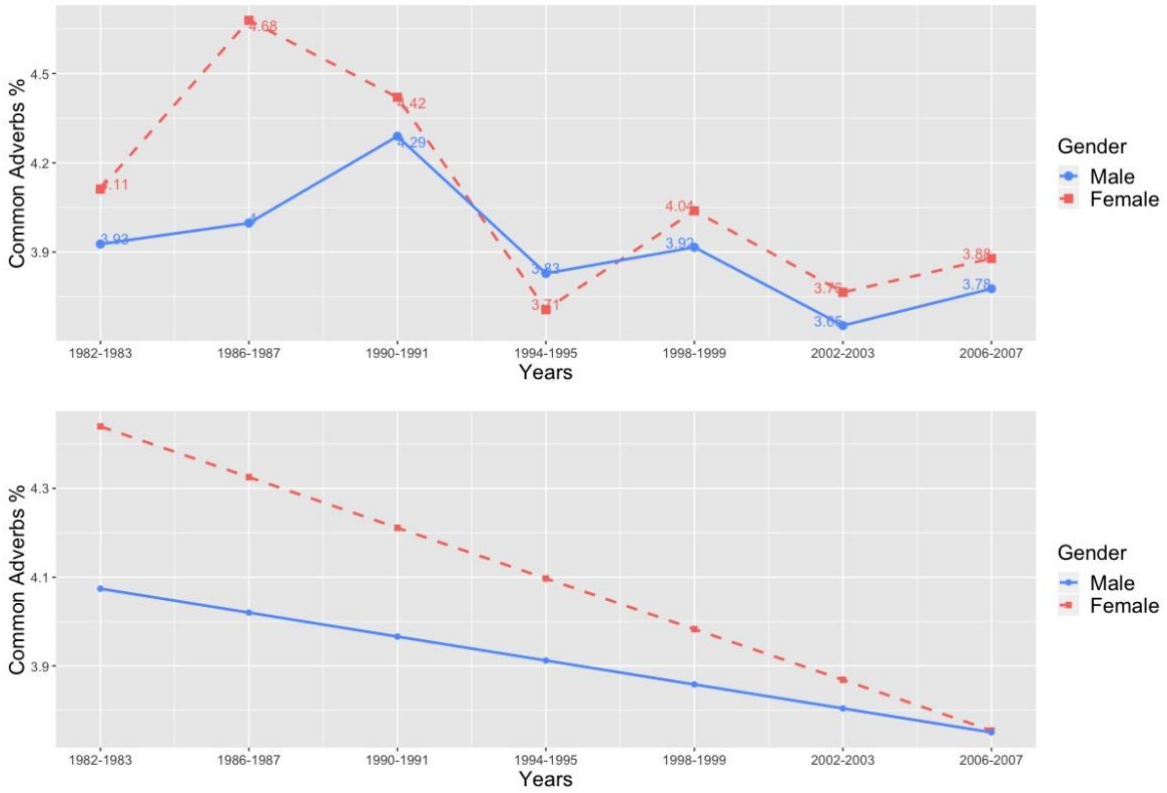
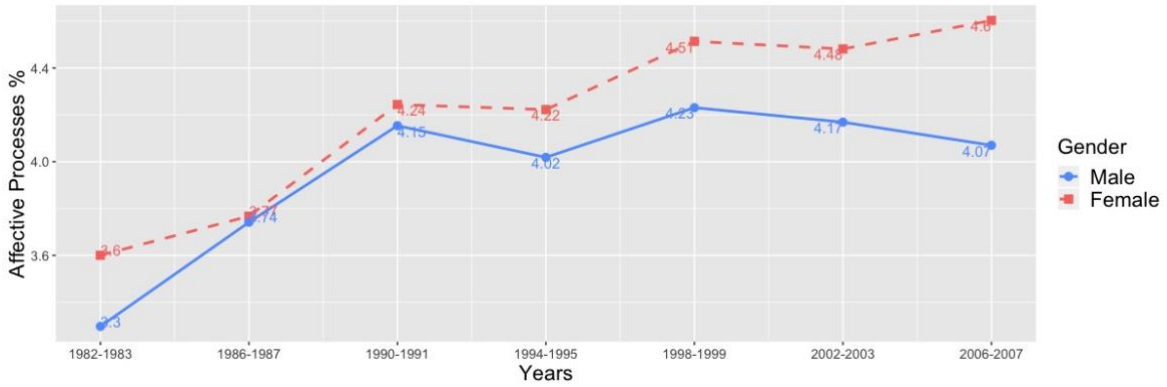


Figure D3. **Frequencies and trend lines of Adverbs by gender over time.**



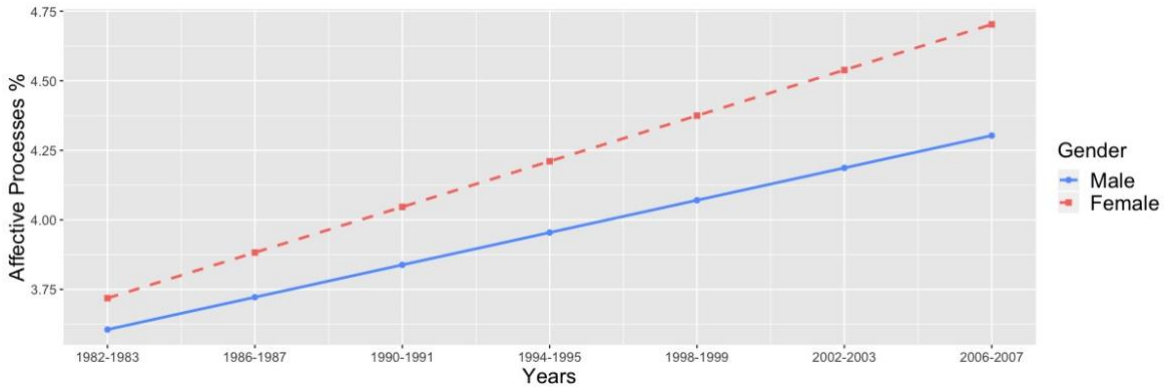


Figure D4. **Frequencies and trend lines of Affective Processes by gender over time.**

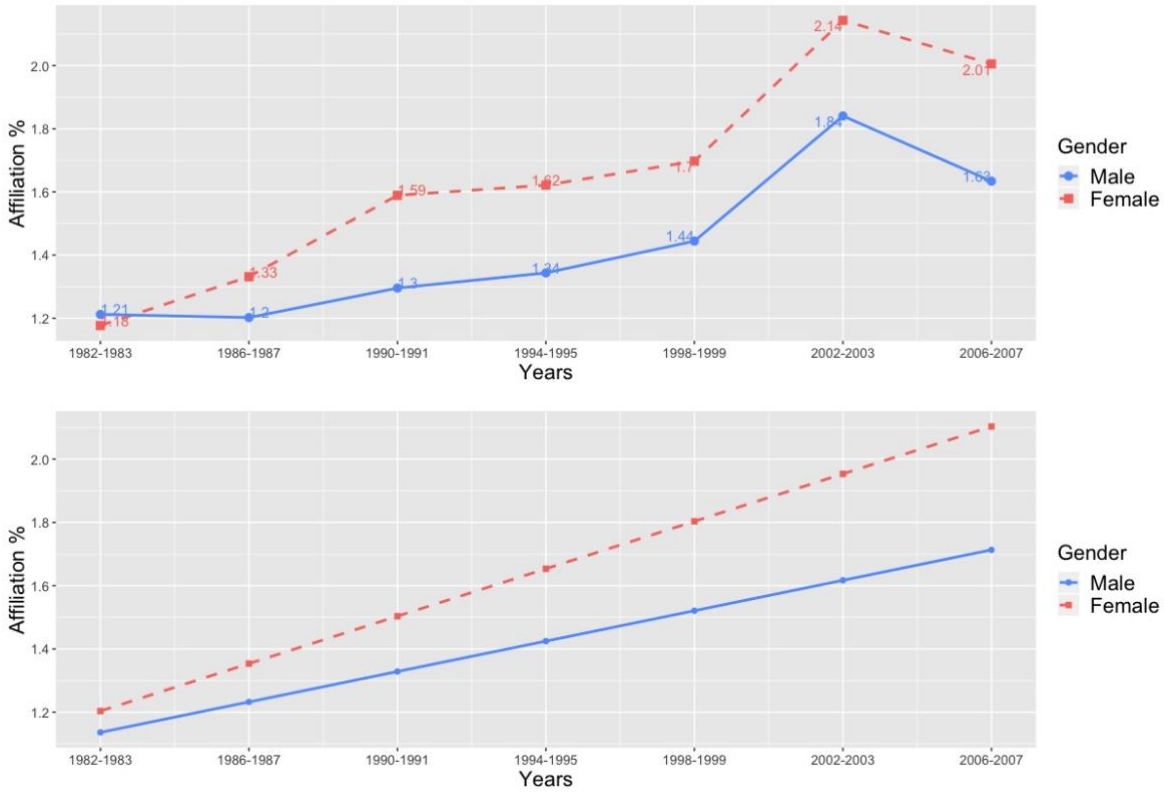
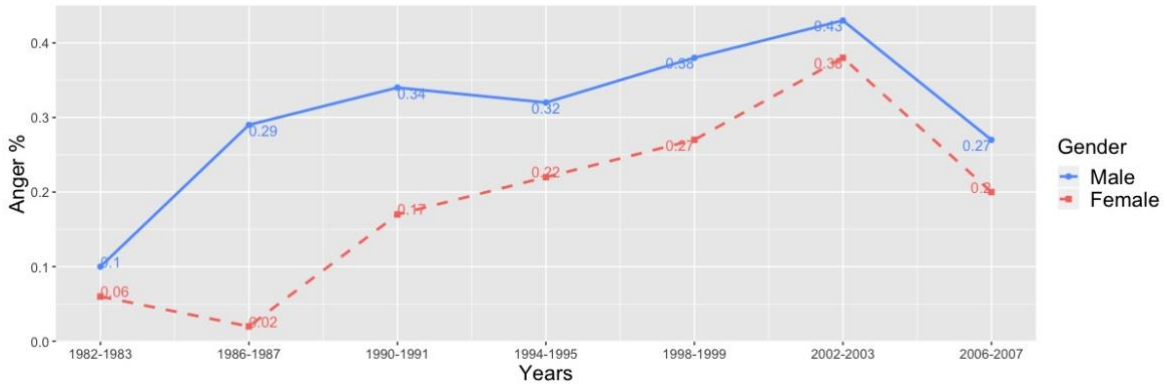


Figure D5. **Frequencies and trend lines of Affiliation by gender over time..**



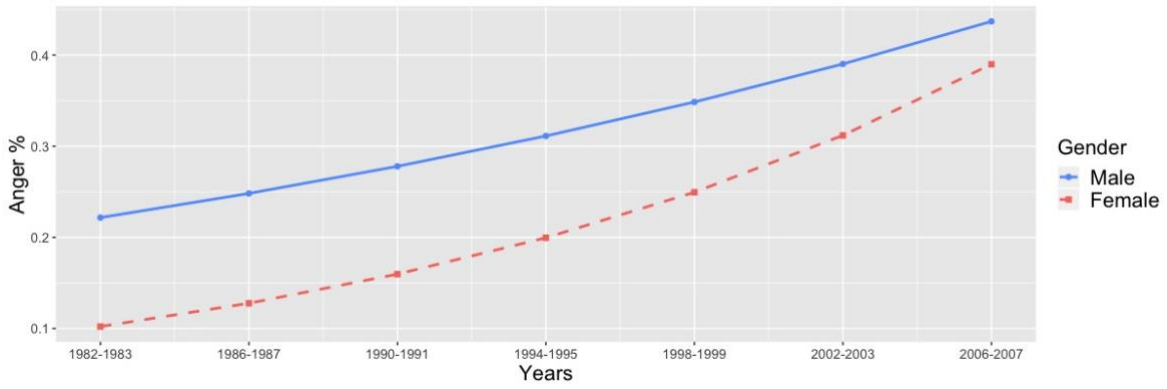


Figure D6. **Frequencies and trend lines of Anger by gender over time.**

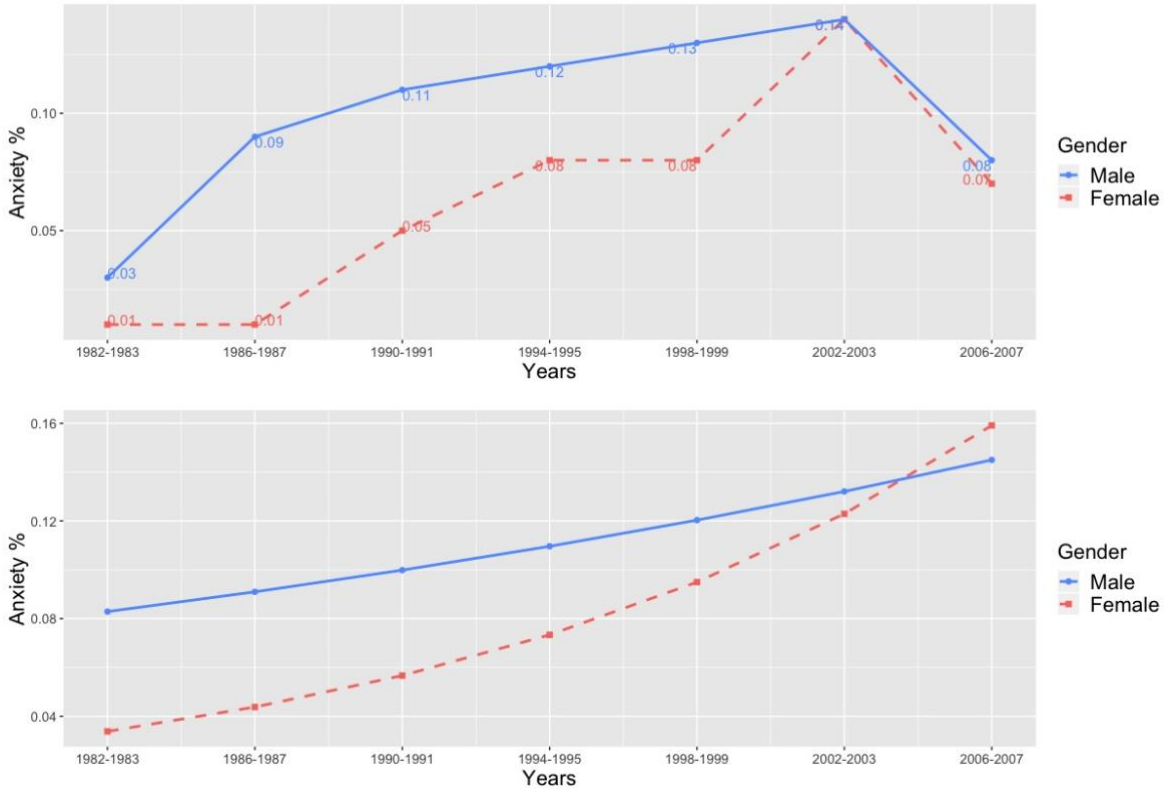
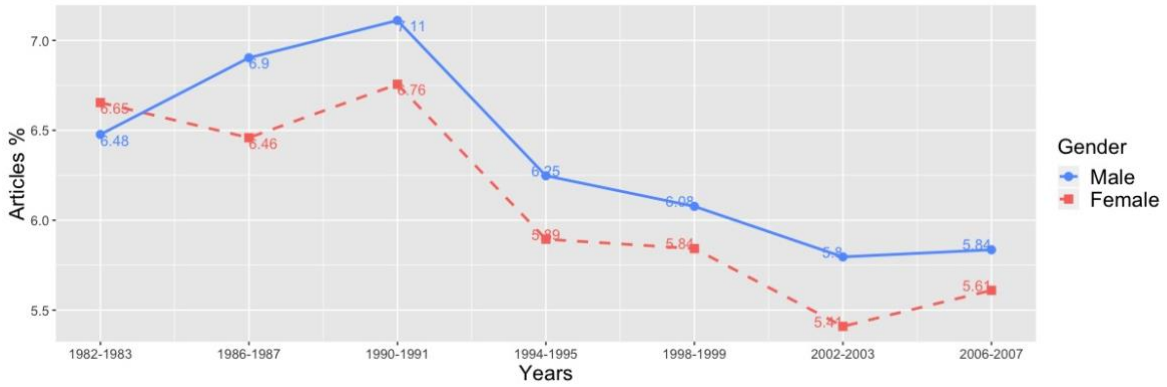


Figure D7. **Frequencies and trend lines of Anxiety by gender over time.**



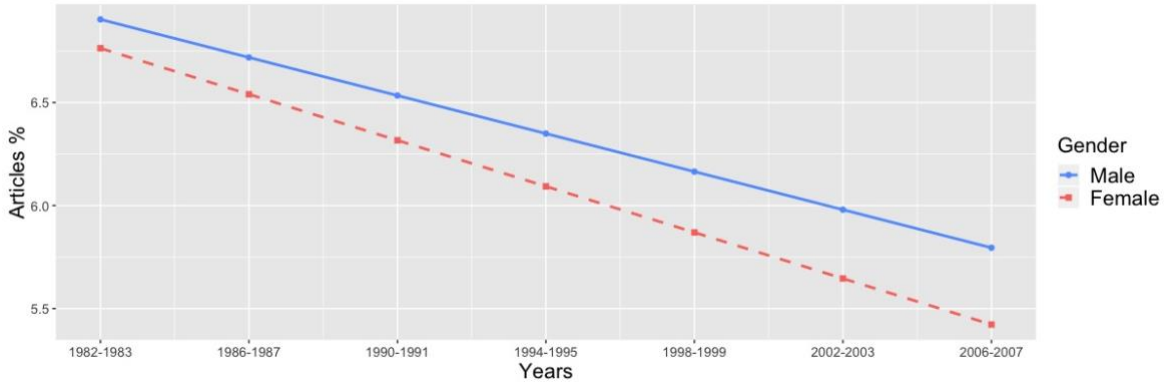


Figure D8. **Frequencies and trend lines of Articles by gender over time.**

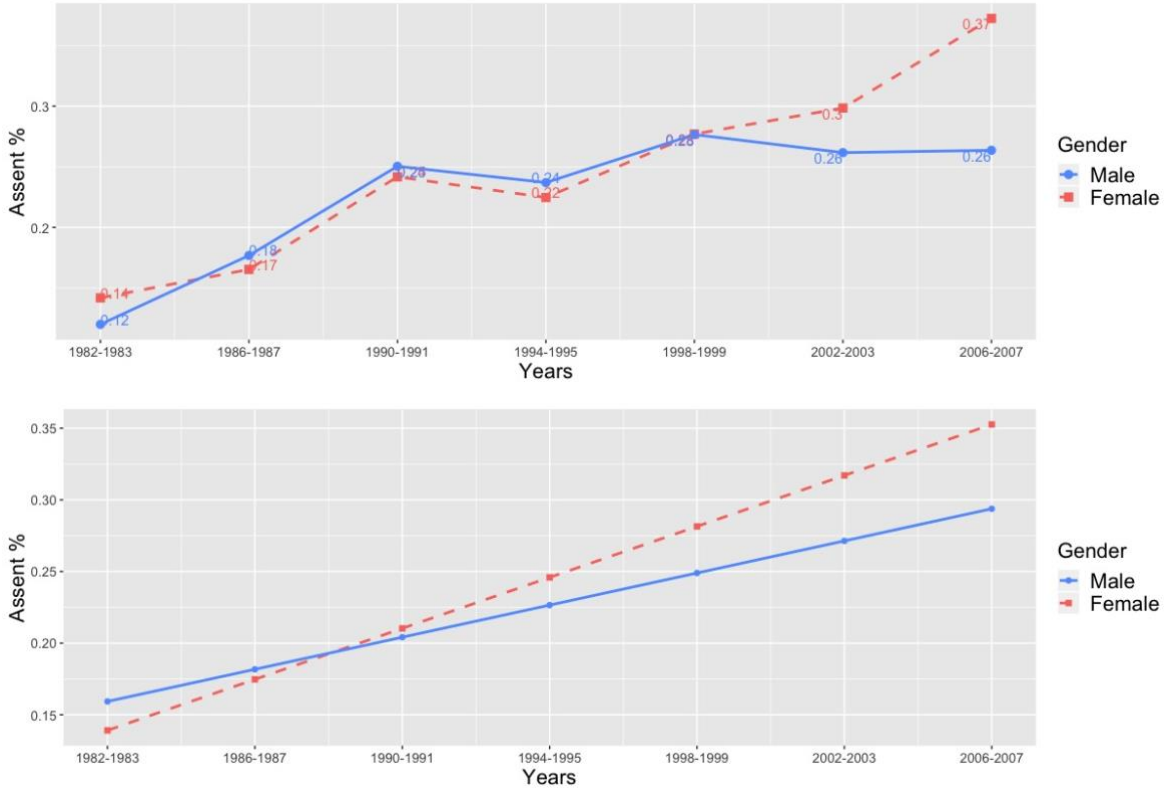
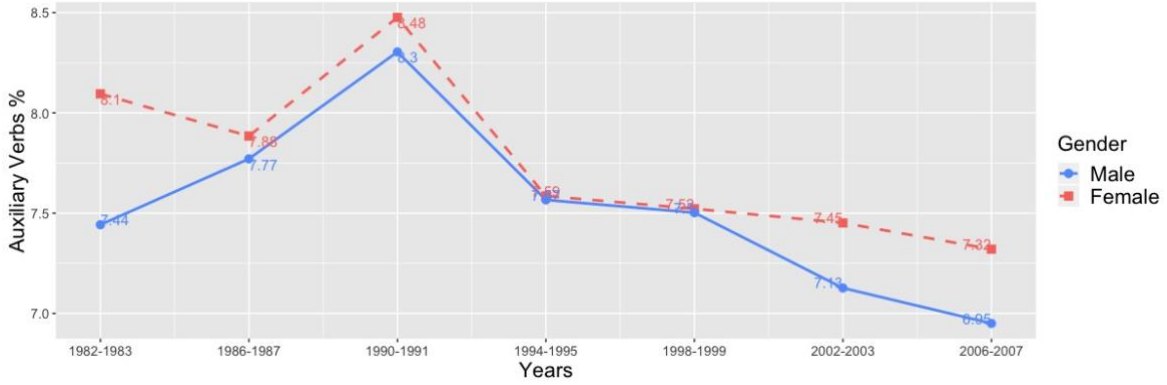


Figure D9. **Frequencies and trend lines of Assent by gender over time.**



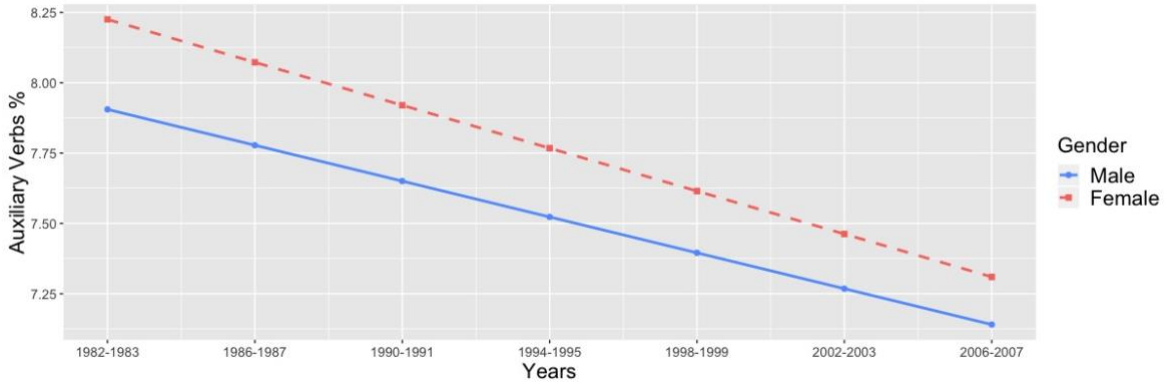


Figure D10. **Frequencies and trend lines of Auxiliary Verbs by gender over time.**

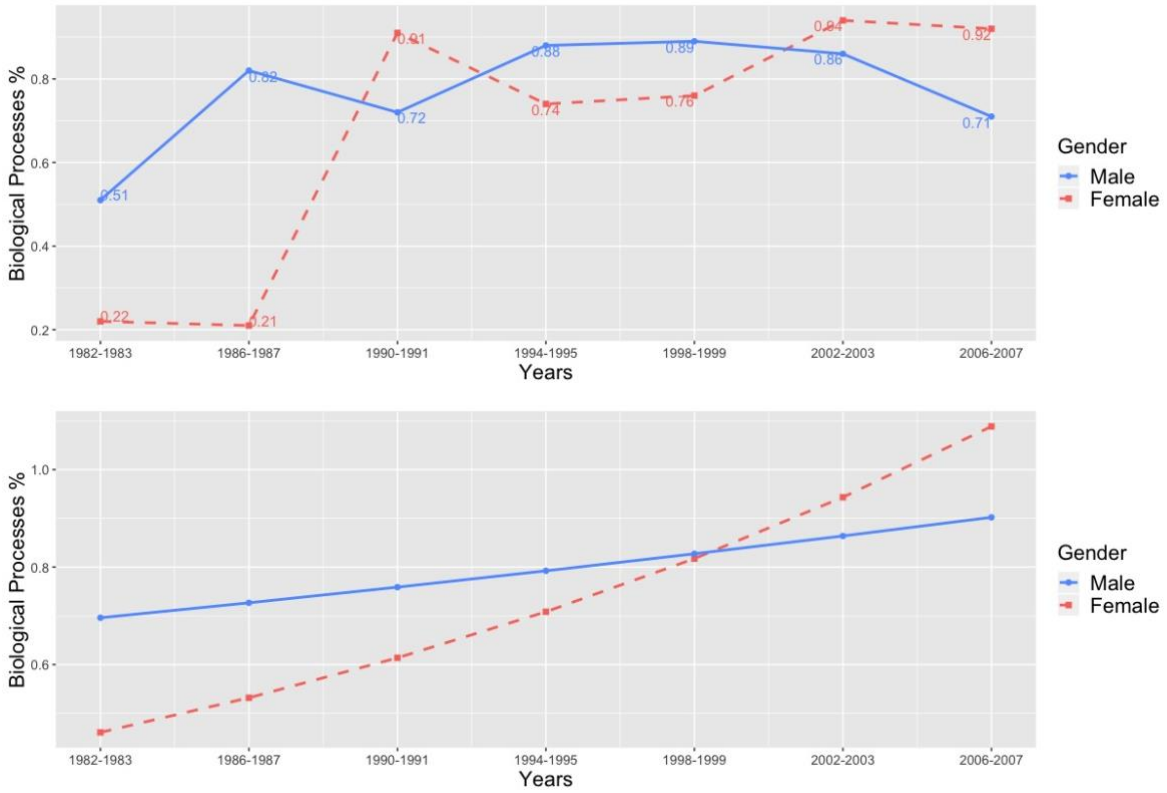
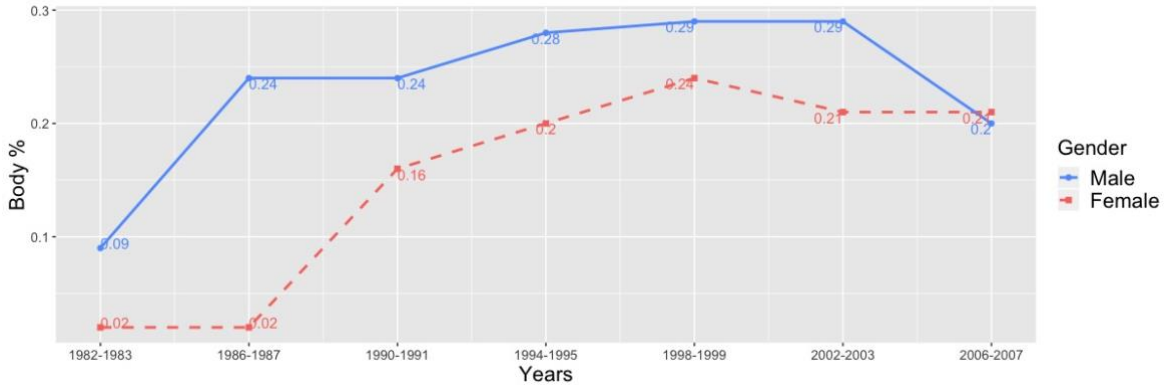


Figure D11. **Frequencies and trend lines of Biological Processes by gender over time.**



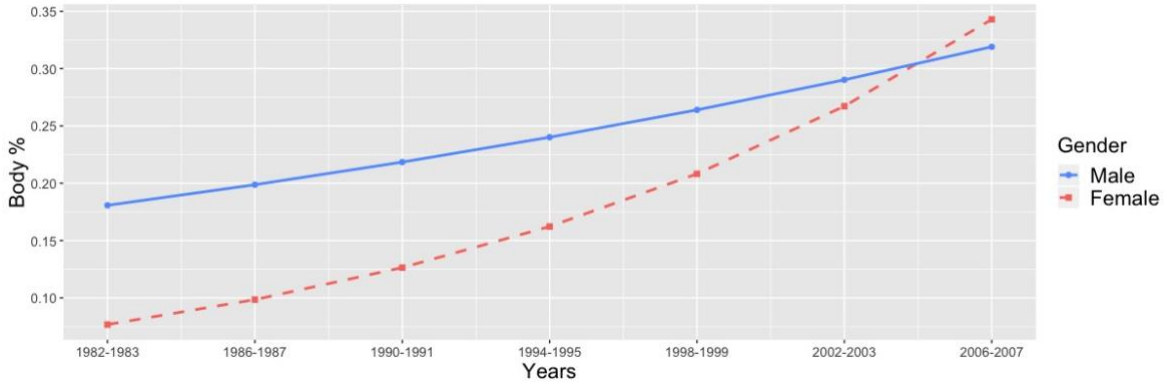


Figure D12. **Frequencies and trend lines of Body by gender over time.**

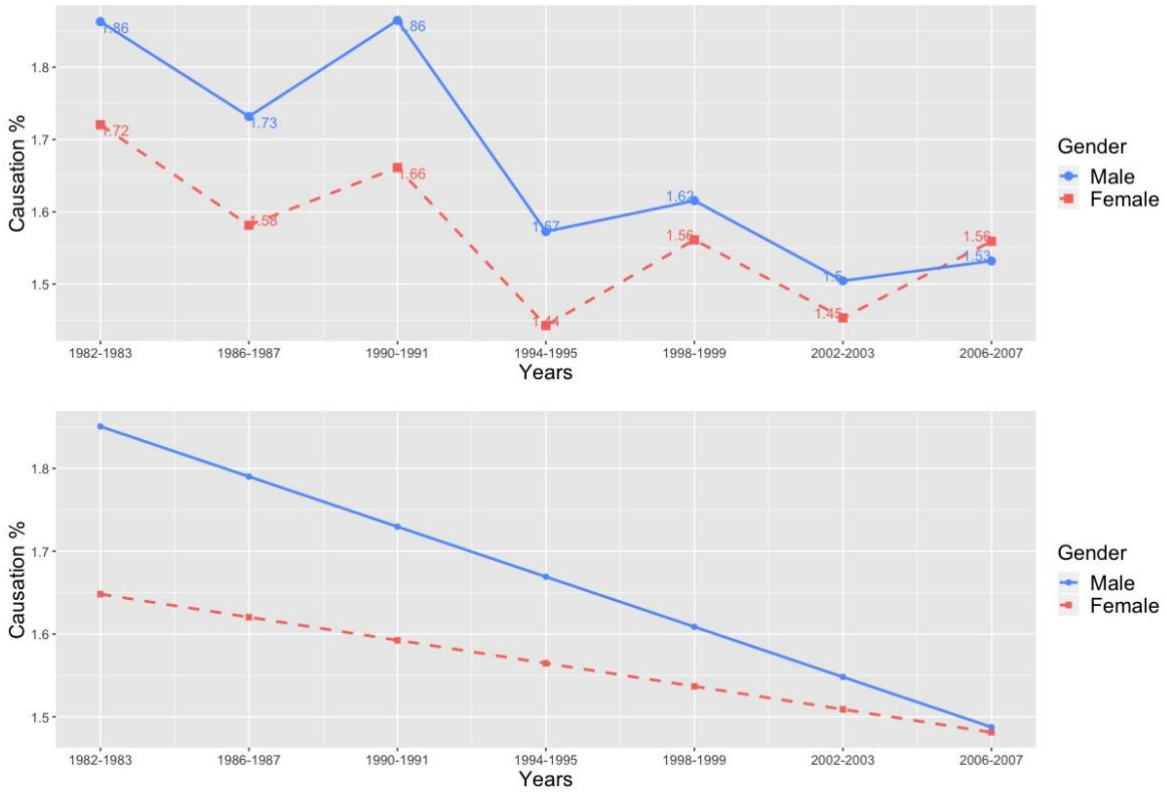
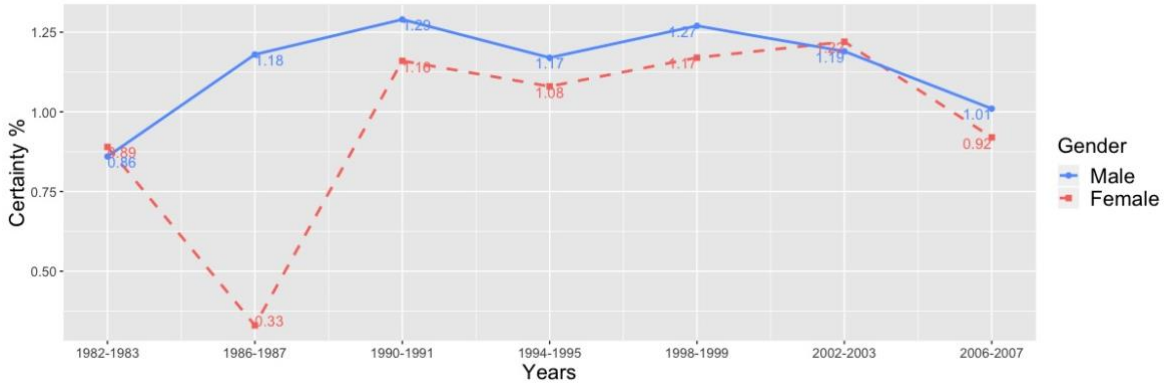


Figure D13. **Frequencies and trend lines of Causation by gender over time.**



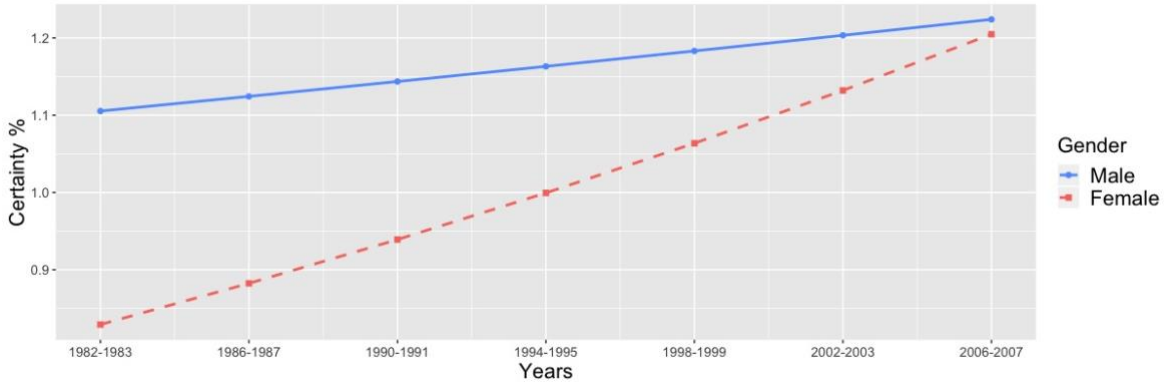


Figure D14. **Frequencies and trend lines of Certainty by gender over time.**

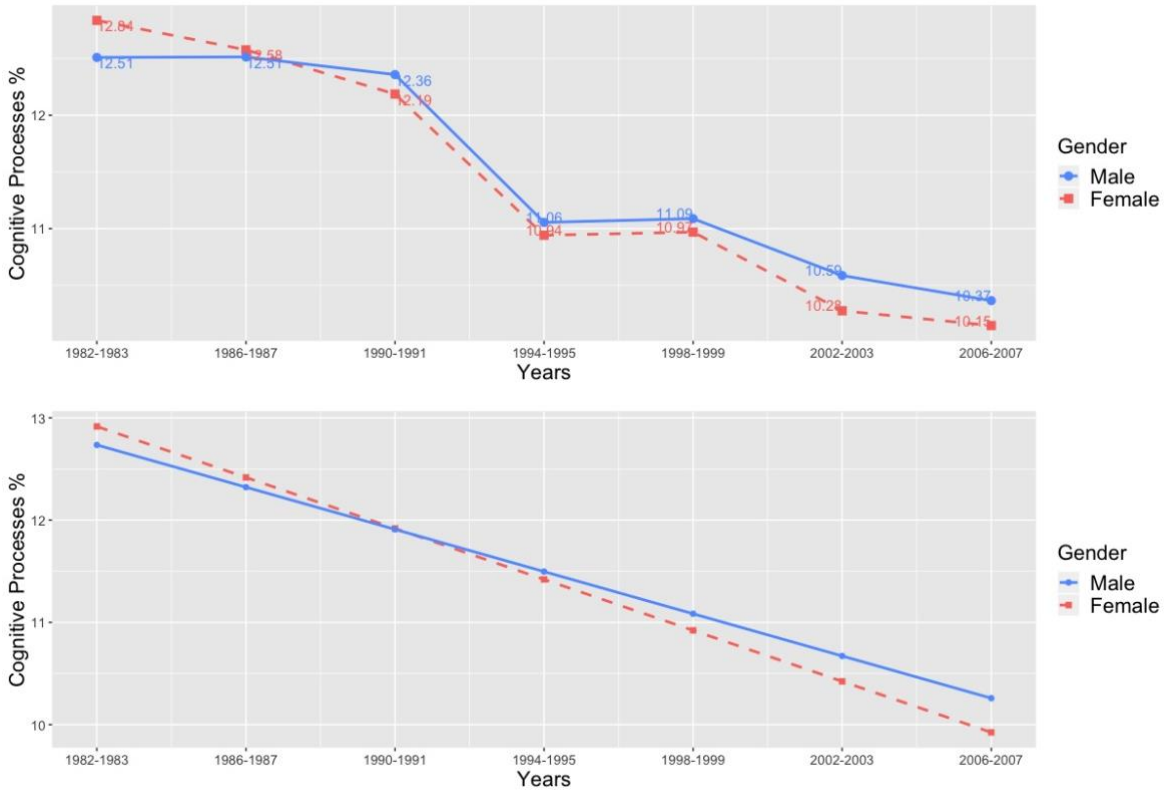
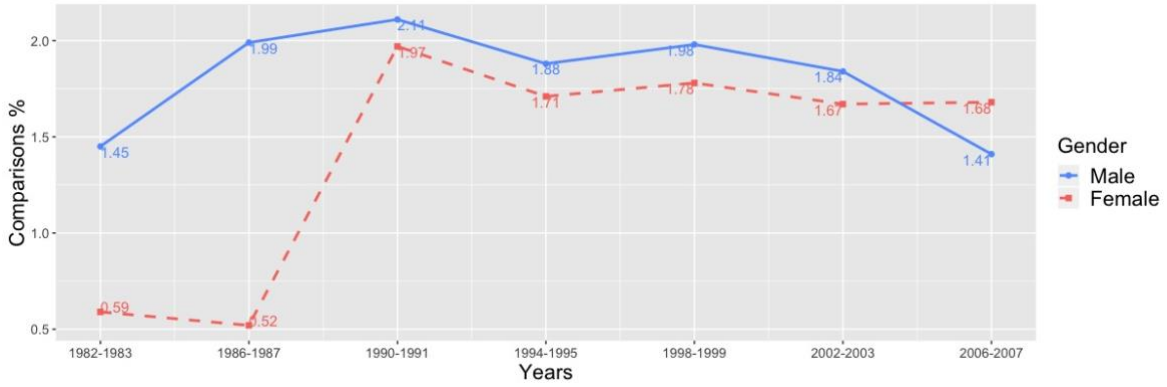


Figure D15. **Frequencies and trend lines of Cognitive Processes by gender over time.**



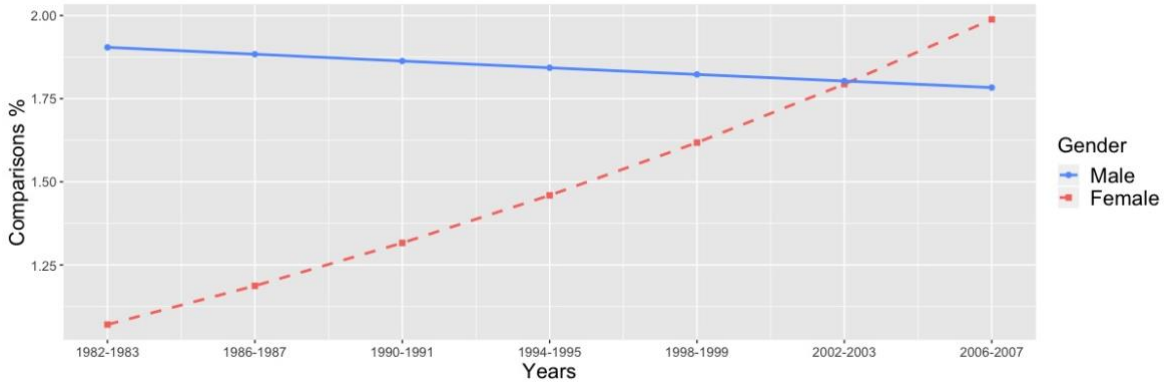


Figure D16. **Frequencies and trend lines of Comparisons by gender over time.**

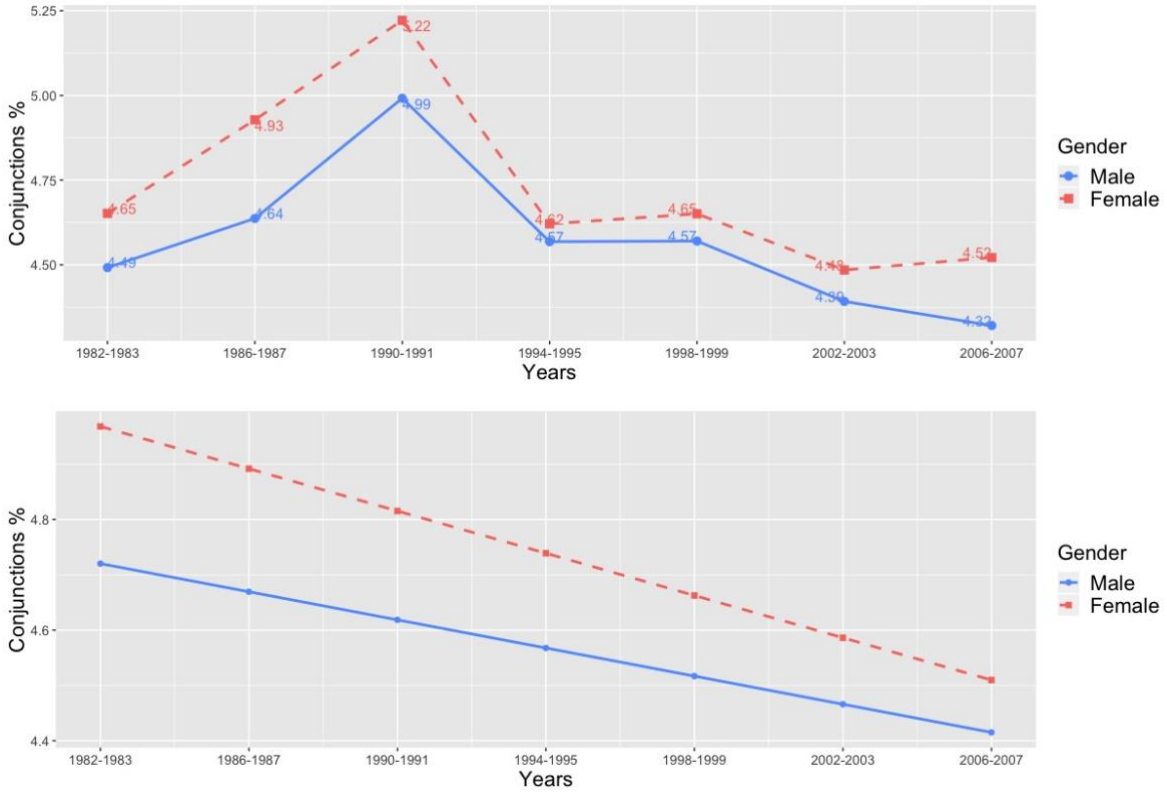
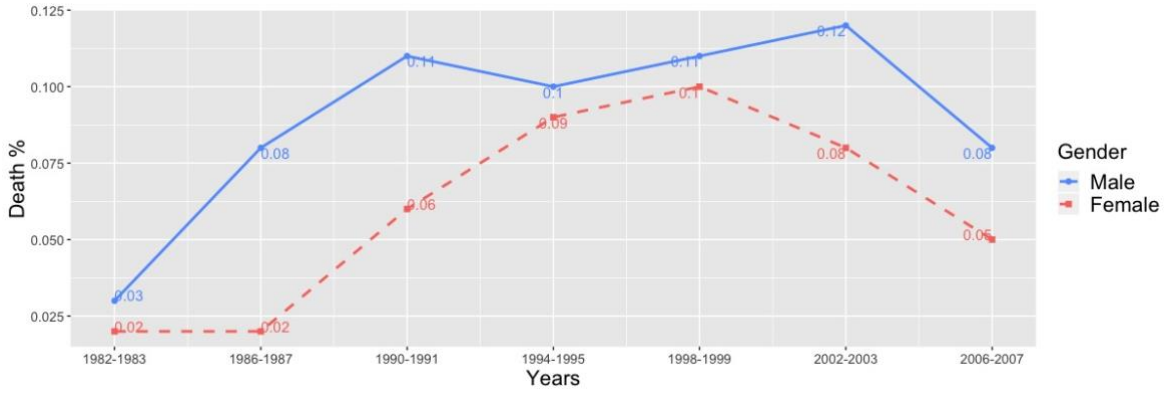


Figure D17. **Frequencies and trend lines of Conjunctions by gender over time.**



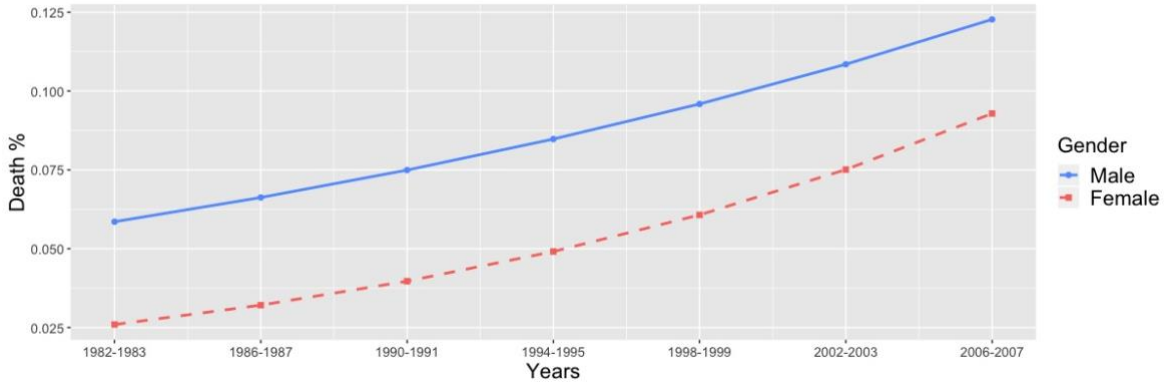


Figure D18. **Frequencies and trend lines of Death by gender over time.**

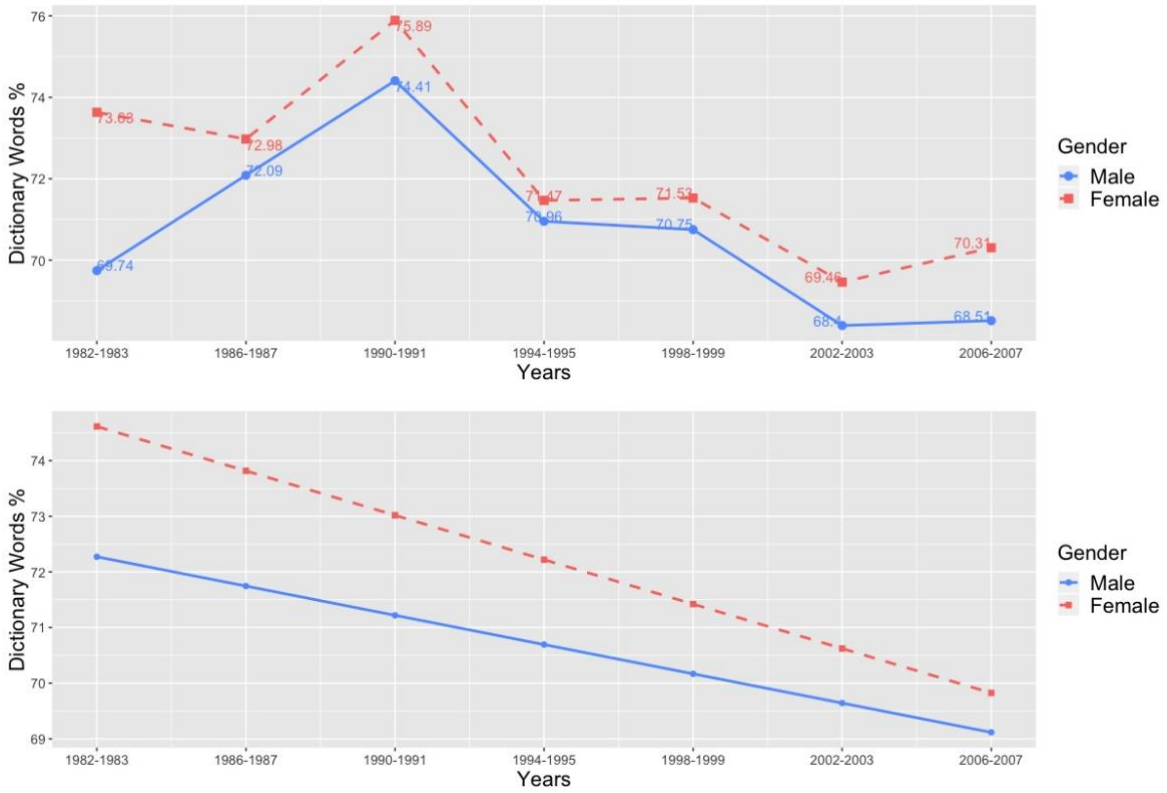
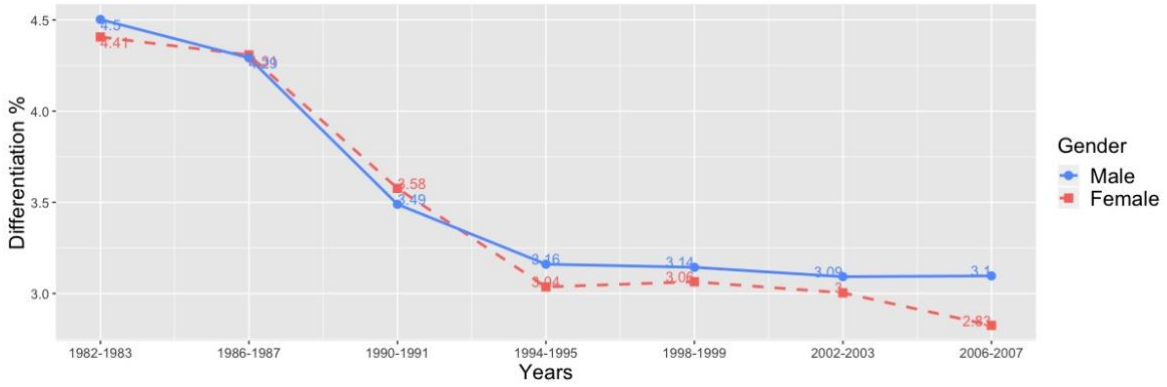


Figure D19. **Frequencies and trend lines of Dictionary Words by gender over time.**



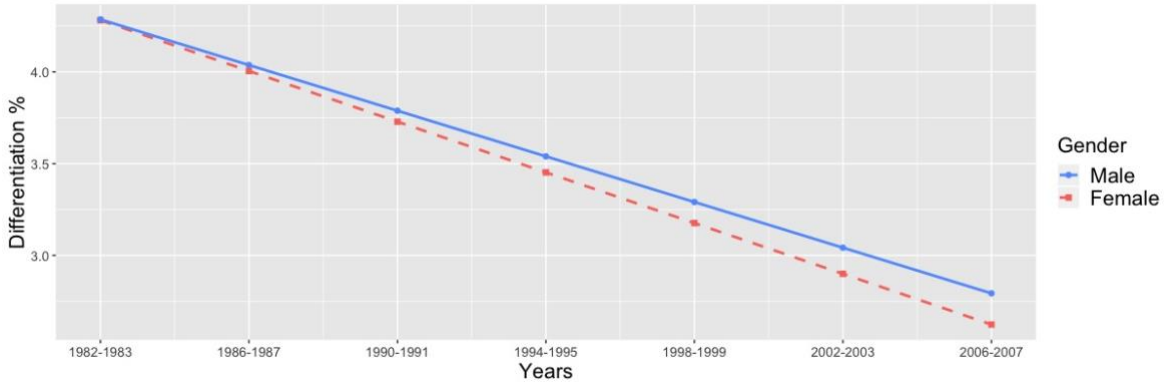


Figure D20. **Frequencies and trend lines of Differentiation by gender over time.**

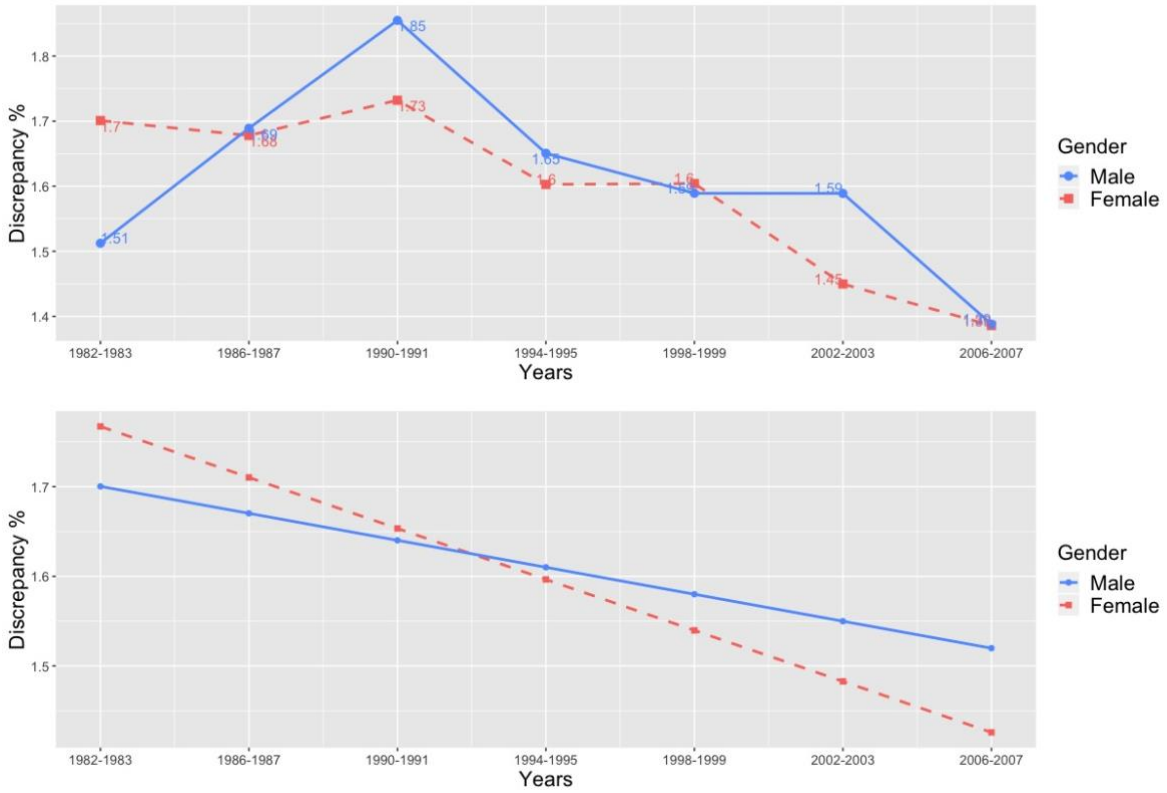
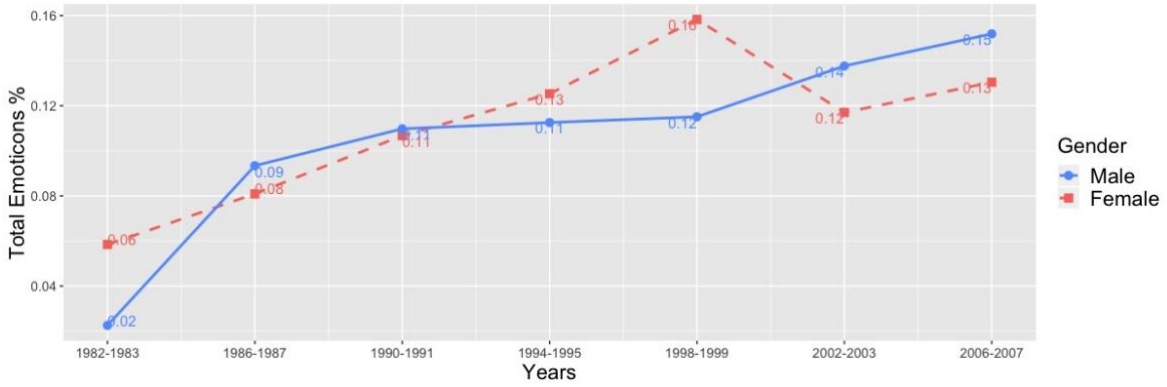


Figure D21. **Frequencies and trend lines of Discrepancy by gender over time.**



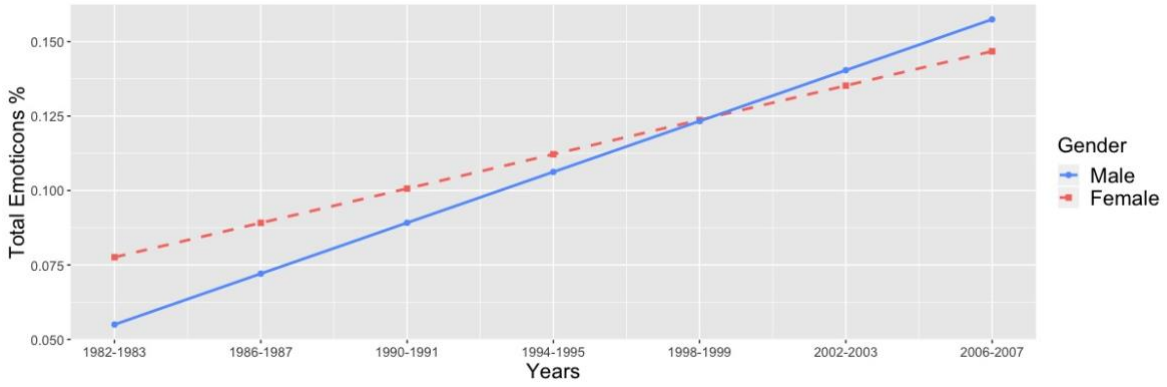


Figure D22. **Frequencies and trend lines of Total Emoticons by gender over time.**

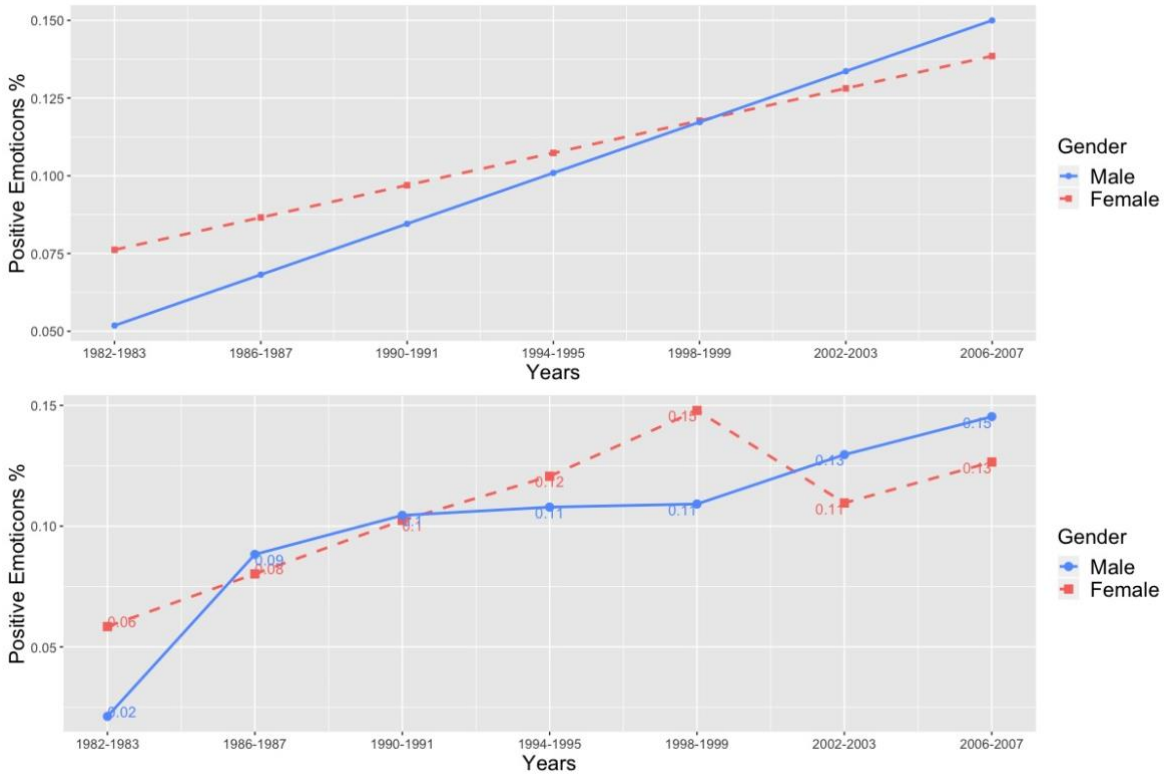
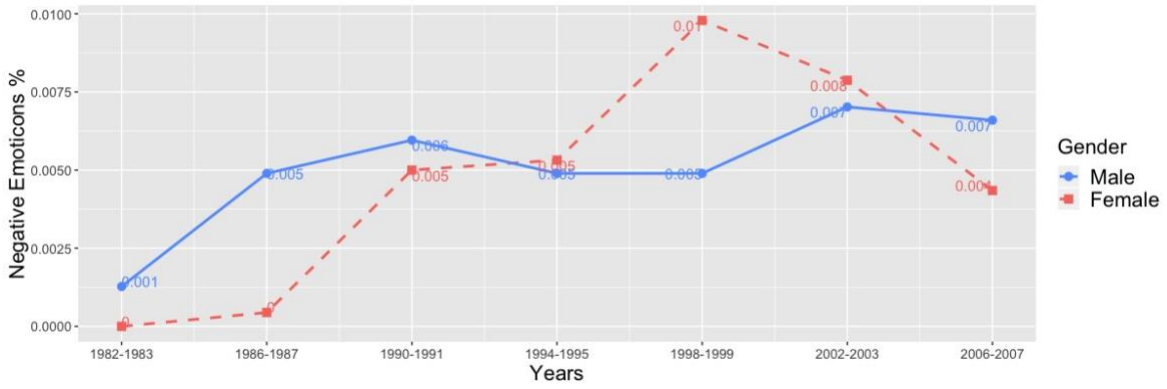


Figure D23. **Frequencies and trend lines of Positive Emoticons by gender over time.**



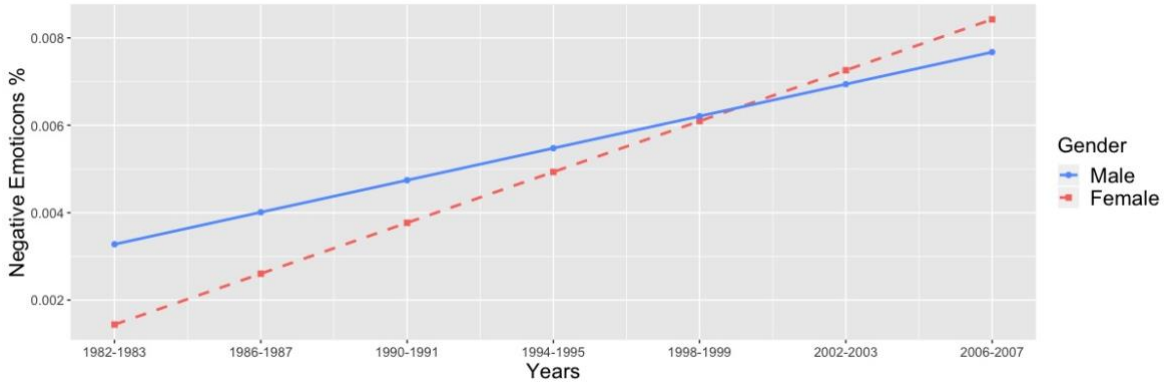


Figure D24. **Frequencies and trend lines of Negative Emoticons by gender over time.**

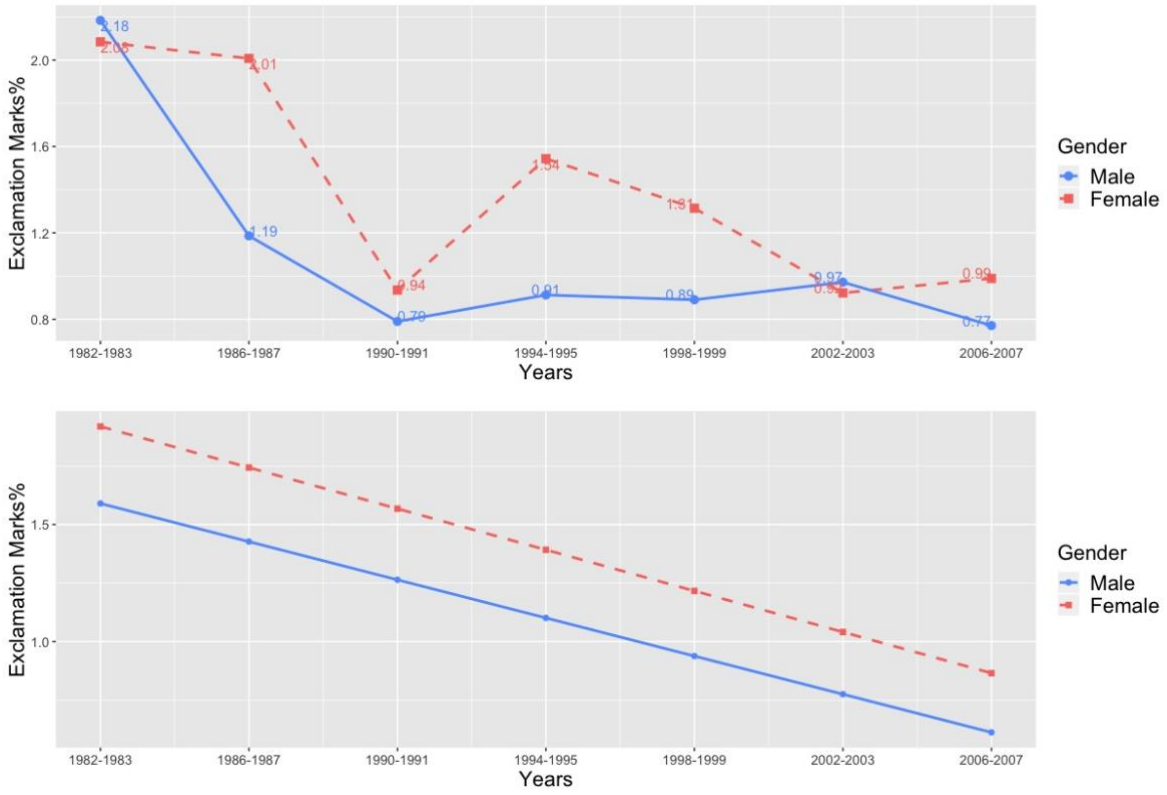
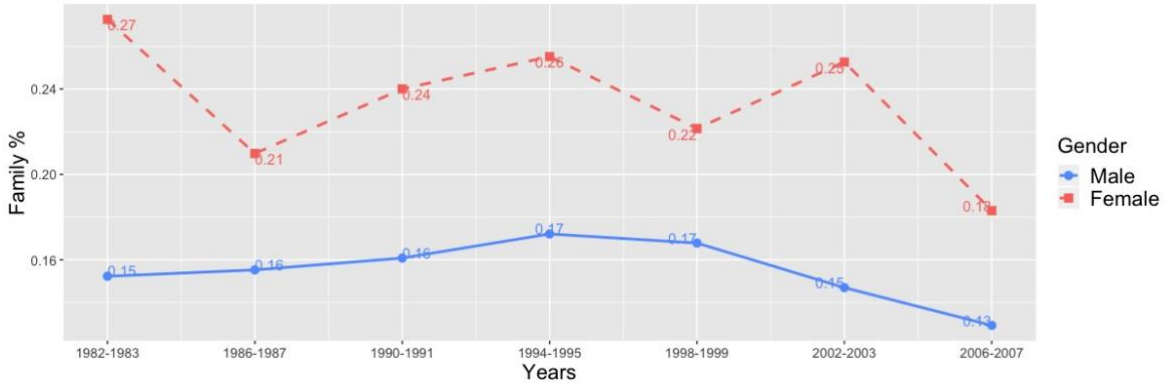


Figure D25. **Frequencies and trend lines of Exclamation Marks by gender over time.**



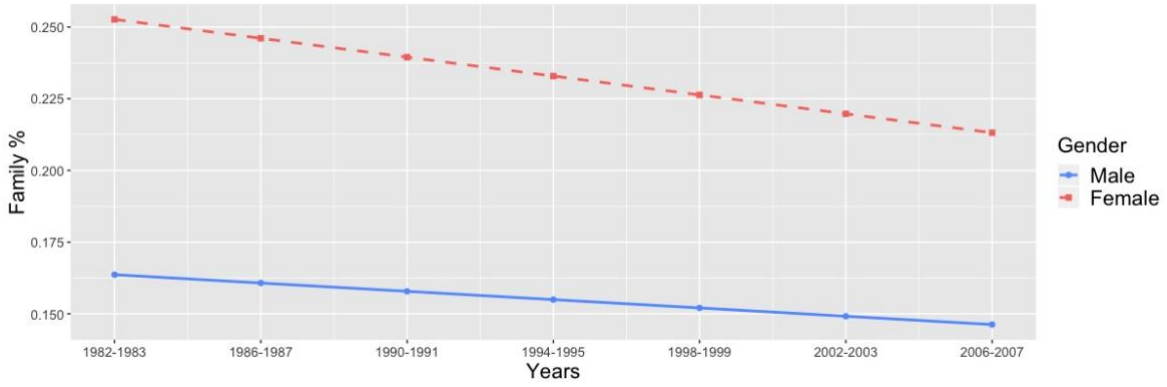


Figure D26. **Frequencies and trend lines of Family by gender over time.**

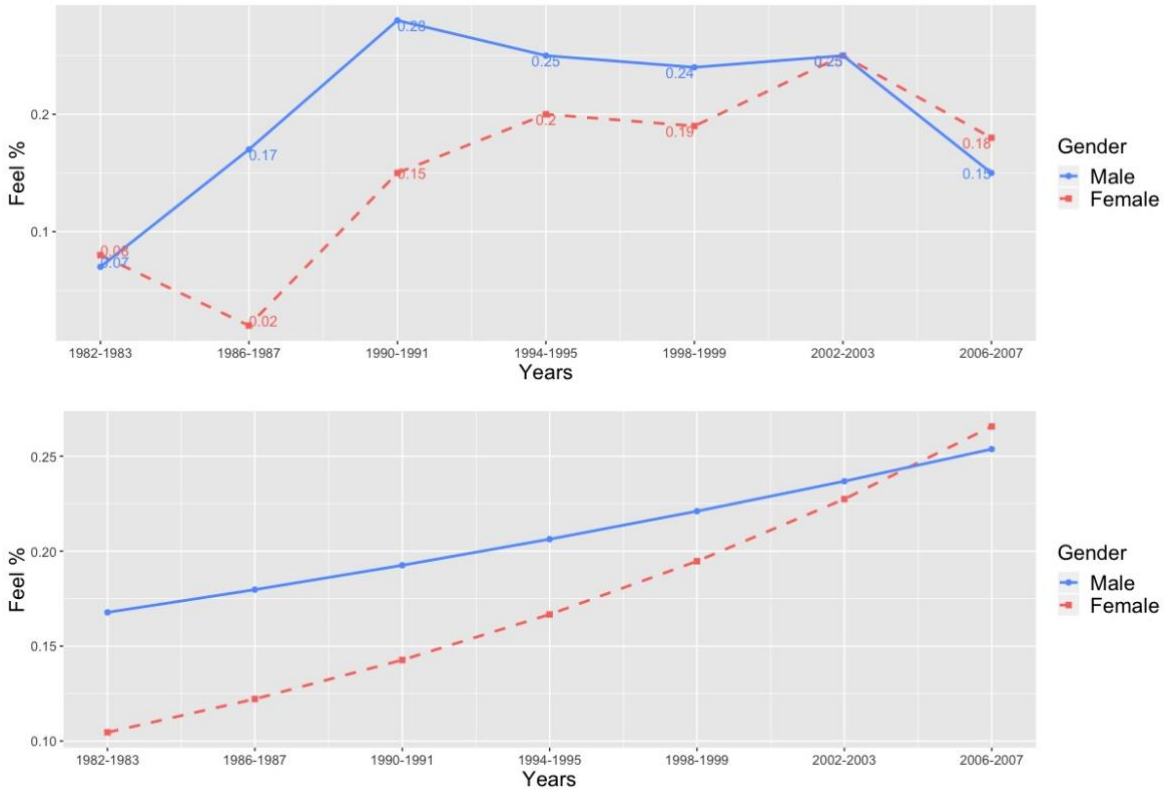
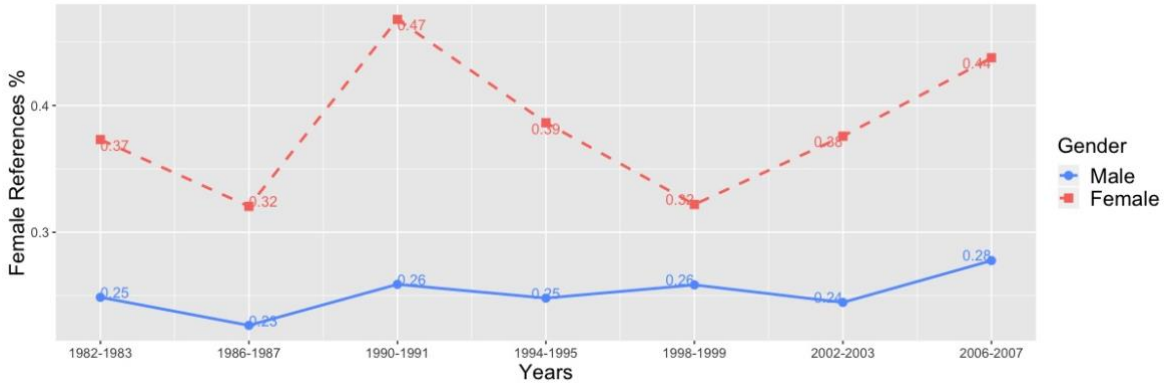


Figure D27. **Frequencies and trend lines of Feel by gender over time.**



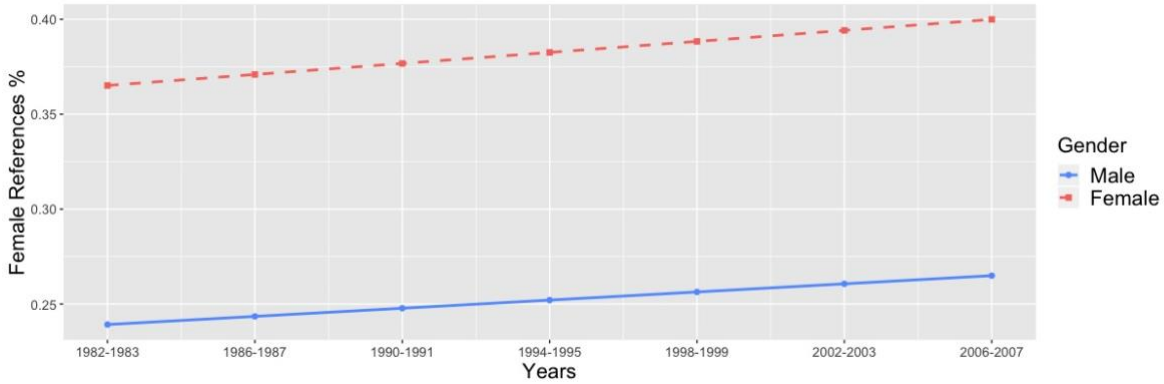


Figure D28. **Frequencies and trend lines of Female References by gender over time.**

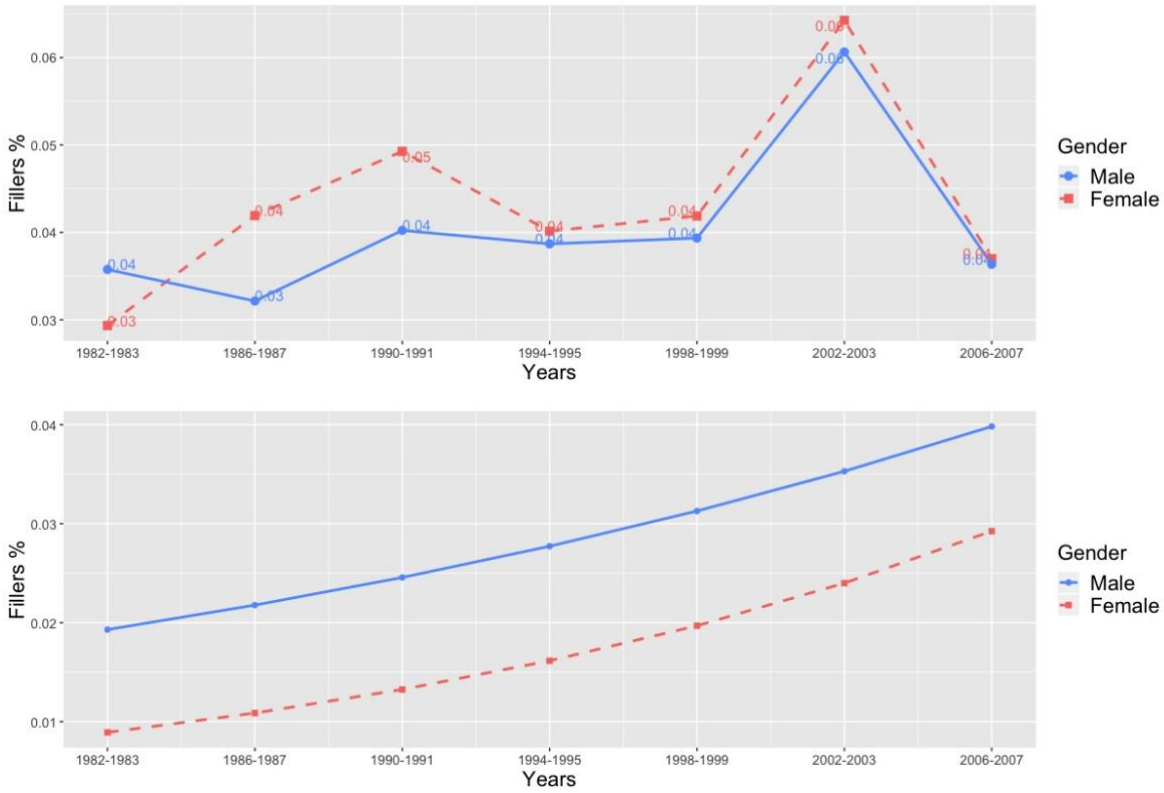
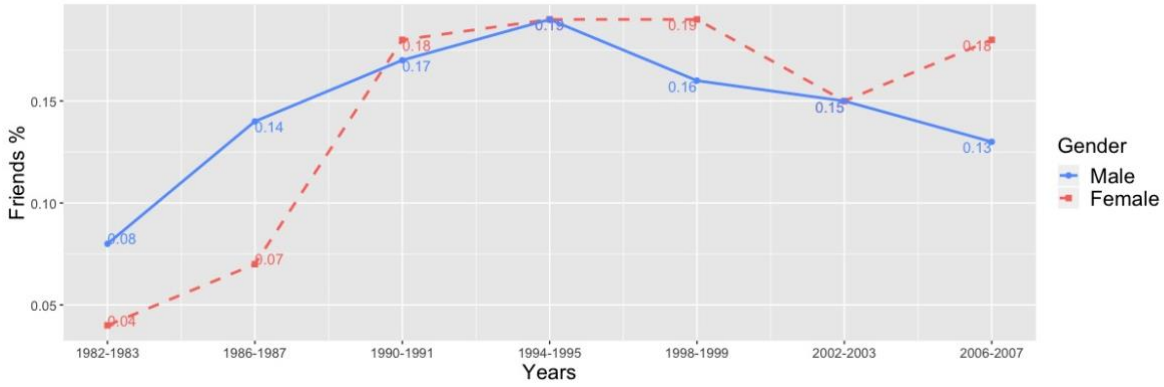


Figure D29. **Frequencies and trend lines of Fillers by gender over time.**



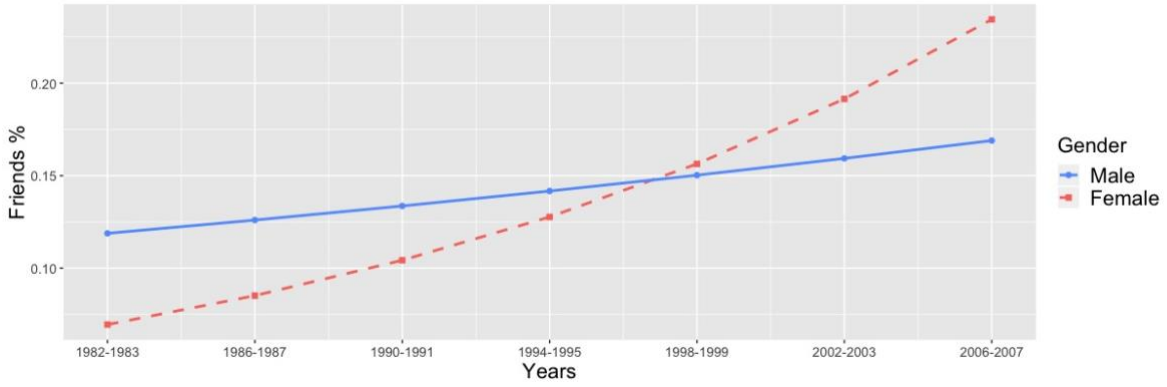


Figure D30. **Frequencies and trend lines of Friend by gender over time.**

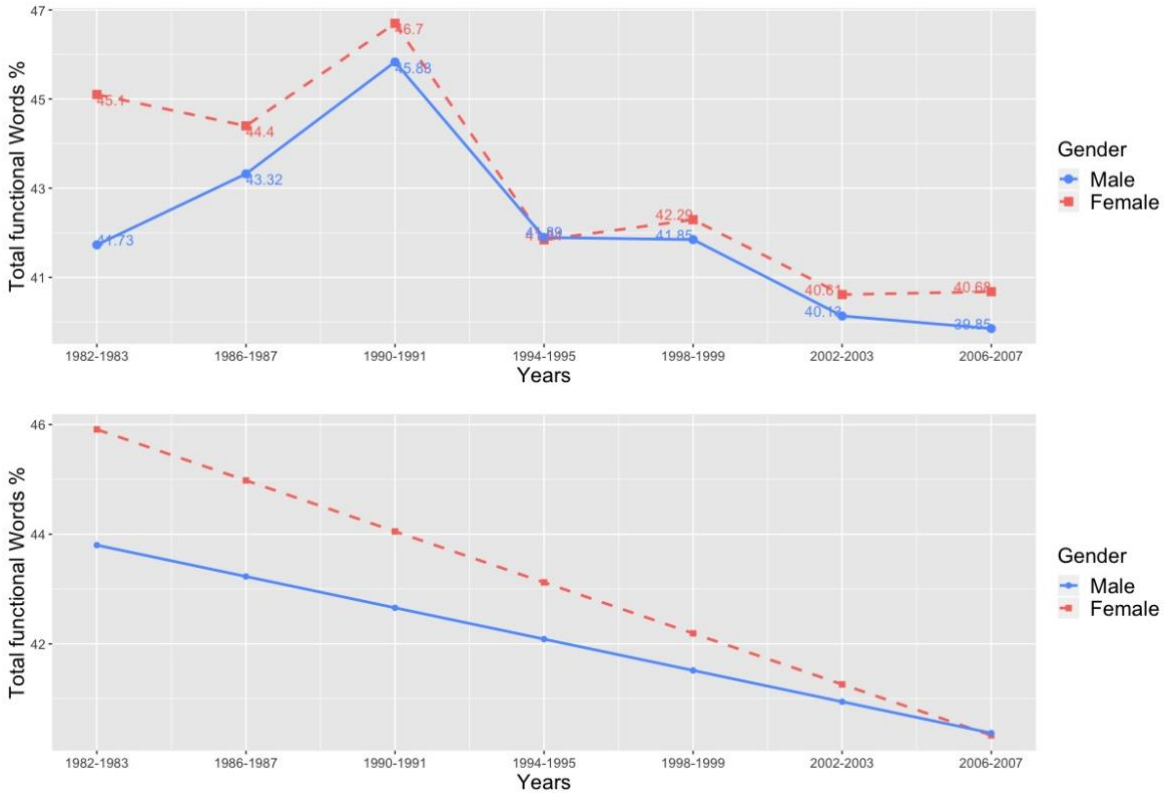
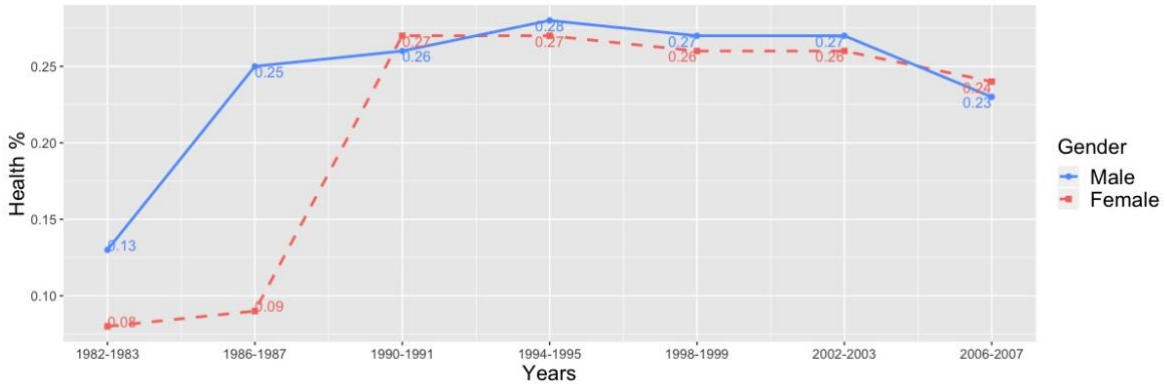


Figure D31. **Frequencies and trend lines of Total Function Words by gender over time.**



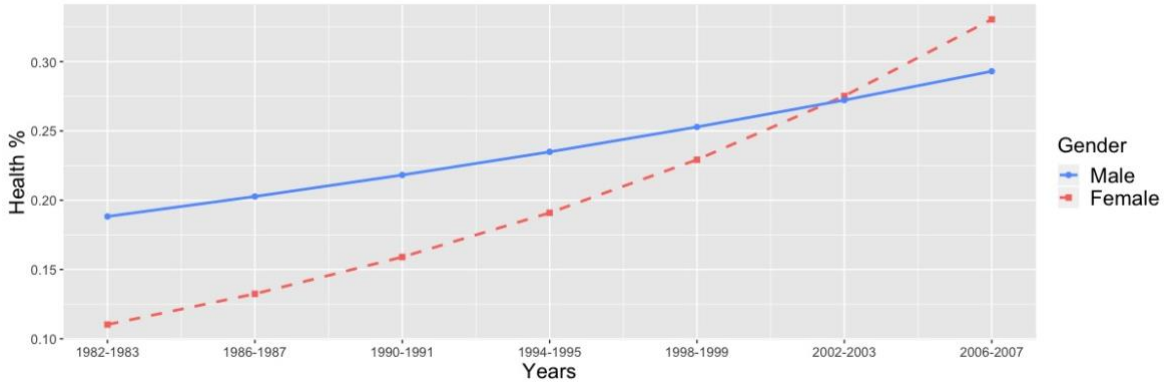


Figure D32. **Frequencies and trend lines of Health by gender over time.**

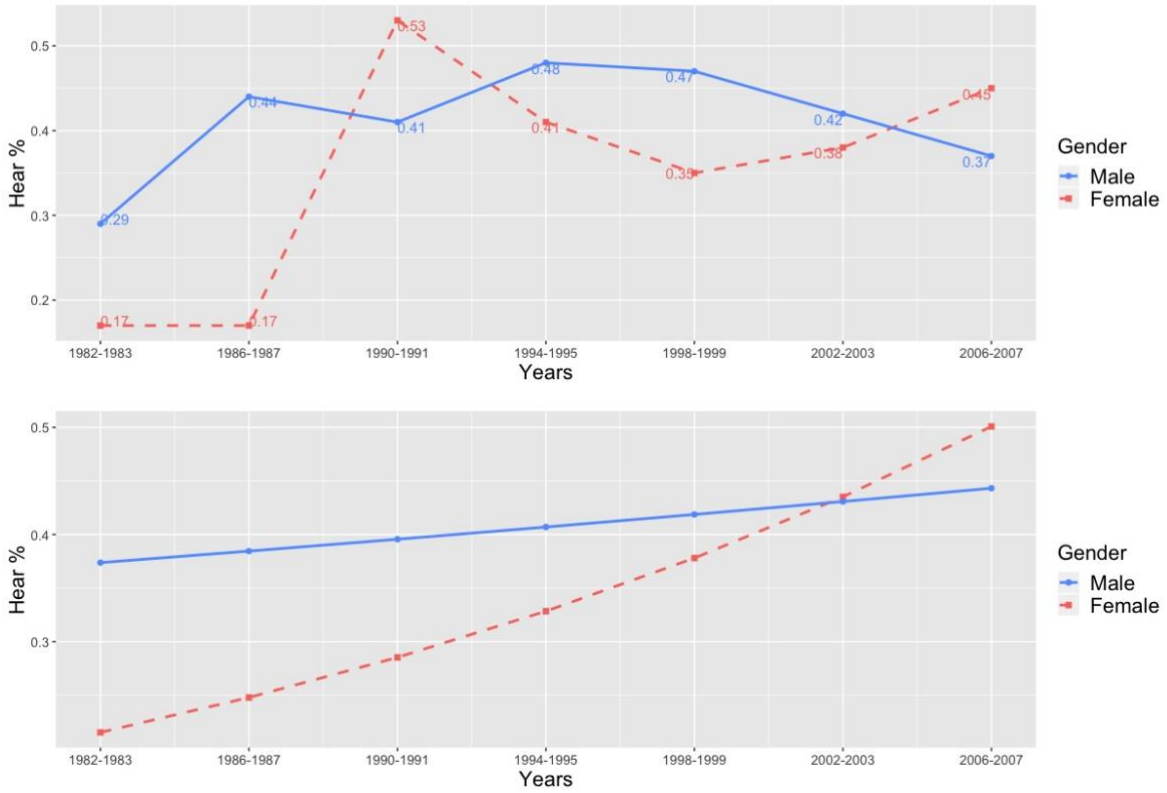
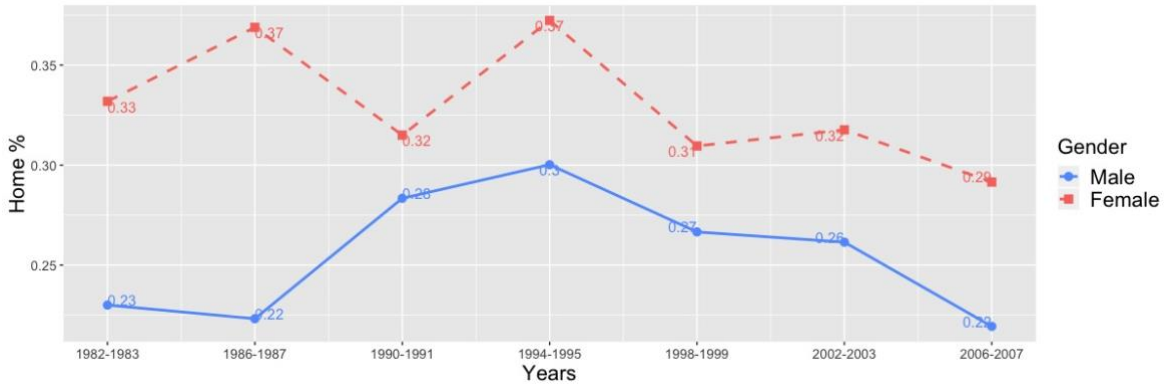


Figure D33. **Frequencies and trend lines of Hear by gender over time.**



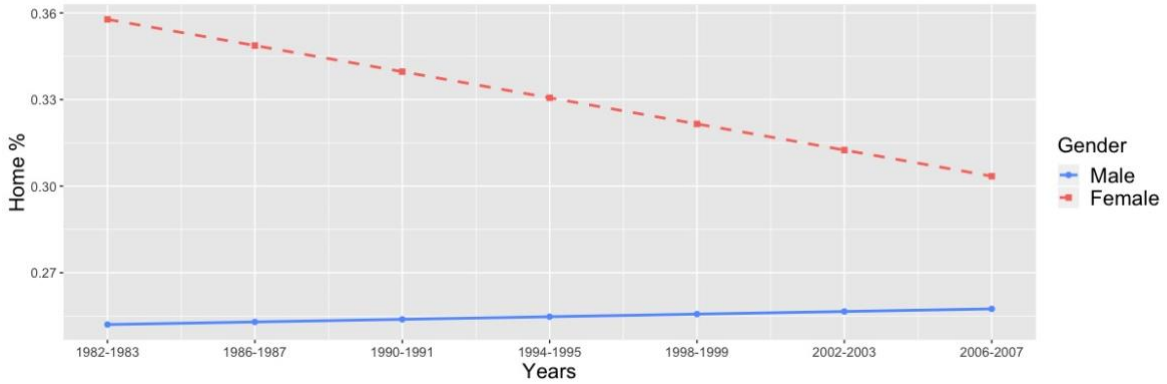


Figure D34. **Frequencies and trend lines of Home by gender over time.**

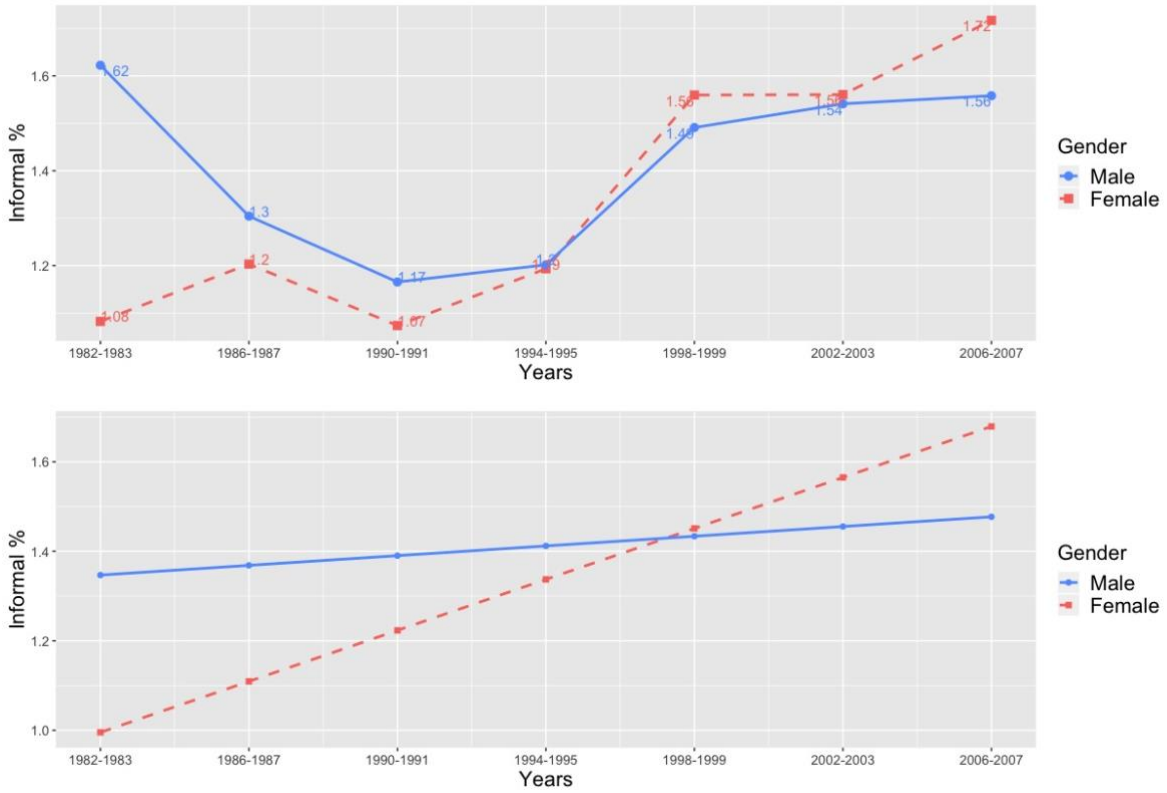
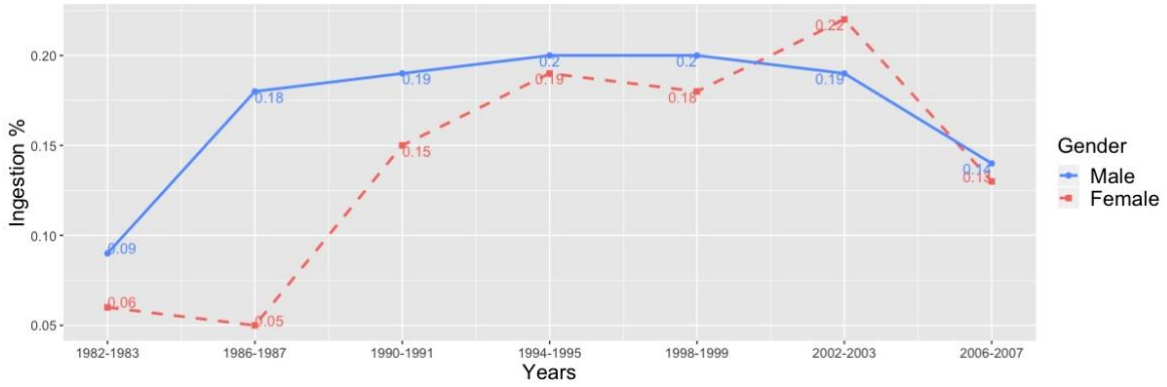


Figure D35. **Frequencies and trend lines of Informal Language by gender over time.**



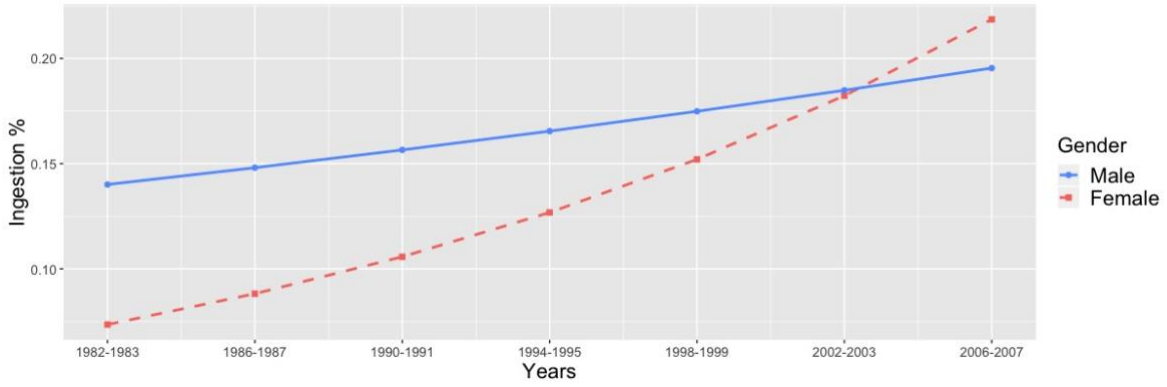


Figure D36. **Frequencies and trend lines of Ingestion by gender over time.**

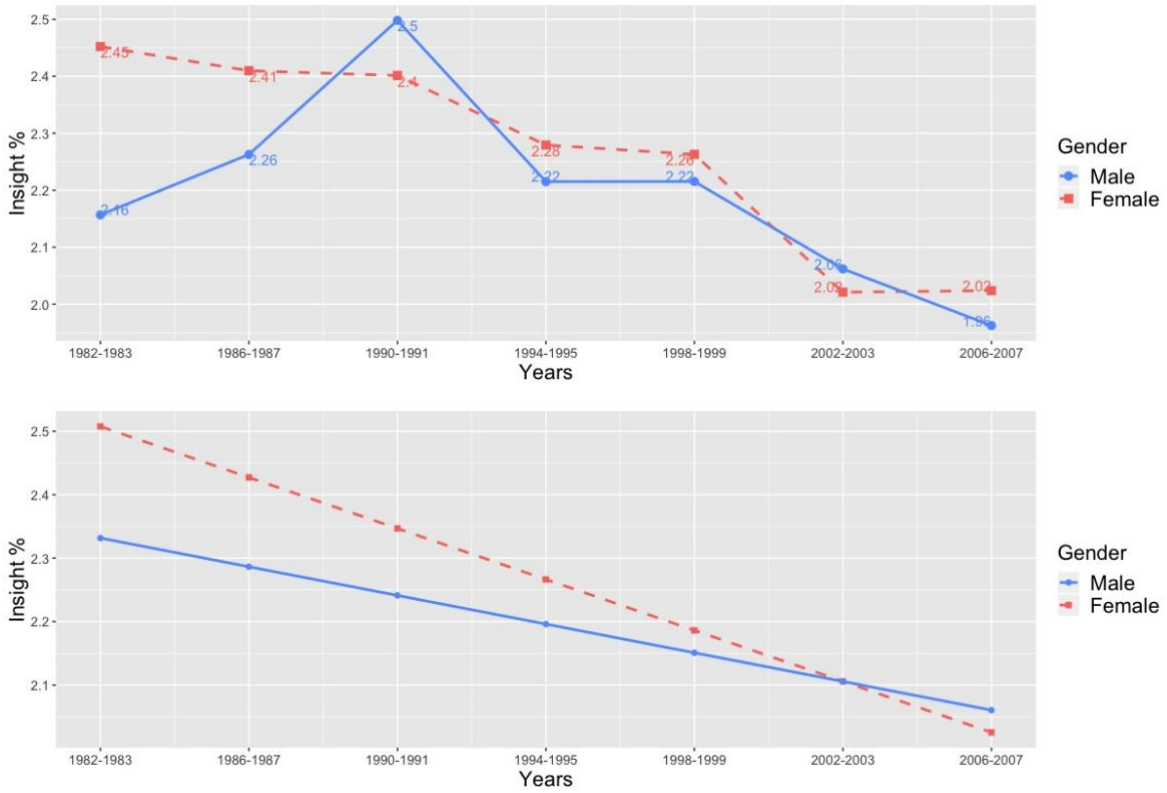
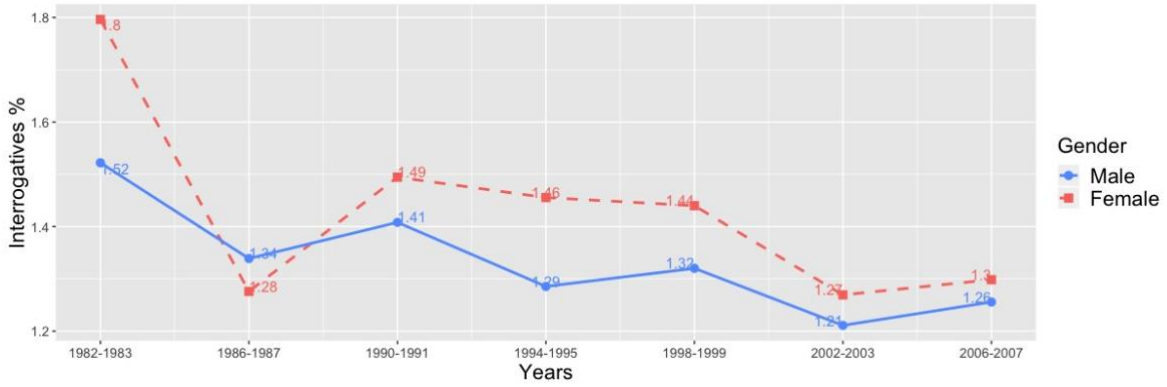


Figure D37. **Frequencies and trend lines of Insight by gender over time.**



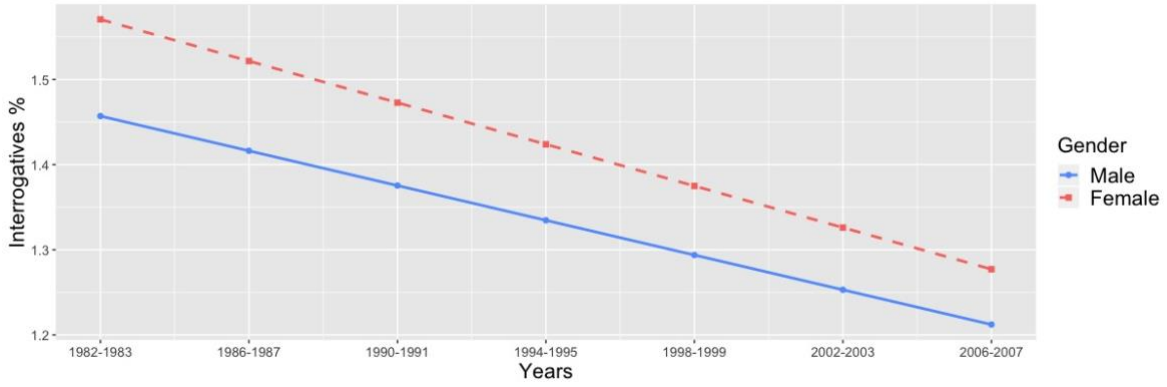


Figure D38. **Frequencies and trend lines of Interrogatives by gender over time.**

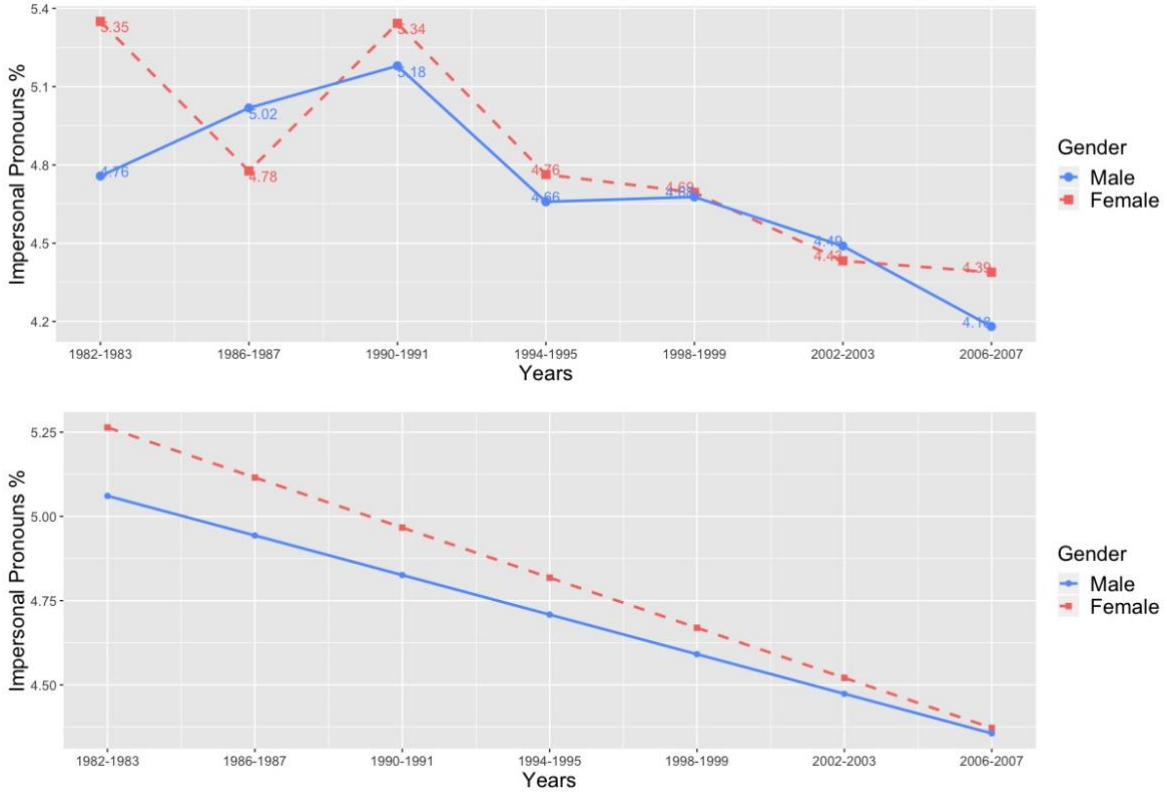


Figure D39. **Frequencies and trend lines of Money by gender over time.**

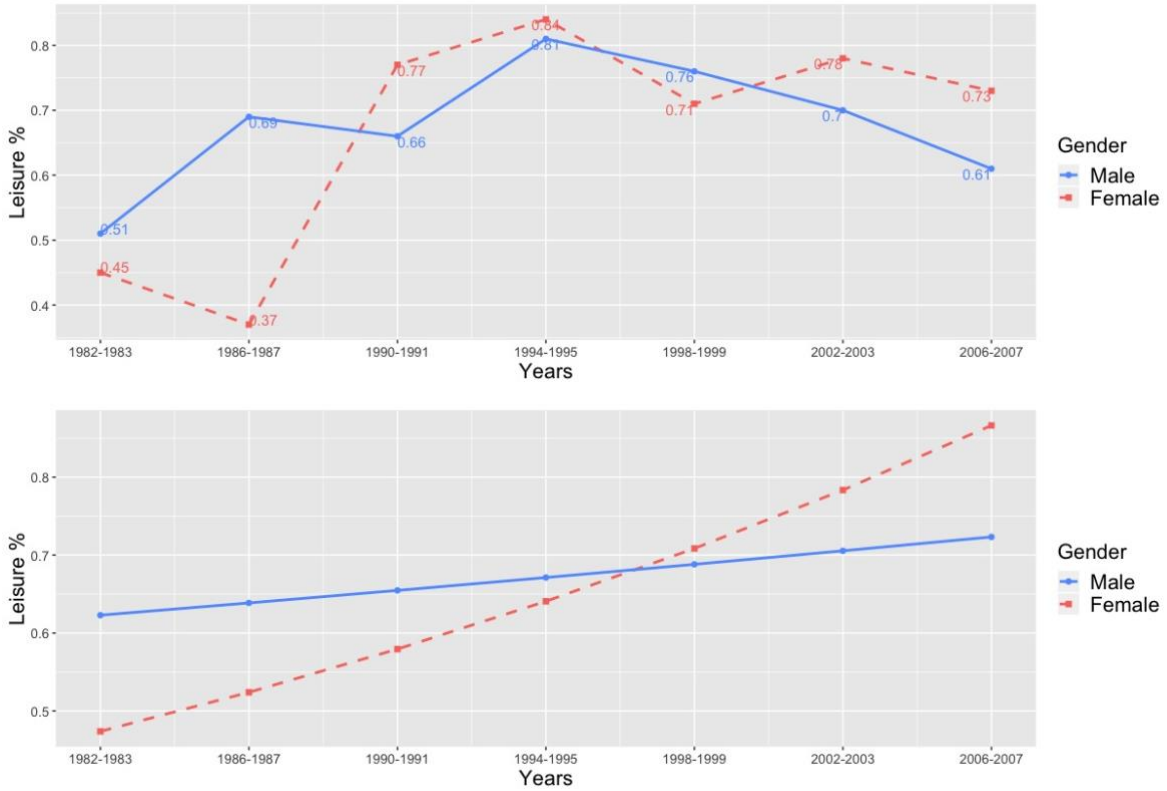
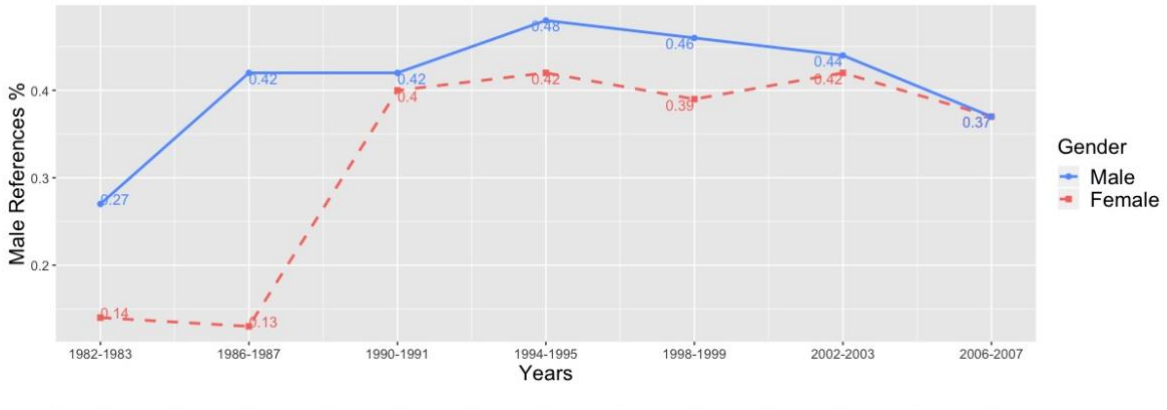


Figure D40. **Frequencies and trend lines of Leisure by gender over time.**



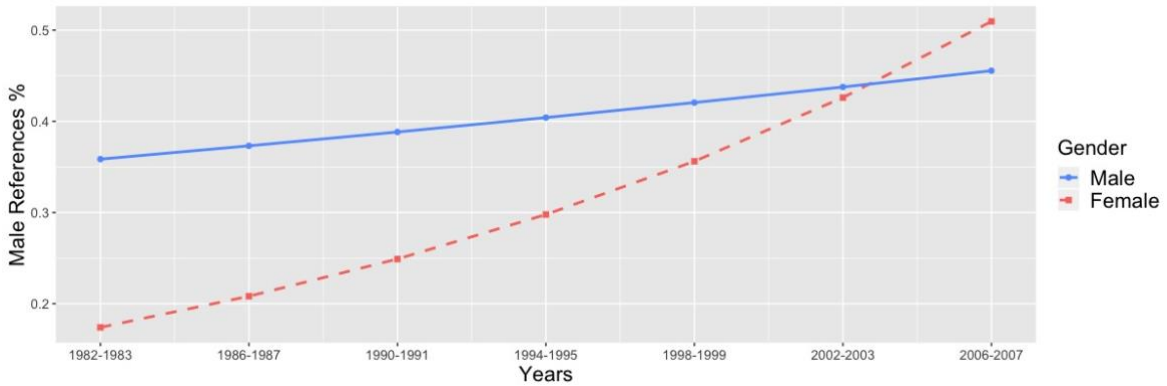


Figure D41. **Frequencies and trend lines of Male References by gender over time.**

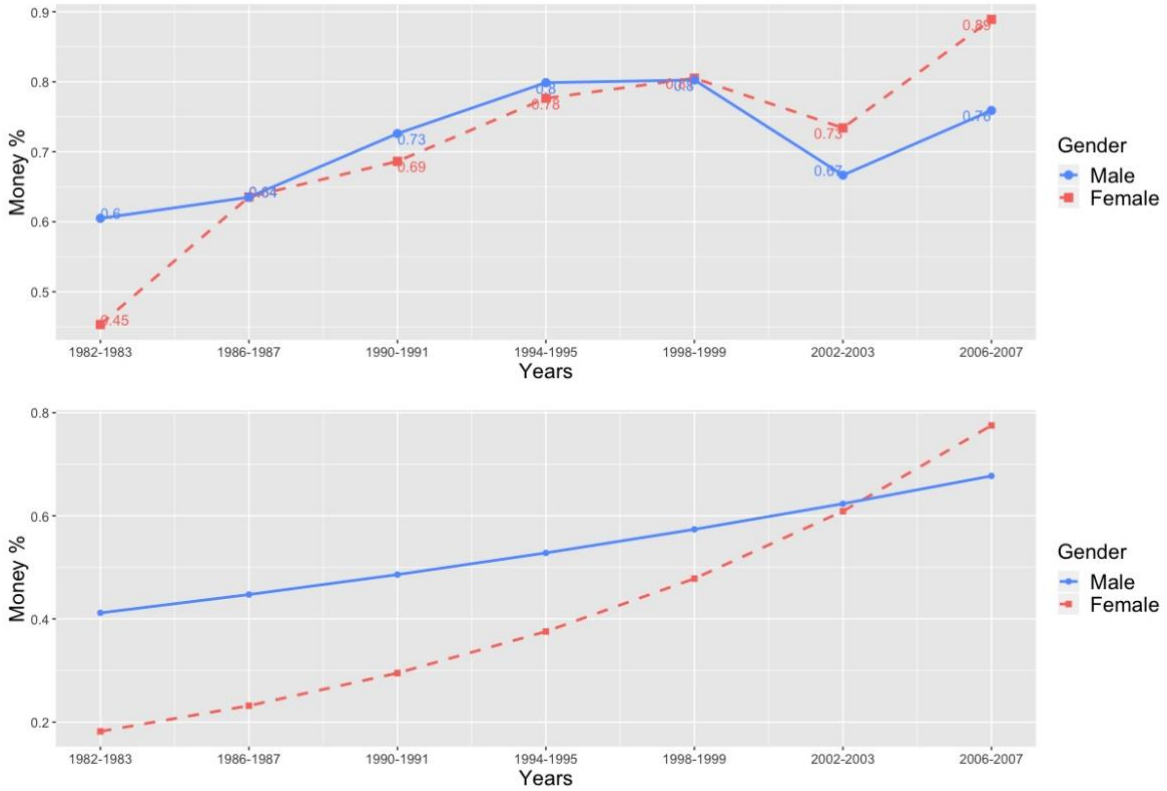


Figure D42. **Frequencies and trend lines of Money by gender over time.**

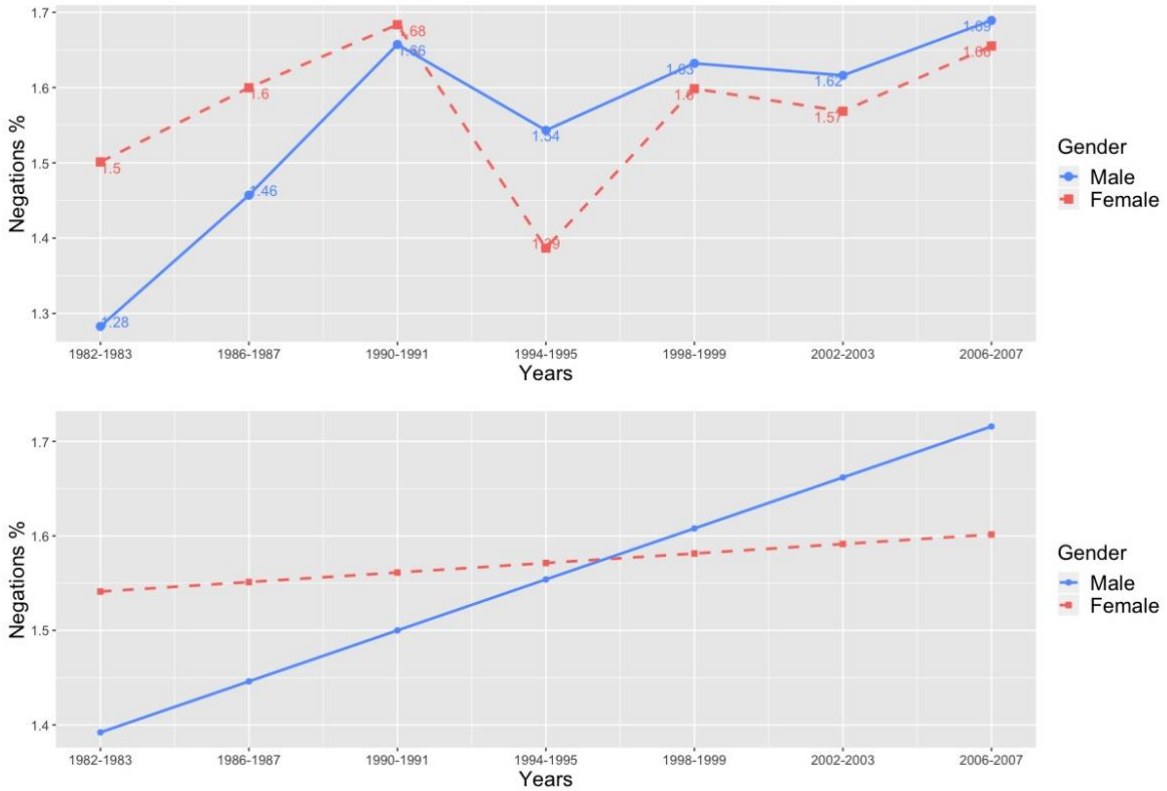


Figure D43. **Frequencies and trend lines of Negation by gender over time.**

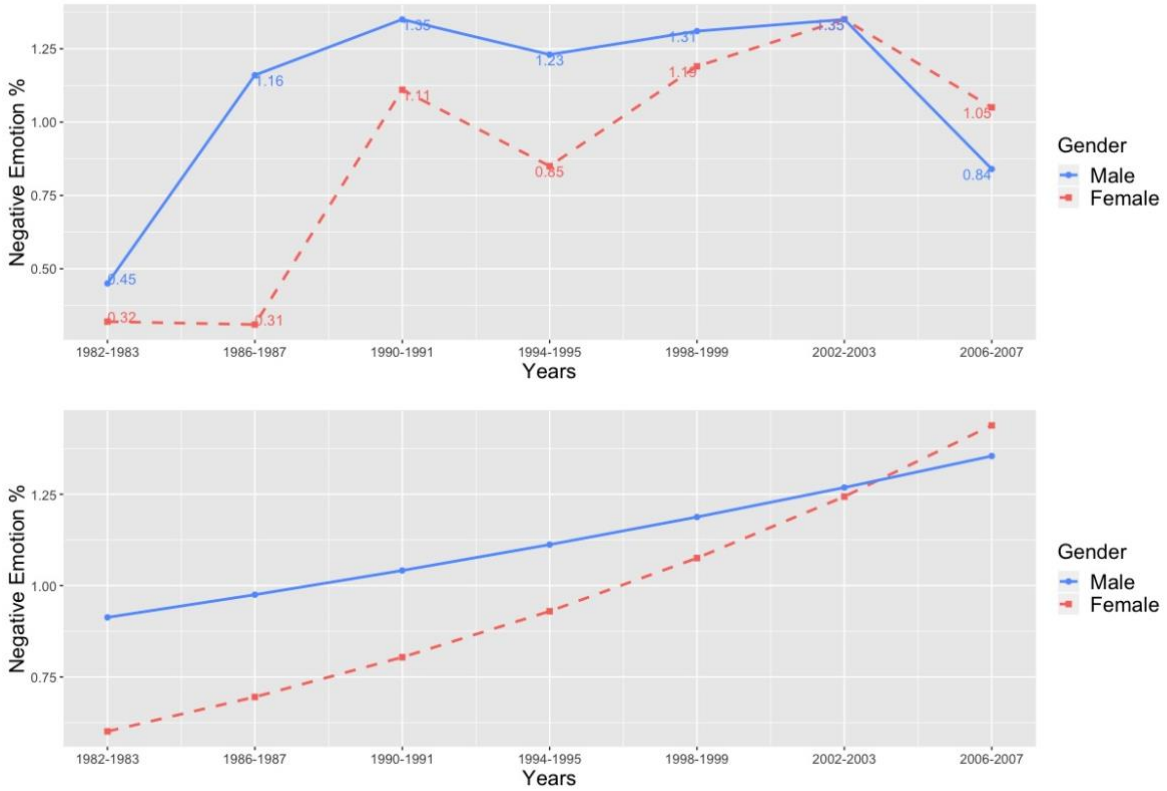
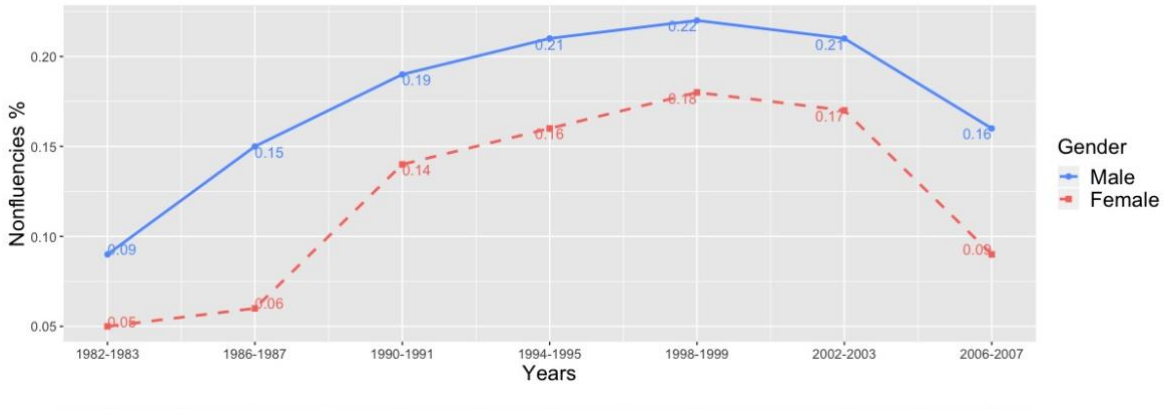


Figure D44. **Frequencies and trend lines of Negative Emotion by gender over time.**



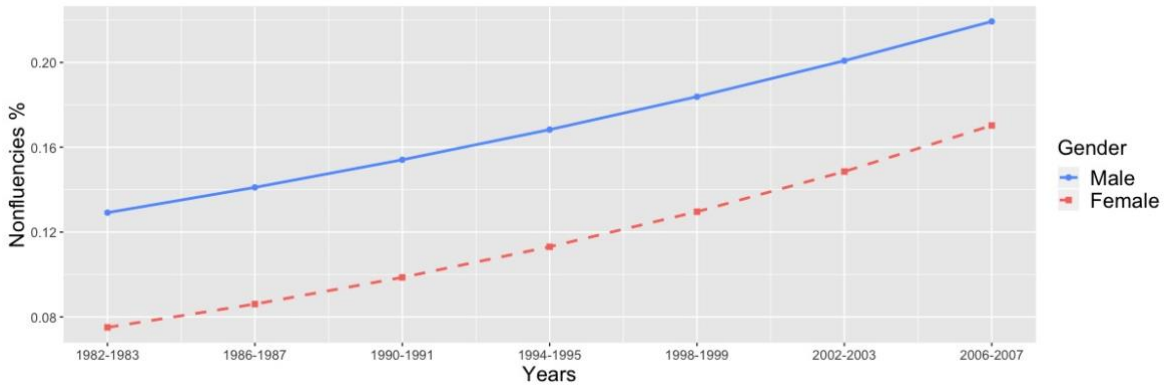


Figure D45. **Frequencies and trend lines of Nonfluencies by gender over time.**

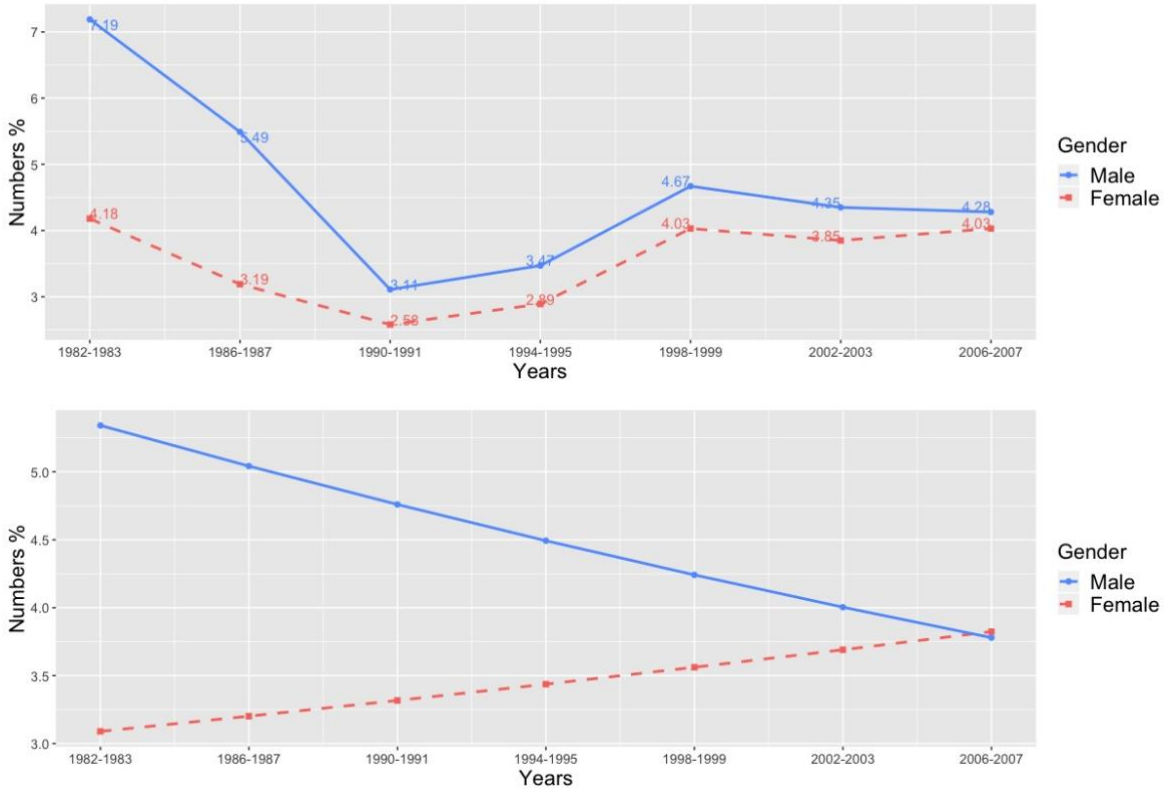


Figure D46. **Frequencies and trend lines of Numbers by gender over time.**

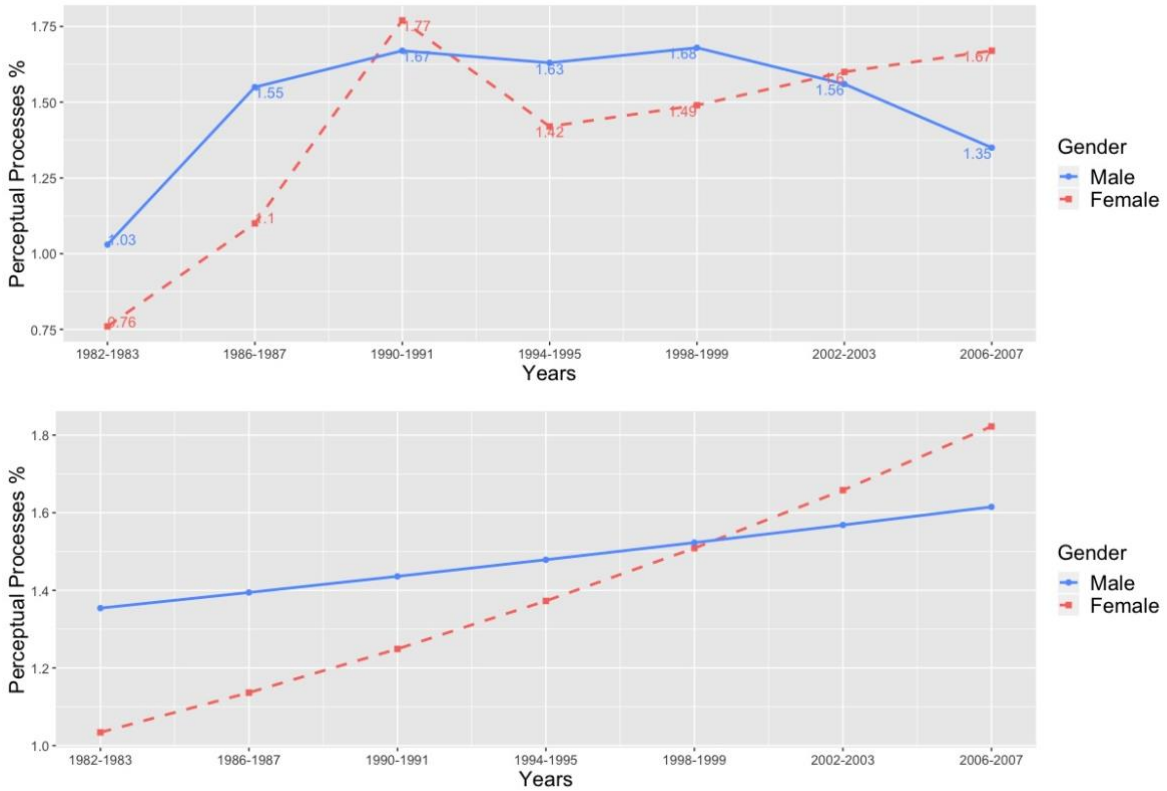


Figure D47. **Frequencies and trend lines of Perceptual Processes by gender over time.**

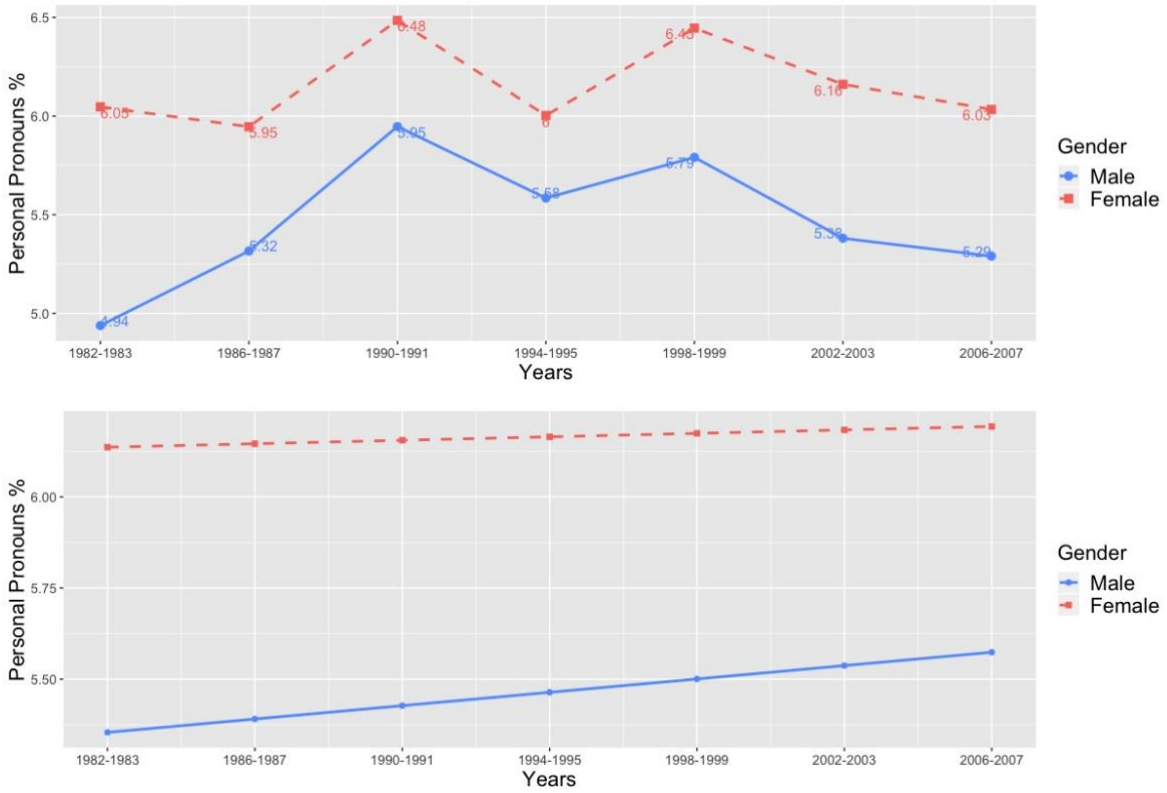


Figure D48. **Frequencies and trend lines of Personal Pronouns by gender over time.**

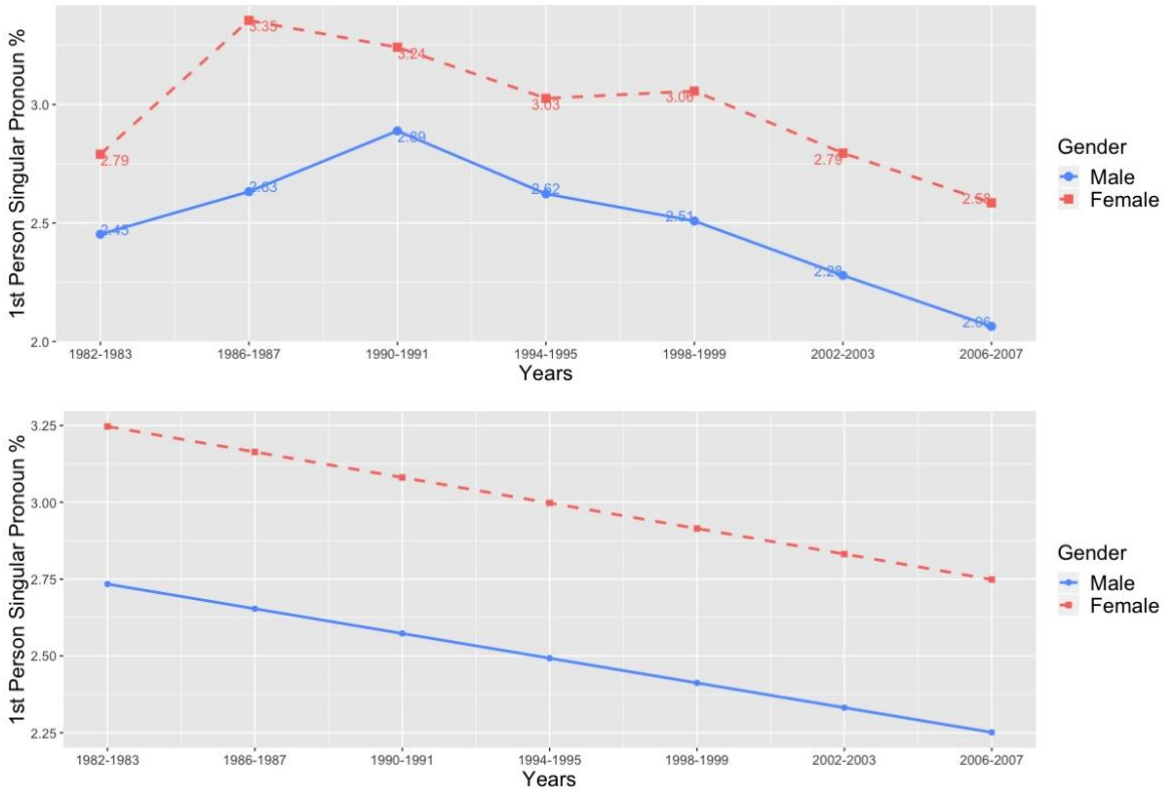


Figure D49. **Frequencies and trend lines of 1st Person Singular Pronoun by gender over time.**

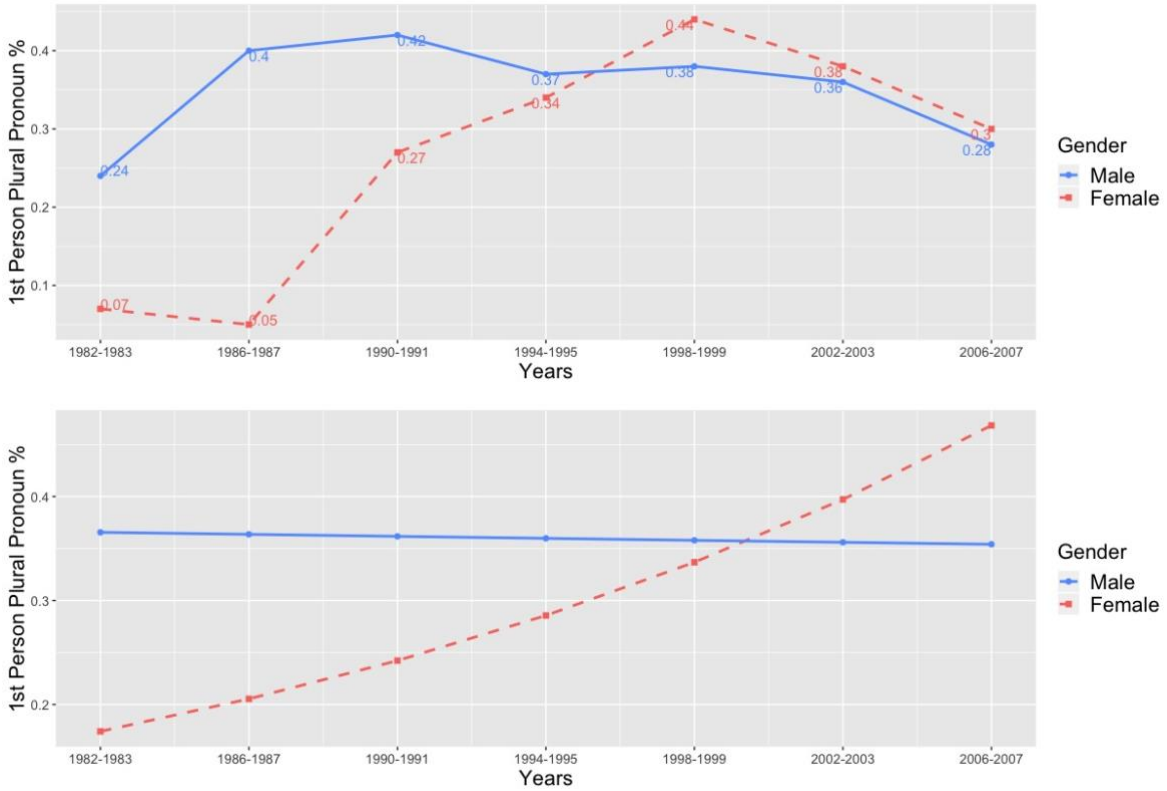
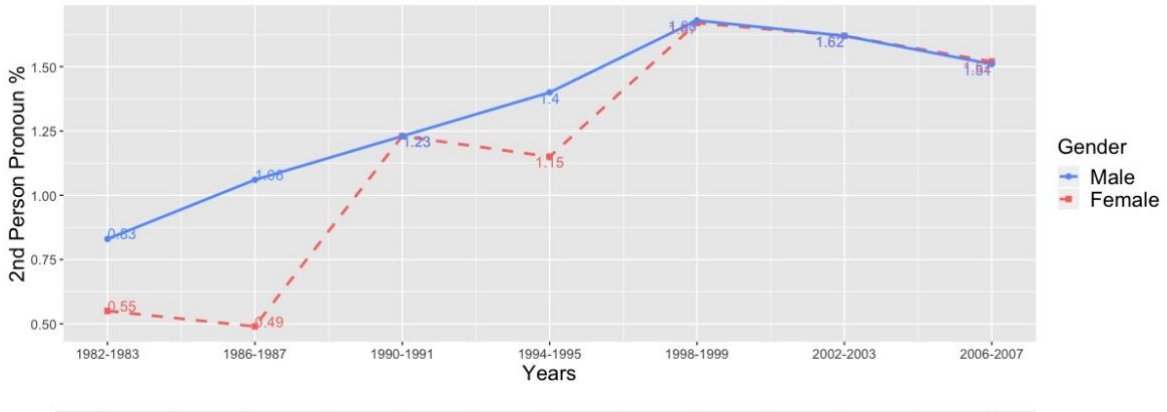


Figure D50. **Frequencies and trend lines of 1st Person Plural Pronoun by gender over time.**



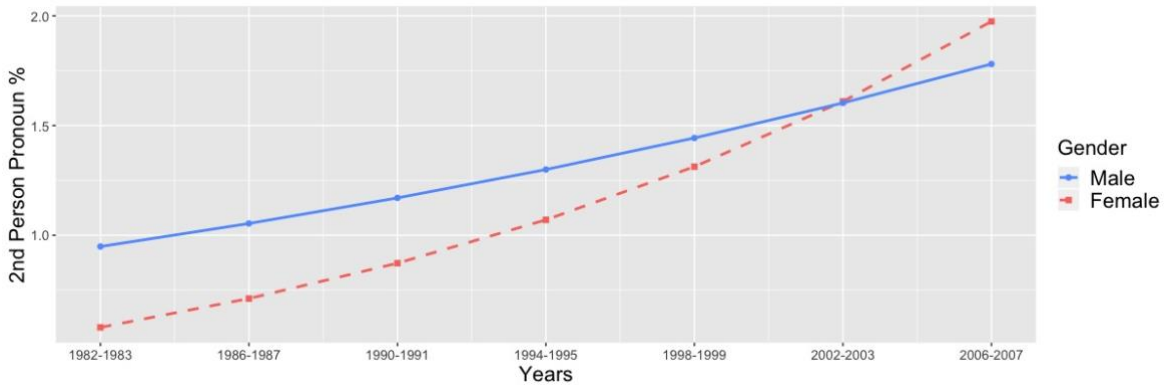


Figure D51. **Frequencies and trend lines of 2nd Person Pronoun by gender over time.**

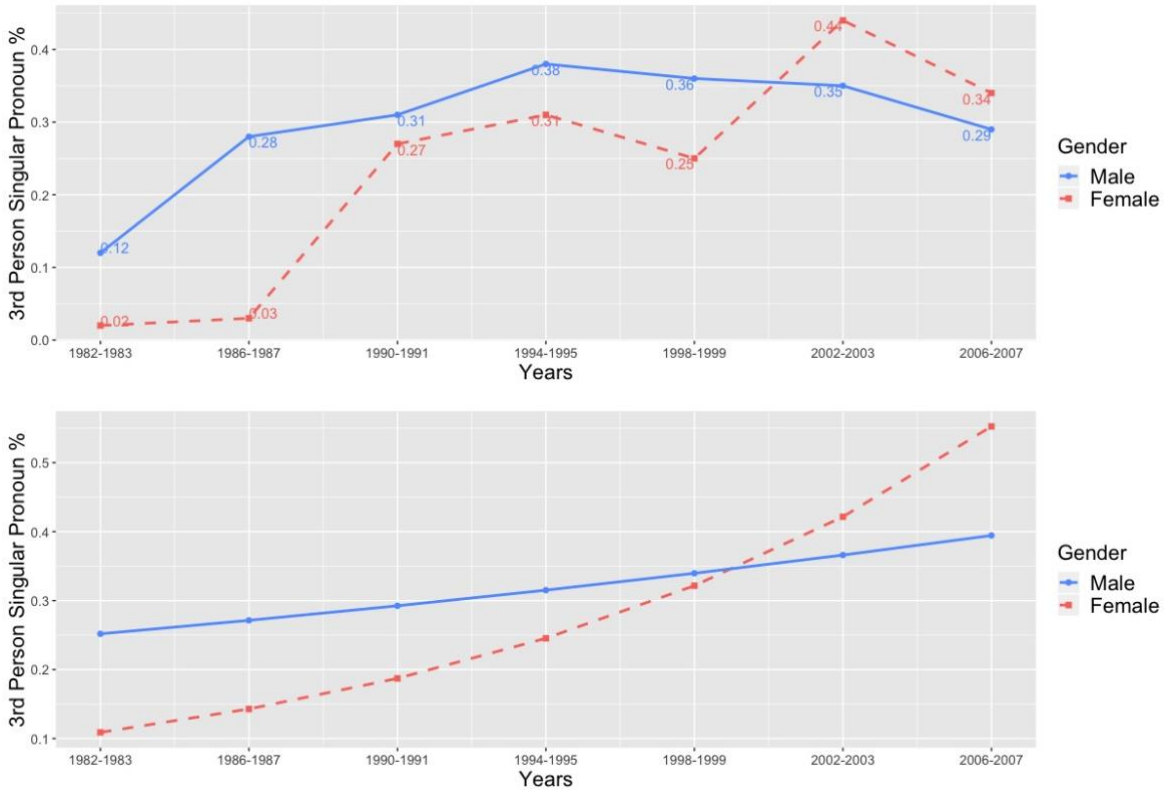


Figure D52. **Frequencies and trend lines of 3rd Person Singular Pronoun by gender over time.**

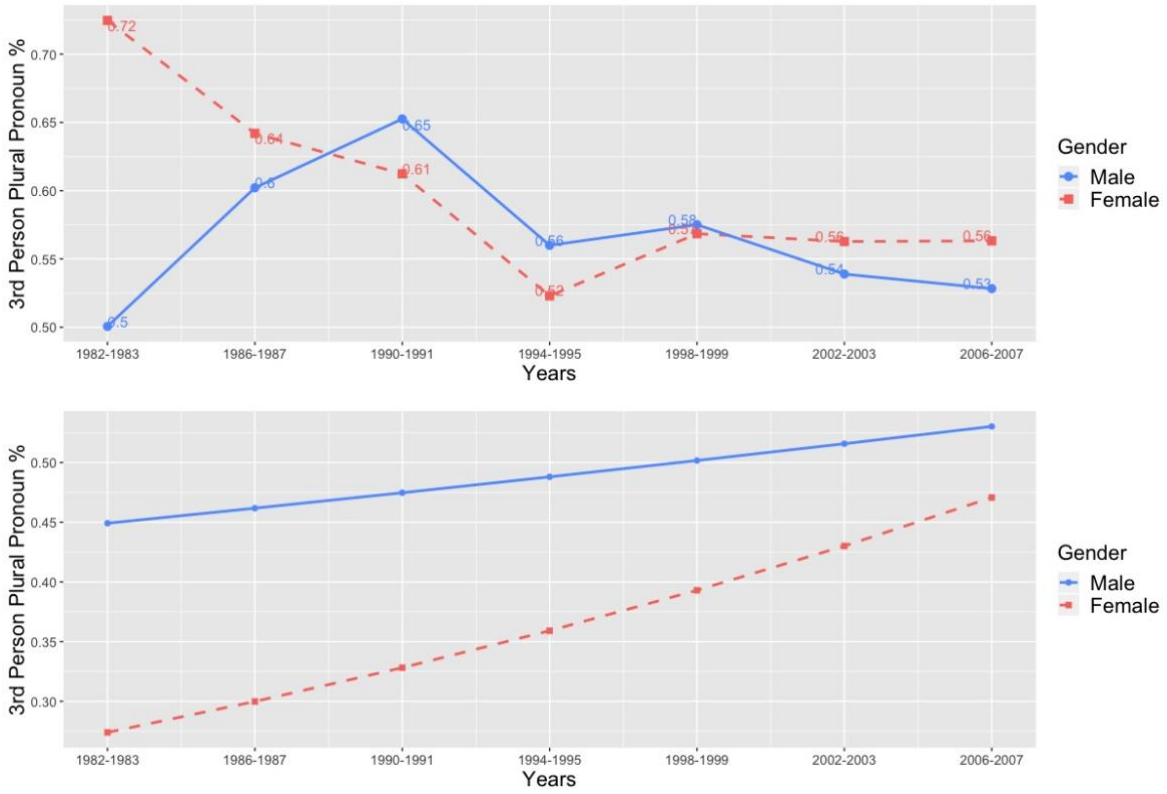


Figure D53. **Frequencies and trend lines of 3rd Person Plural Pronoun by gender over time.**

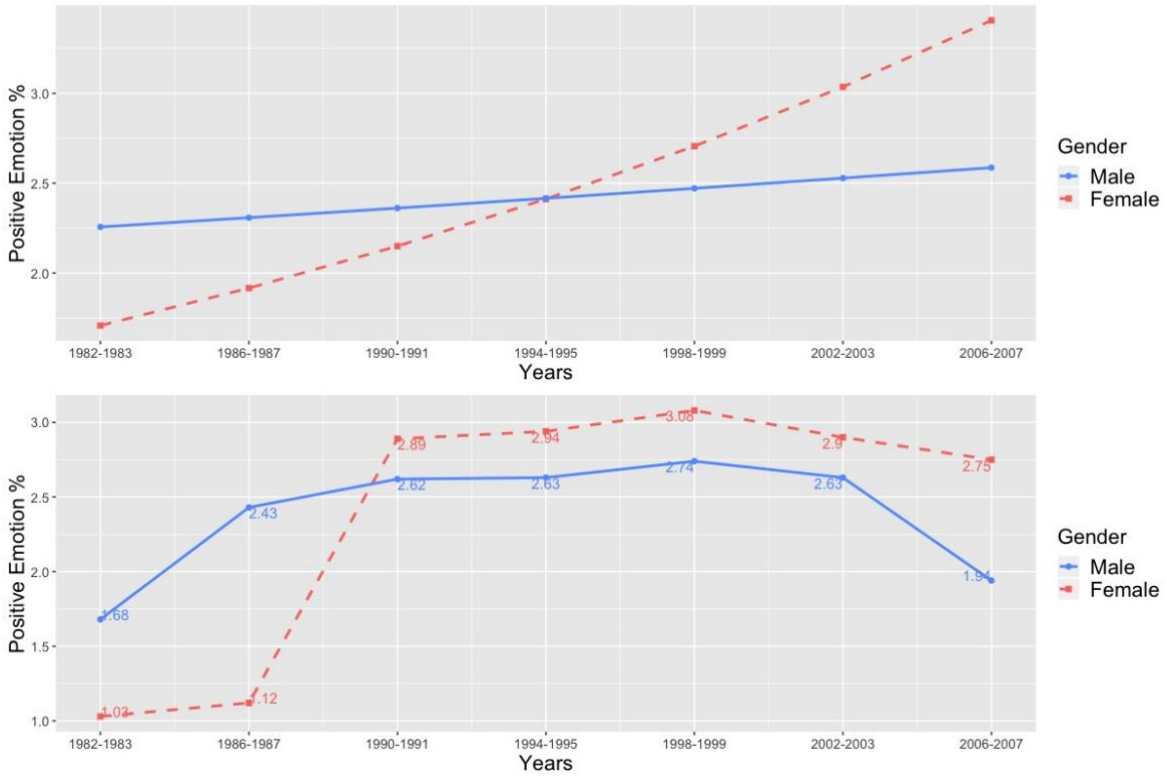


Figure D54. **Frequencies and trend lines of Positive Emotion by gender over time.**

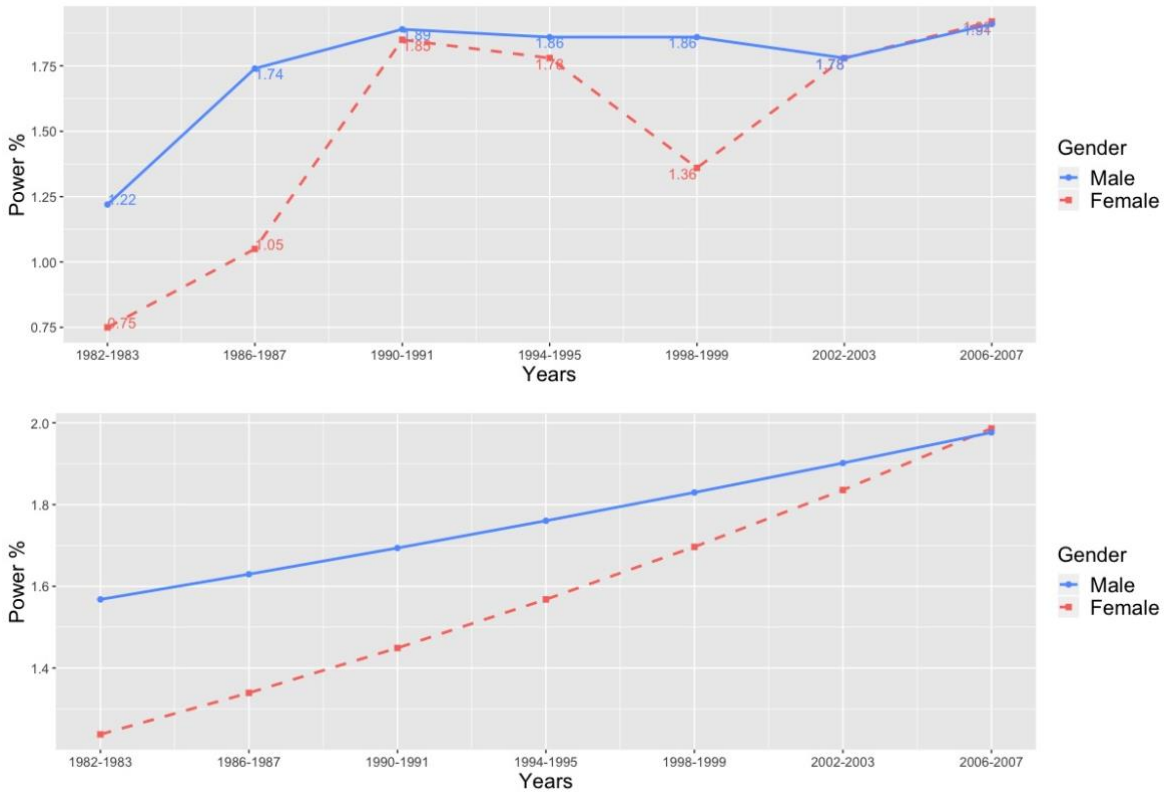


Figure D55. **Frequencies and trend lines of Power by gender over time.**

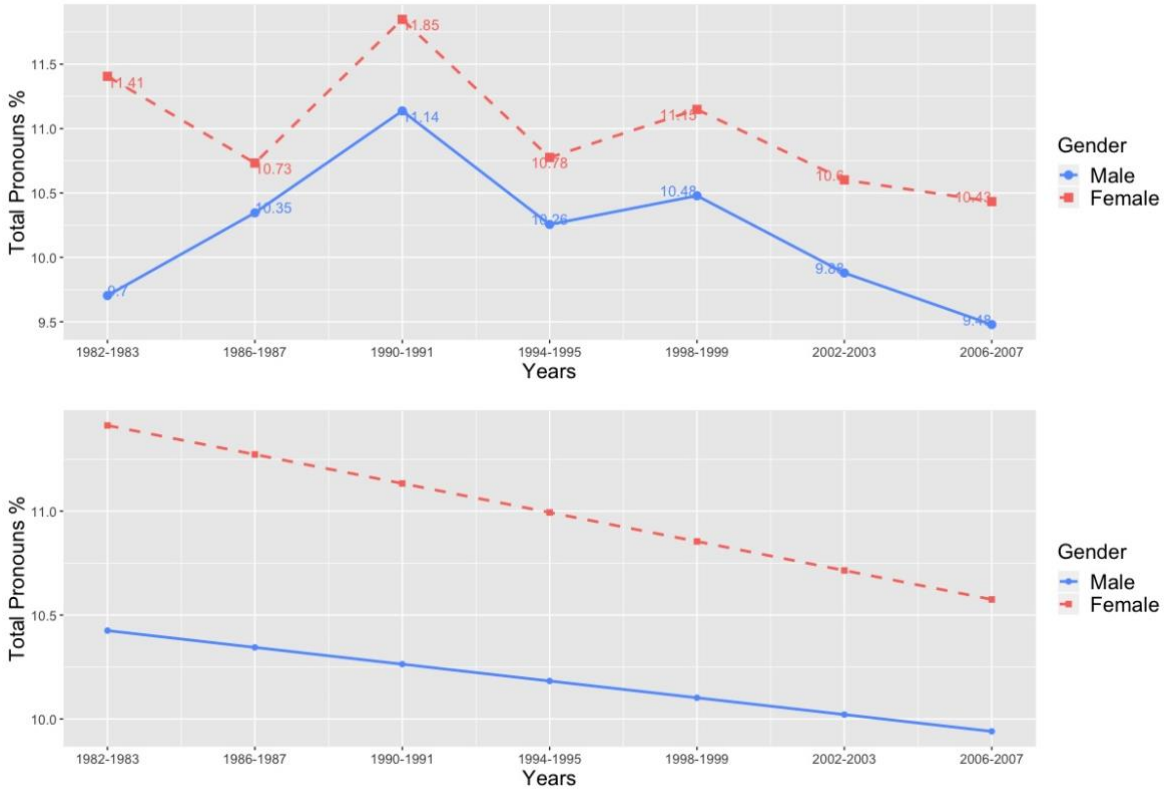


Figure D56. **Frequencies and trend lines of Total Pronouns by gender over time.**

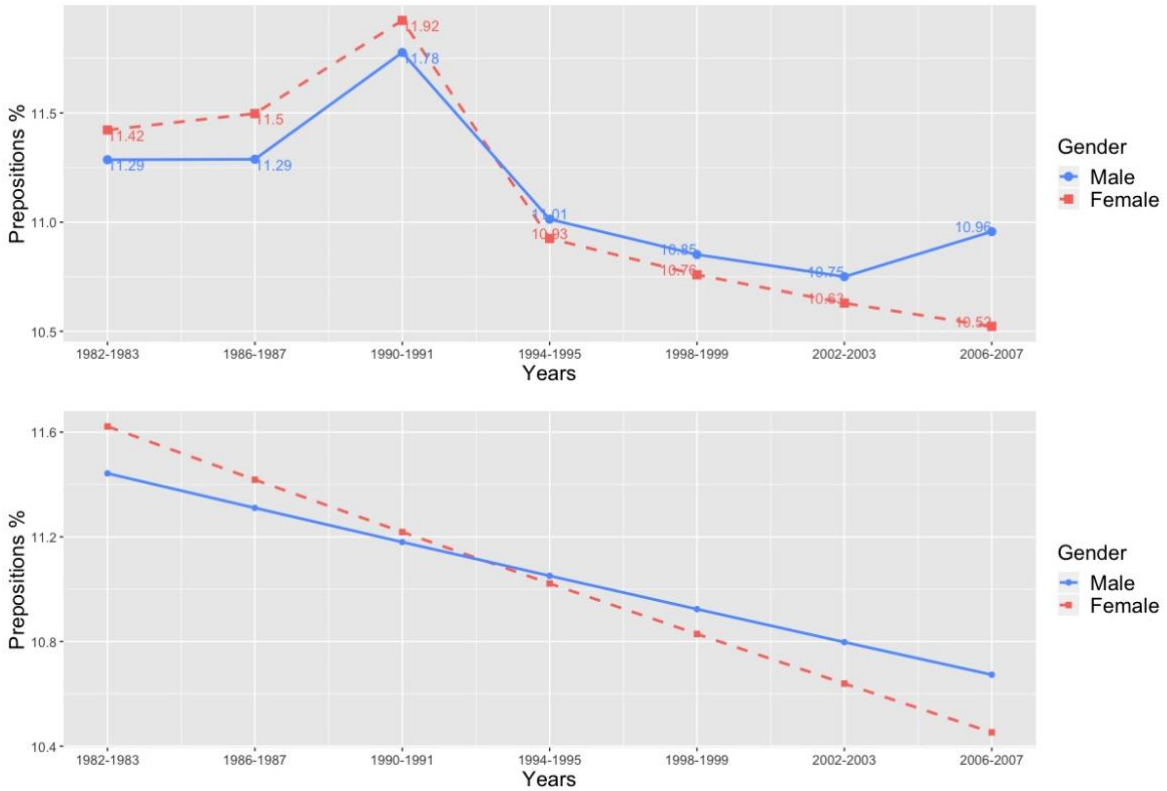


Figure D57. **Frequencies and trend lines of Prepositions by gender over time.**

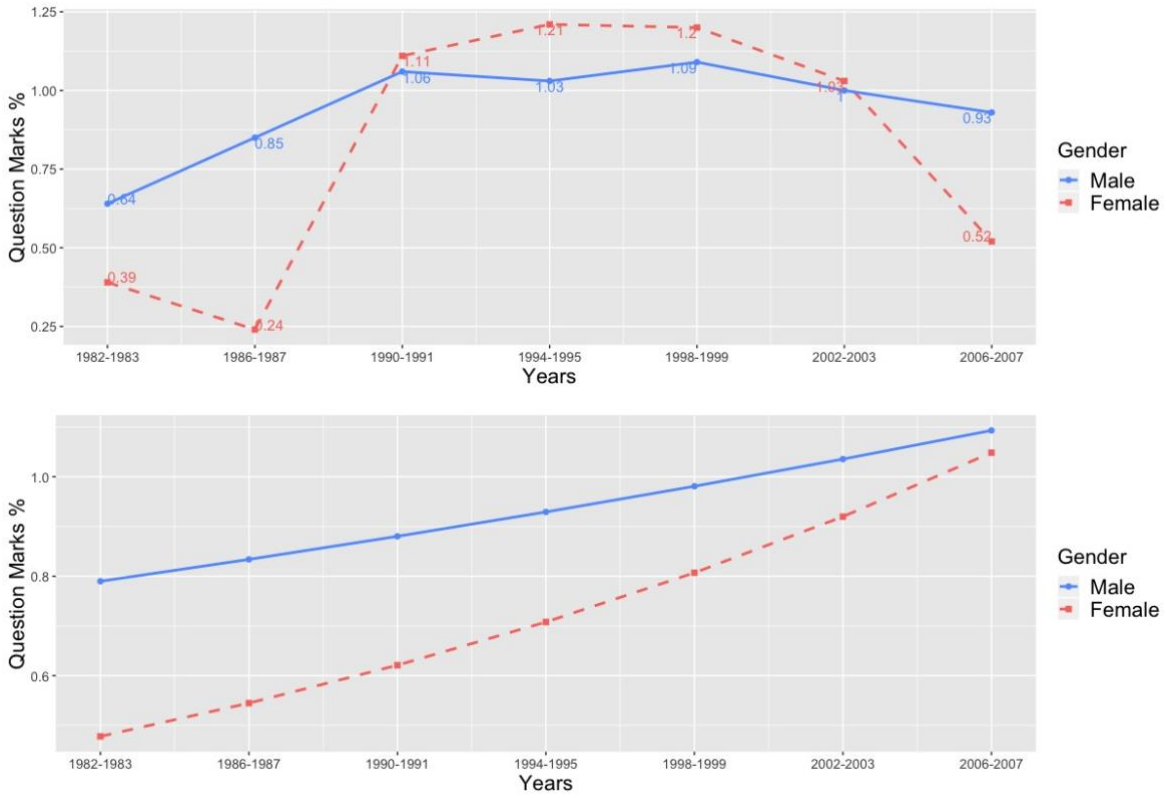


Figure D58. **Frequencies and trend lines of Question Marks by gender over time.**

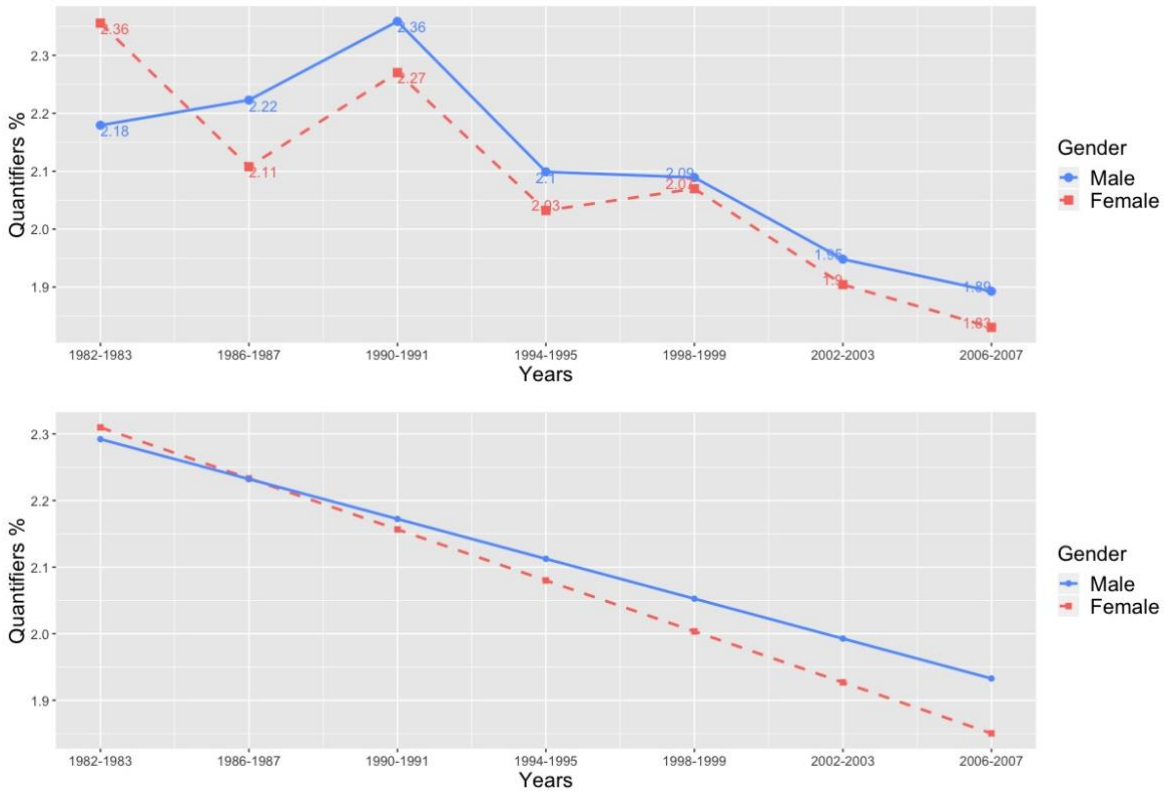


Figure D59. **Frequencies and trend lines of Quantifiers by gender over time.**

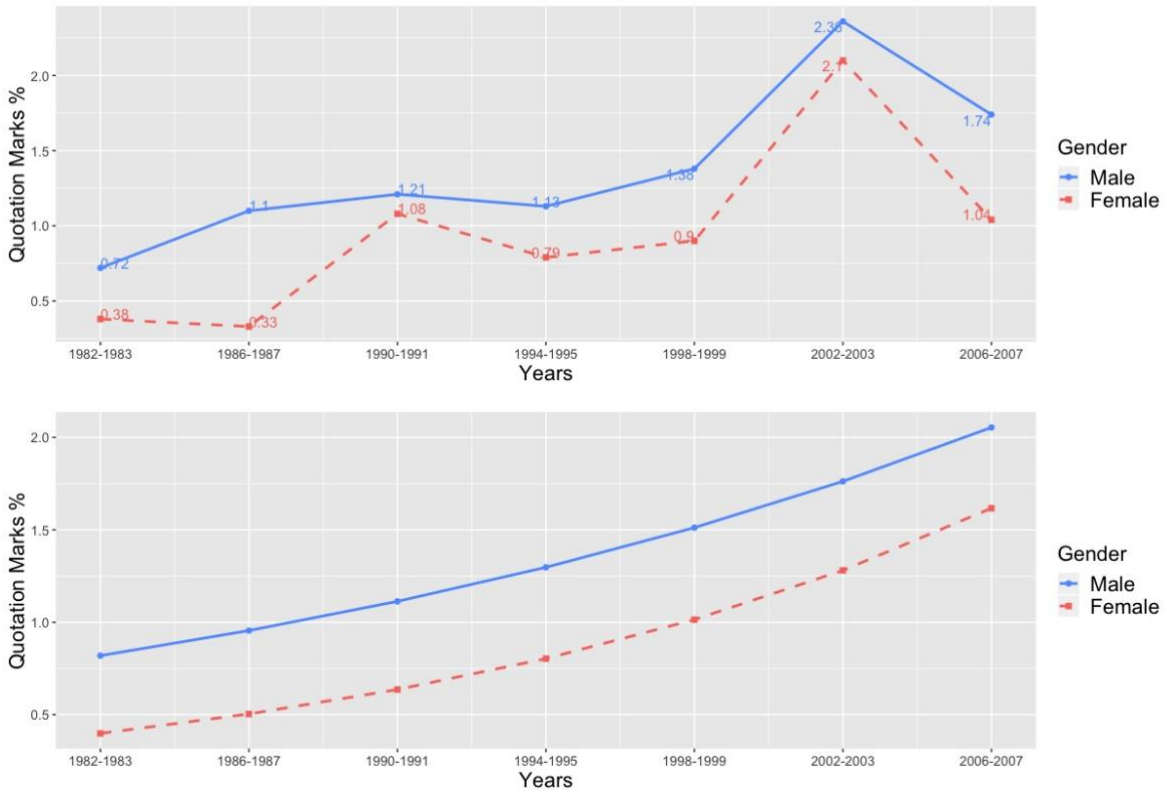


Figure D60. **Frequencies and trend lines of Quotation Marks by gender over time.**

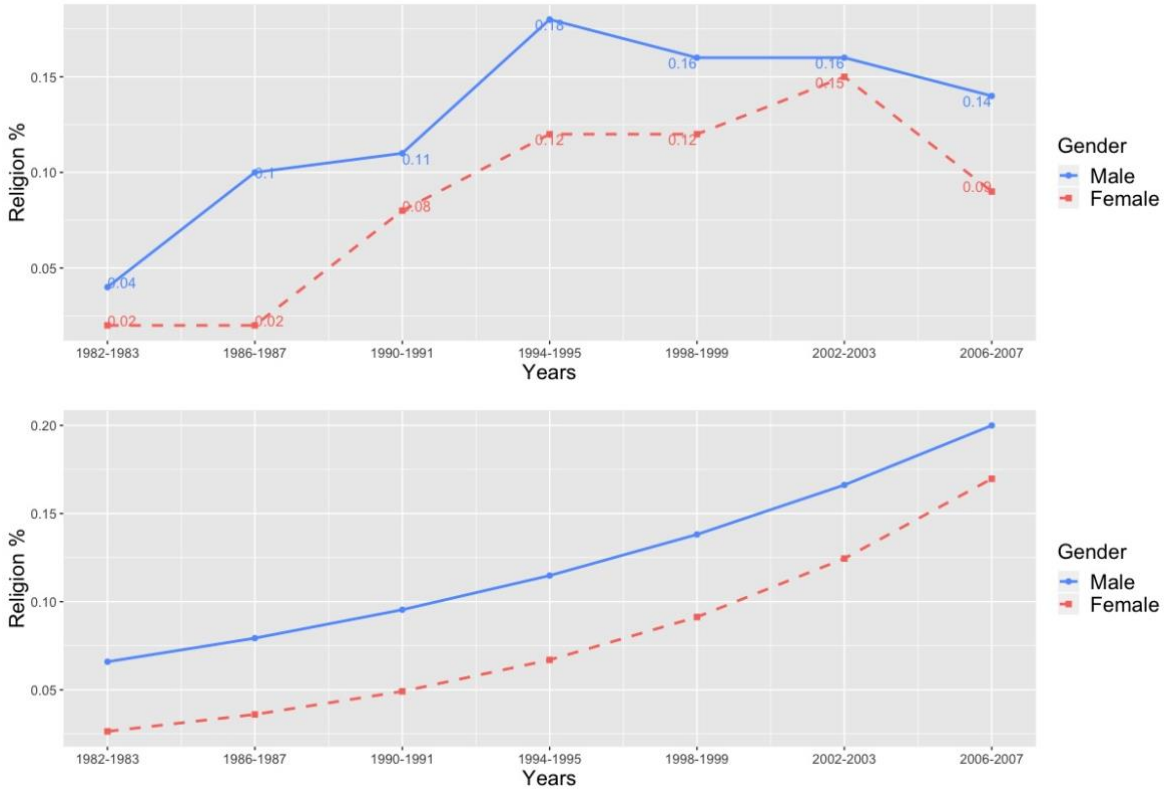


Figure D61. **Frequencies and trend lines of Religion by gender over time.**

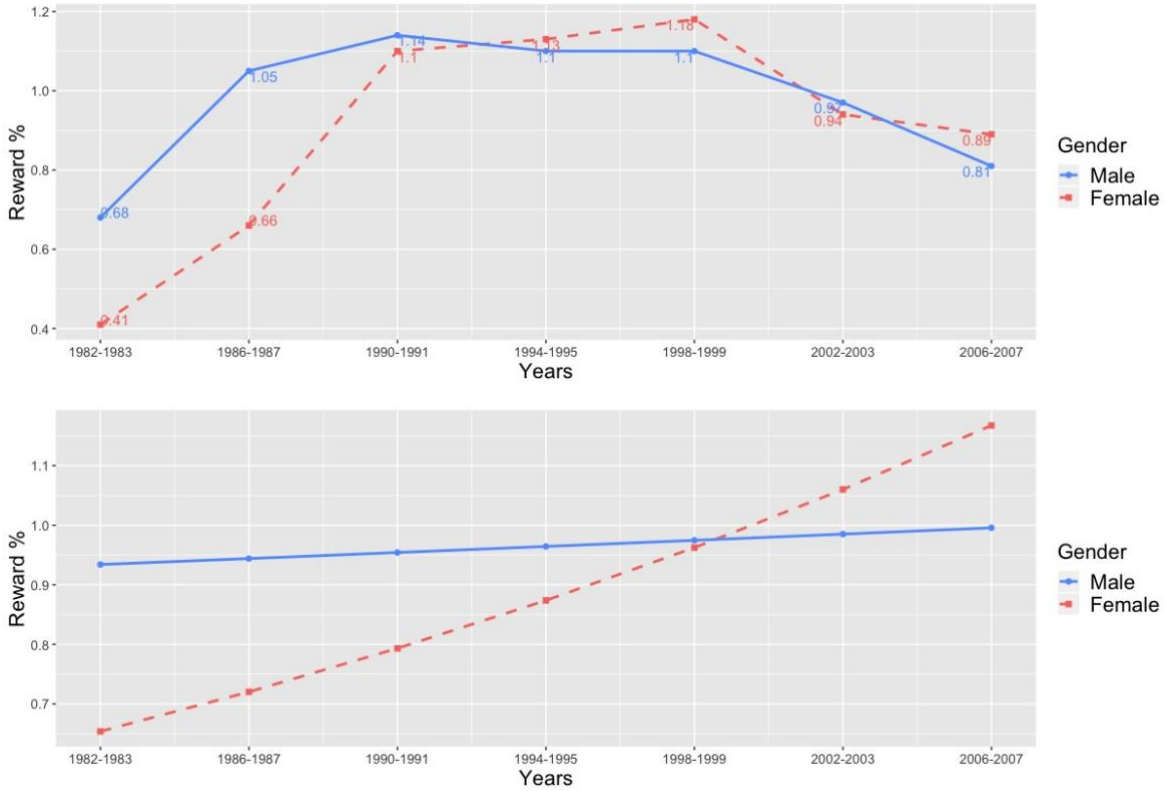


Figure D62. **Frequencies and trend lines of Reward by gender over time.**

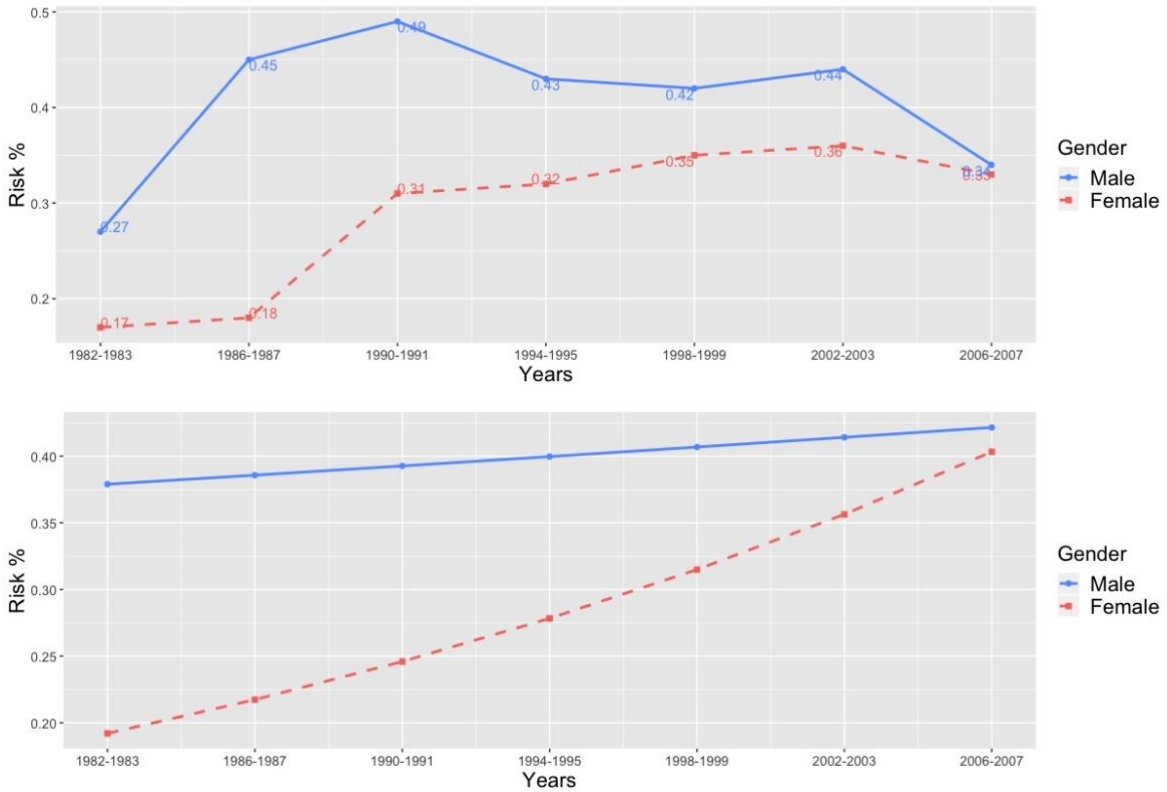


Figure D63. **Frequencies and trend lines of Risk by gender over time.**

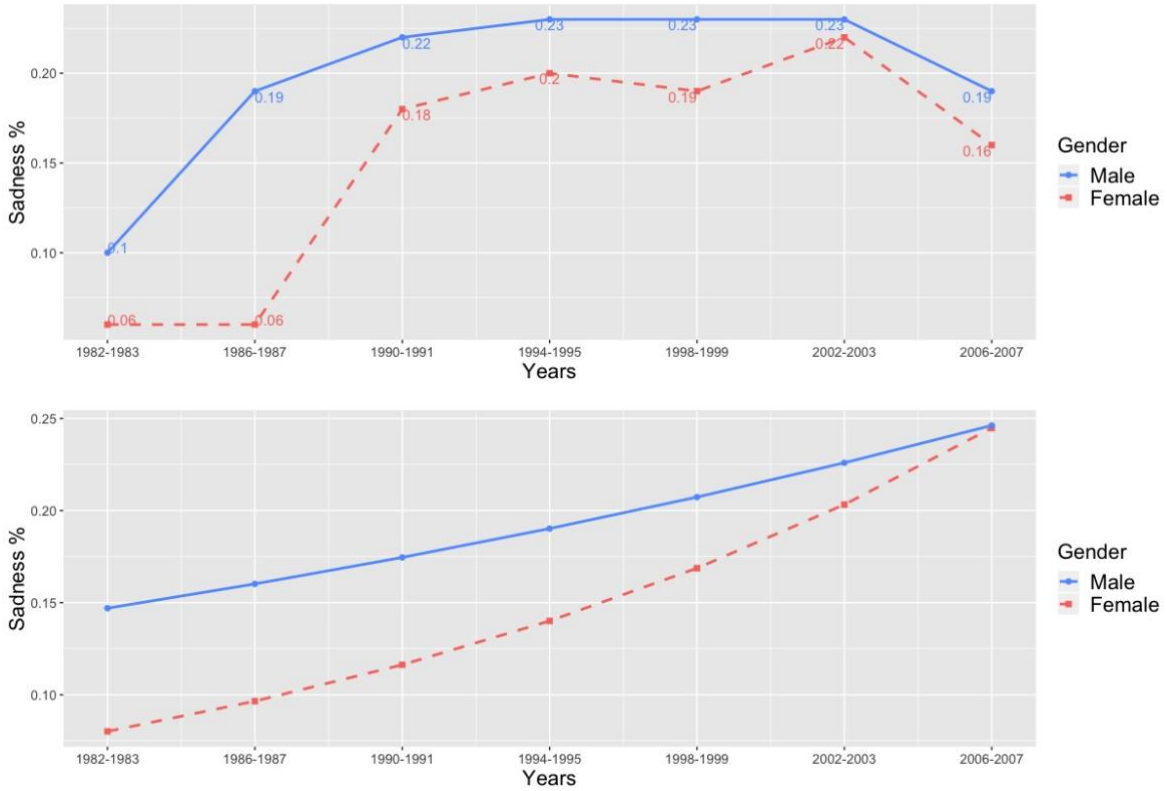


Figure D64. **Frequencies and trend lines of Sadness by gender over time.**

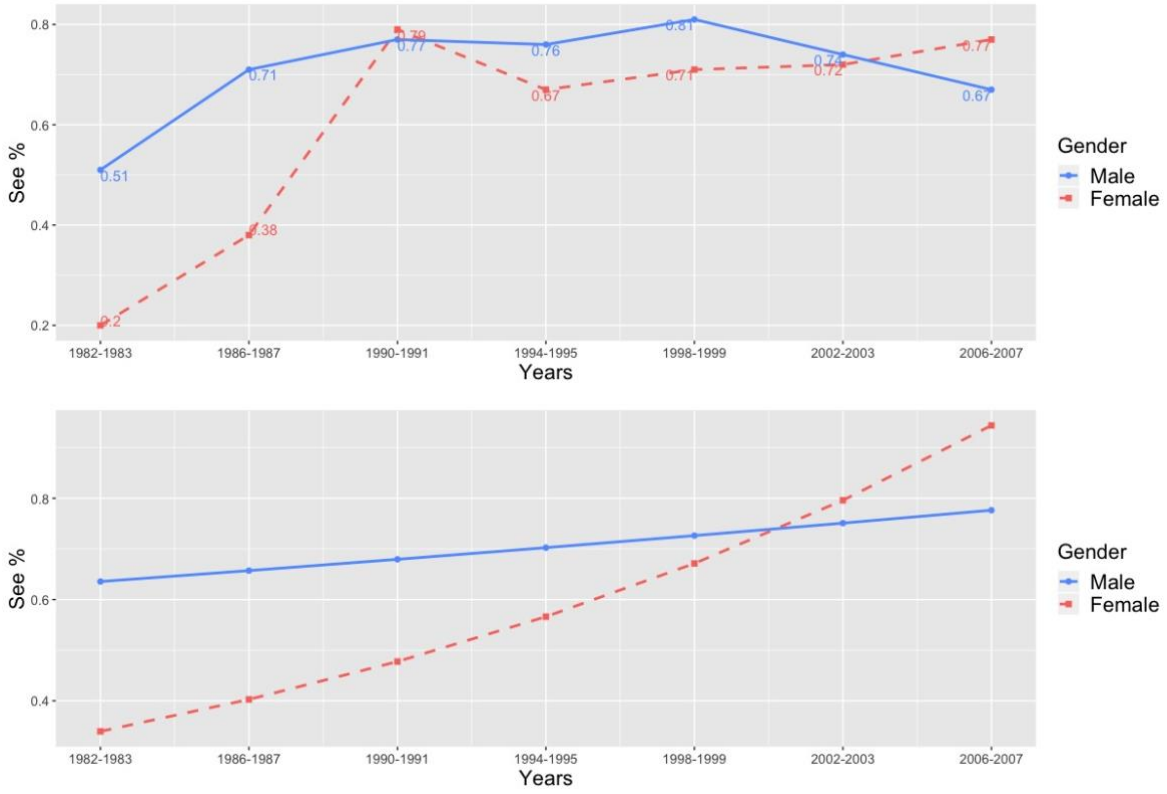


Figure D65. **Frequencies and trend lines of See by gender over time.**

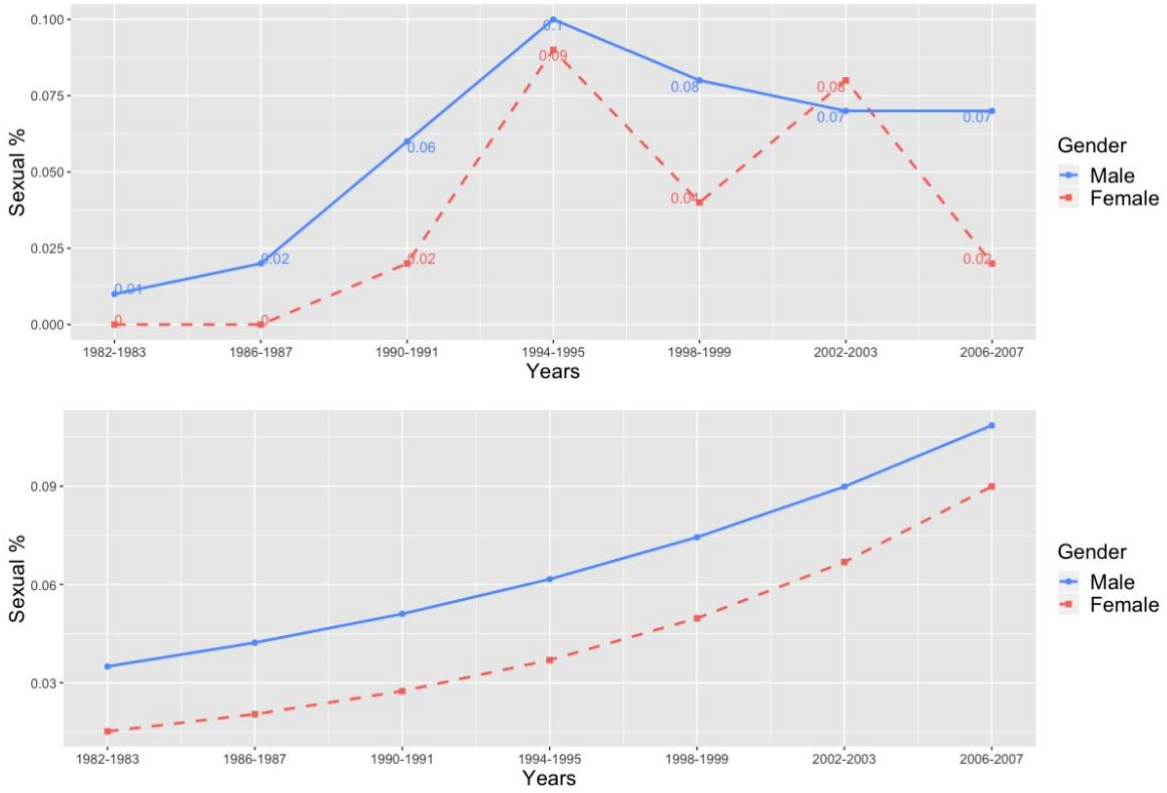


Figure D66. **Frequencies and trend lines of Sexual Words by gender over time.**

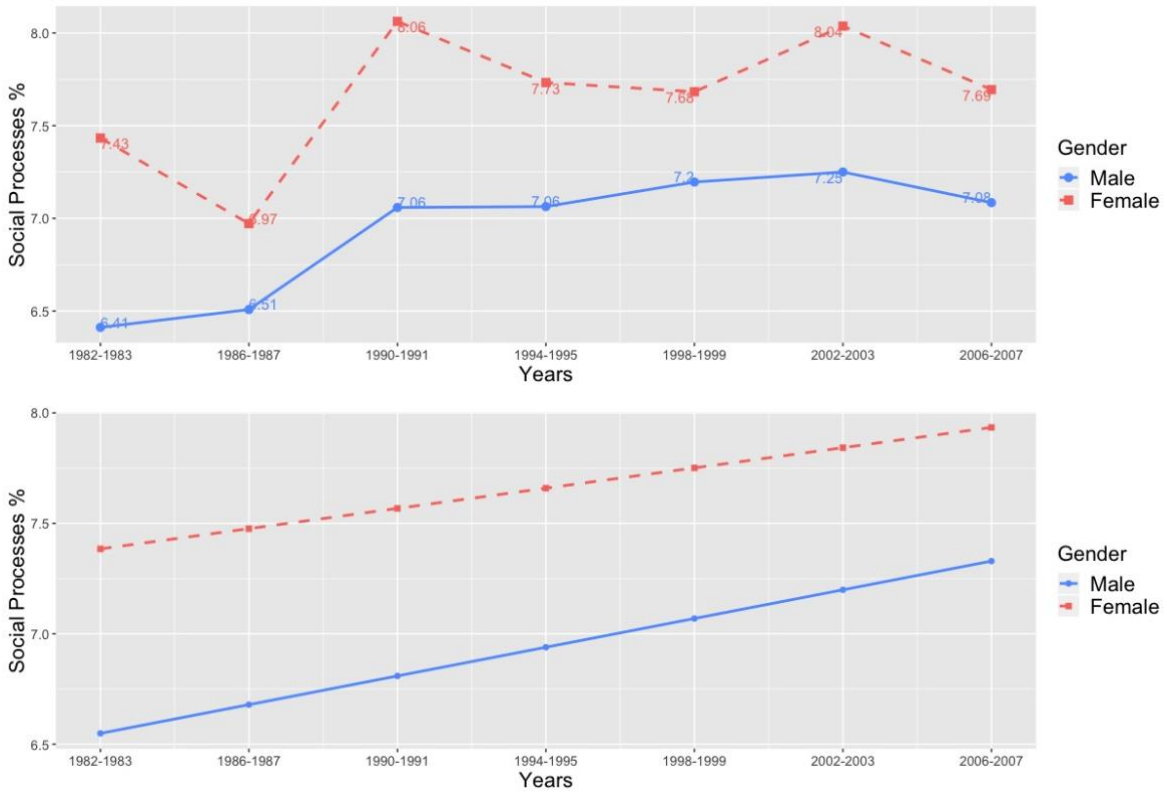


Figure D67. **Frequencies and trend lines of Social Processes by gender over time.**

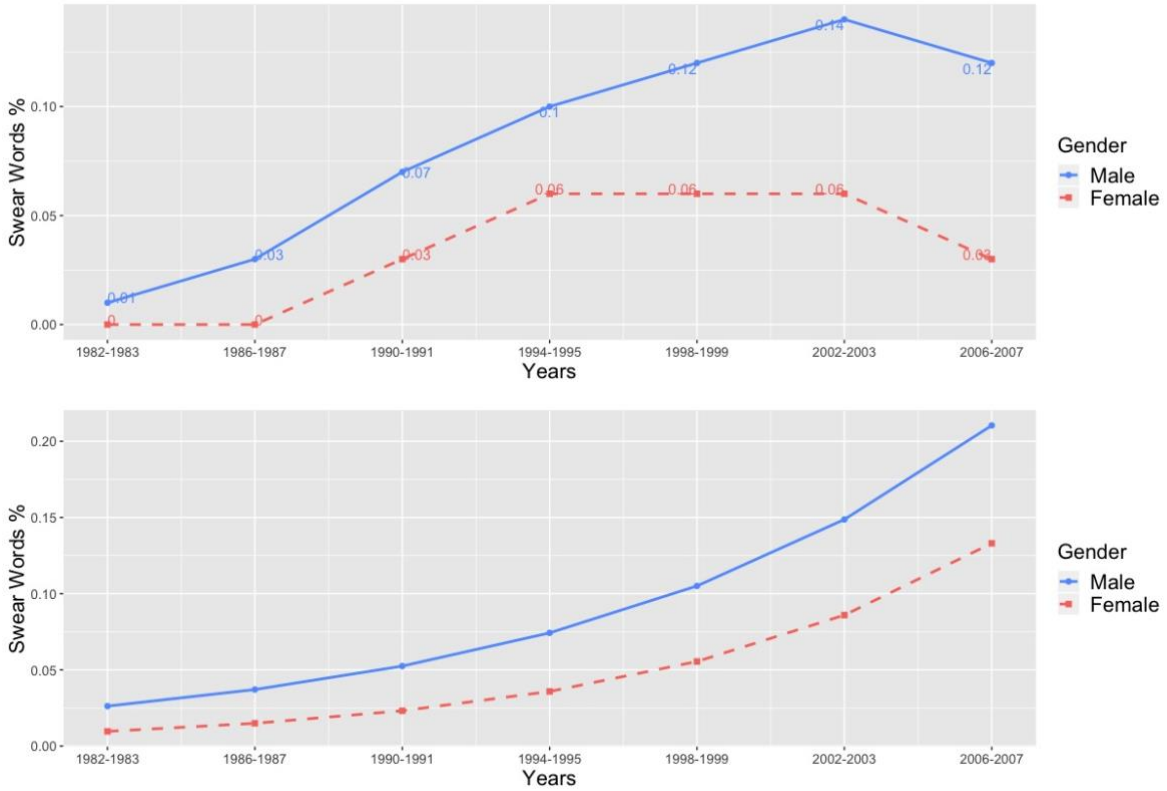


Figure D68. **Frequencies and trend lines of Swear Words by gender over time.**

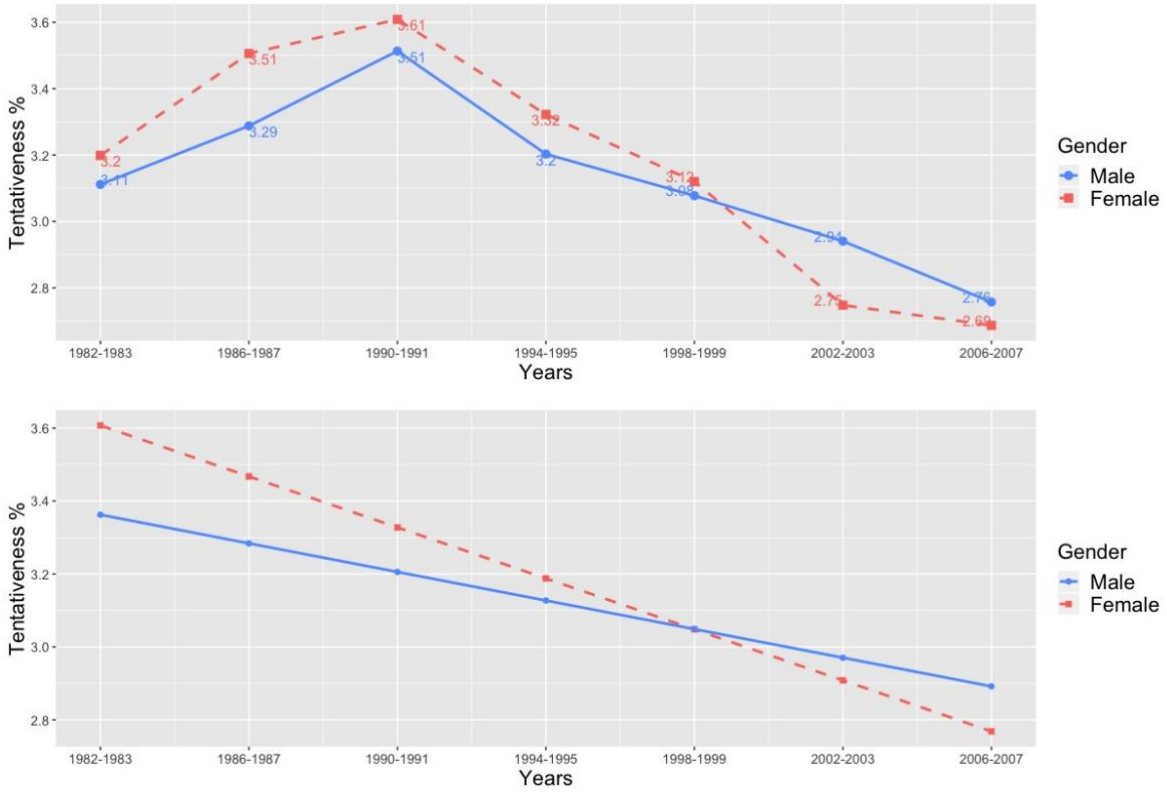


Figure D69. **Frequencies and trend lines of Tentativeness by gender over time.**

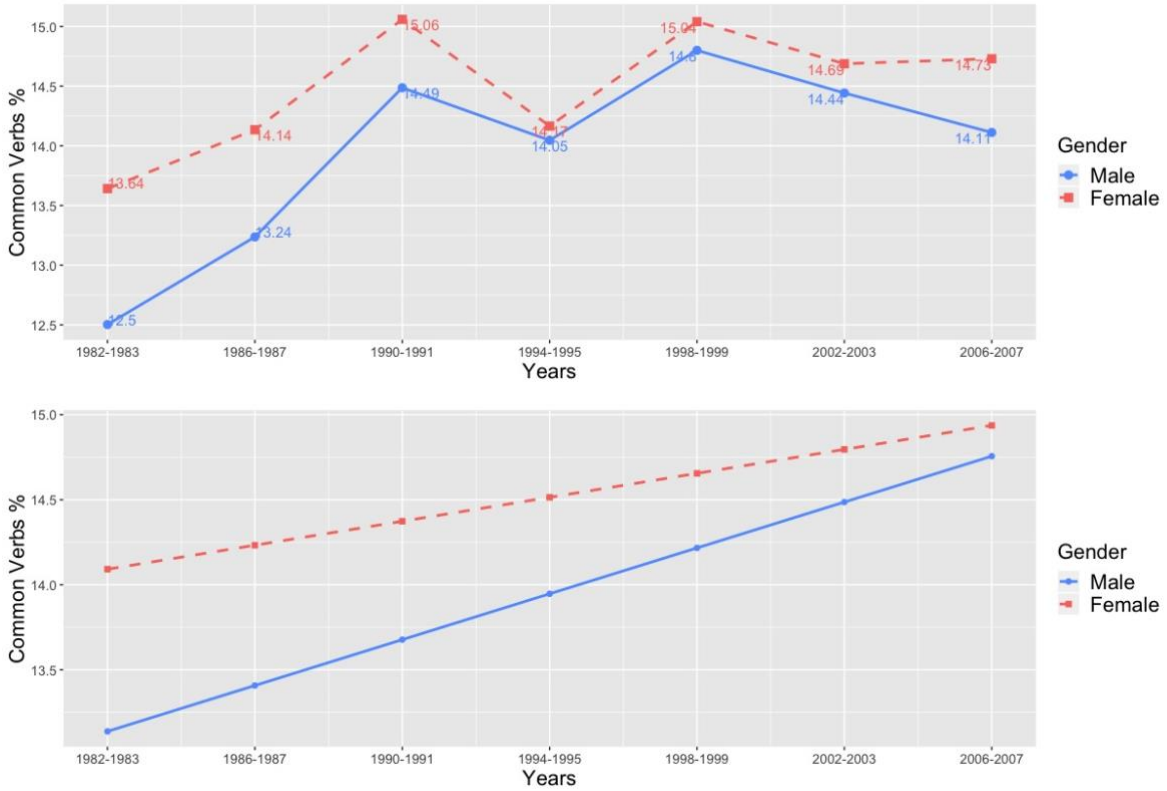
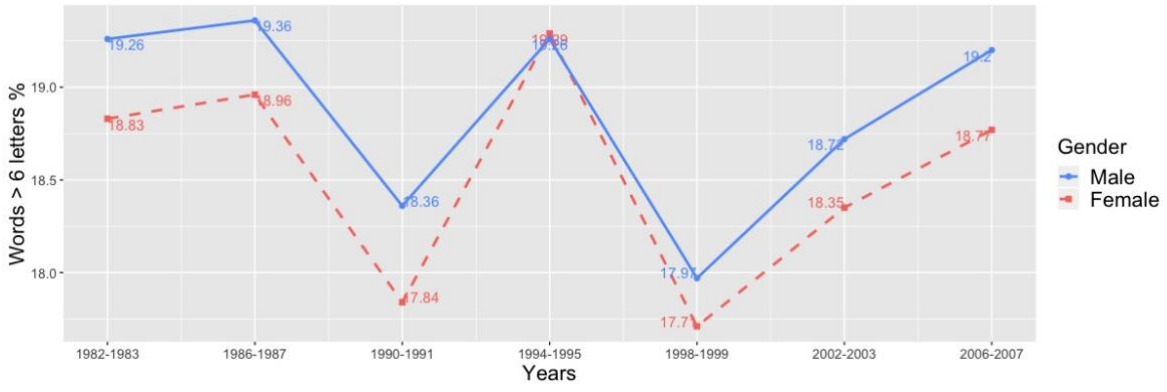


Figure D70. **Frequencies and trend lines of Common Verbs by gender over time.**



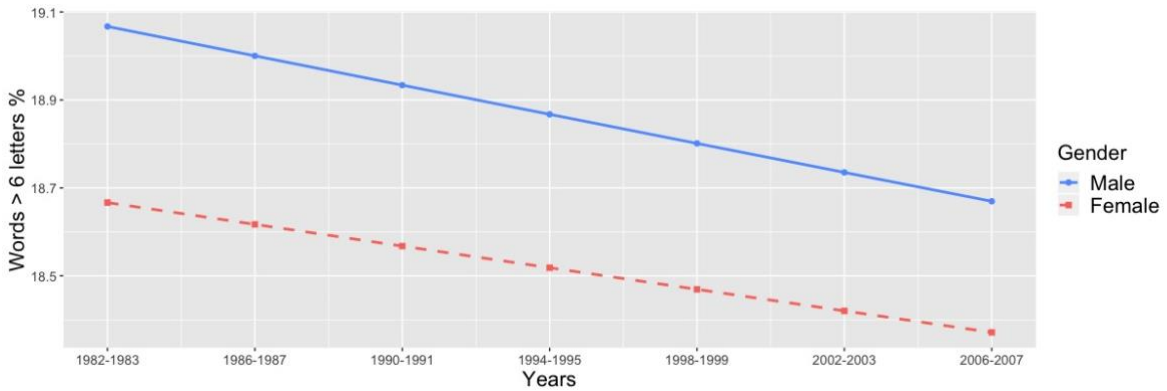


Figure D71. **Frequencies and trend lines of Words > 6 Letters by gender over time.**

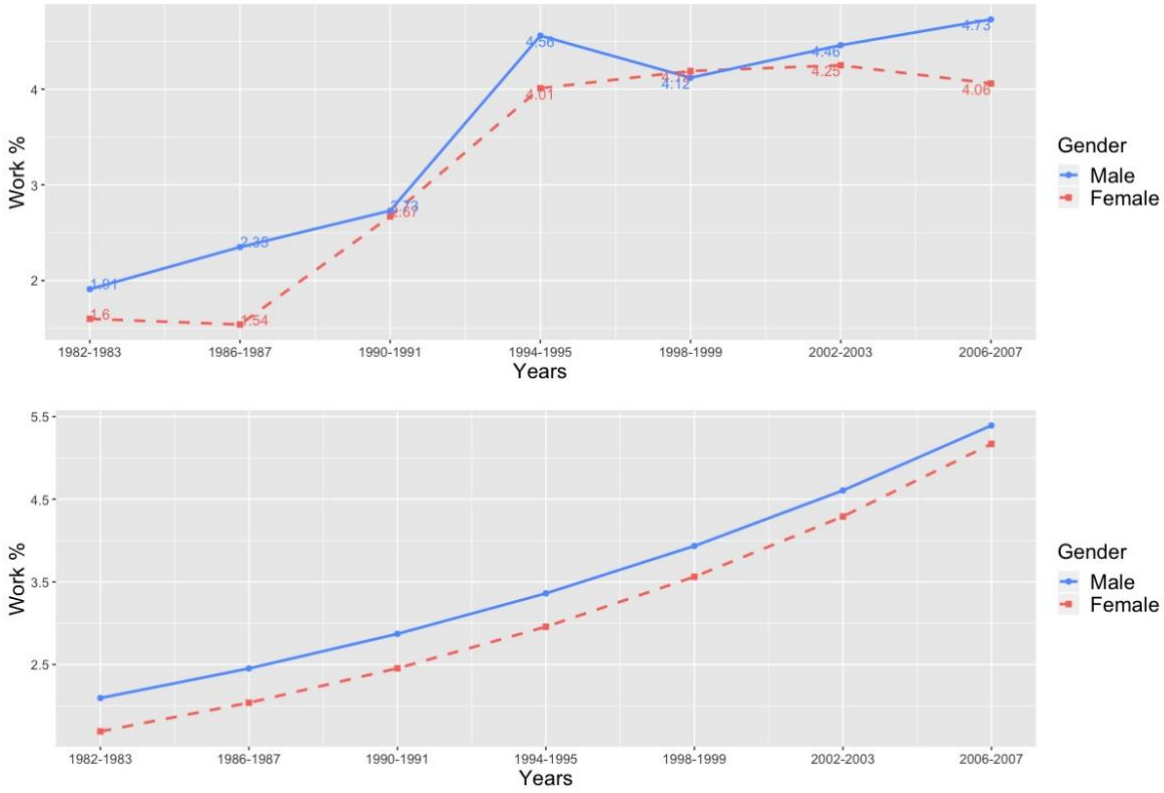


Figure D72. **Frequencies and trend lines of Work by gender over time.**

Elli E. Bourlai

700 N. Woodlawn Avenue
Bloomington, IN 47408 ebourlai.com ebourlai@iu.edu

EDUCATION

Ph.D., Information Science December 2018
(Minor: Computational Linguistics)
Department of Information and Library Science,
School of Informatics, Computing and Engineering,
Indiana University Bloomington.

M.A., Historical Language Studies (Research Track) 2009
(Awarded with Distinction)
University of Sheffield, UK.

B.A., English Language and Literature 2008
(Specialization: Theoretical Linguistics)
Aristotle University of Thessaloniki, Greece.

RESEARCH COMMITTEE MEMBERS

Susan C. Herring, Professor of Information Science (Chair)
Markus Dickinson, Associate Professor of Linguistics (Minor advisor)
Pnina Fichman, Professor of Information Science
Allen Riddell, Assistant Professor of Information Science

PROJECTS AND PUBLICATIONS

- Bourlai, E. E. (in preparation). *Gender and language in CMD: A historical analysis of USENET newsgroups* (Dissertation Thesis).
- Bourlai, E. E., & Gao, Z. (in preparation). The Historical Usenet Newsgroups Corpus (HUNC): A diachronic CMD corpus.
- Bourlai, E. E., Herring, S. C., & Abdul-Mageed, M. (in preparation). Distinguishing functional types of hashtags: A structural approach.
- Bourlai, E. E. (2017). "Comments in tags please!": Tagging practices on Tumblr (Special Issue: Discourse of Social Tagging). *Discourse, Context, and Media*, 22, 46-56.
- Bourlai E. E., & Herring, S. C. (2014). Multimodal communication on Tumblr: "I have so many feels!" *Proceedings of WebSci'14*, June 23-26, Bloomington, IN.

PRESENTATIONS

- Bourlai, E. E., Herring, S. C., & Abdul-Mageed, M. (2016). Distinguishing functional types of hashtags: A structural approach. *AMPRA 2016*, November 4-6, Indiana University Bloomington, US.
- Bourlai, E. E. (2016). "Comments in tags please!": Tagging practices on Tumblr. The 14th ILS Annual Doctoral Research Forum, October 22, 2016, Bloomington, IN.
- Bourlai, E. E. (2016). Gender differences in *soc.men* and *soc.women*: A diachronic study. Diachronic Corpora, Genre, and Language Change, April 8-9, University of Nottingham, UK.
- Bourlai, E., Herring, S.C., & Abdul-Mageed, M. (2015). Distinguishing functional types of hashtags: A structural approach. *Spring 2015 CCMC Symposia*, March 28, Bloomington, IN
- Bourlai, E. E. (2014). Distinguishing topic and evaluative hashtags: A structural approach. *The 12th ILS Annual Doctoral Research Forum*. October 25, 2014, Bloomington, IN.
- Bourlai E. E., & Herring, S. C. (2014). Multimodal communication on Tumblr: "I have so many feels!" *WebSci'14*, June 23-26, Bloomington, IN.
- Bourlai, E. E. (2013). Multimodal communication on Tumblr: "I have so many feels!" *The 11th ILS Annual Doctoral Research Forum*. October 12, 2013, Bloomington, IN.

GUEST LECTURES / INVITED TALKS

- *Introduction to LIWC*, Z641: Computer-Mediated discourse Analysis, January 23, 2018. Indiana University, Bloomington.
- *Introduction to LIWC*, Z641: Computer-Mediated discourse Analysis, January 30, 2017. Indiana University, Bloomington.
- *Gender and Language in CMC: A historical analysis of USENET newsgroups*, L715: Author Profiling, September 15, 2016. Indiana University, Bloomington.
- *Data Modeling*, Z556: Systems Analysis & Design, March 7, 2016. Indiana University, Bloomington.
- *Mining Social Meaning in Computer-Mediated Communication*, Z639: Social Media Mining, February 18, 2016. Indiana University, Bloomington.
- *Features for Sentiment Analysis in Microblogs*. Z543: Computer-Mediated Communication. November 19, 2014. Indiana University, Bloomington.
- *Twitter Hashtags*. S543: Computer-Mediated Communication. April 15, 2013. Indiana University, Bloomington.
- *Discourse Analysis for Social Science Research.* S706: Introduction to Research. October 31, 2012. Indiana University, Bloomington, IN.

TEACHING

(Mixed Undergraduate/Graduate Level)

Z544: Gender and Computerization. Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington Spring 2016, Fall 2016, Fall 2017, Fall 2018.

Z543: Computer-Mediated Communication. Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington Spring 2016, Fall 2017.

(Graduate Level)

Z511: Database Design. Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington Fall 2013 (co-taught with Prof. Ying Ding), Spring 2014, Summer 2014, Fall 2014, Spring 2015, Summer 2015, Fall 2015, Fall 2016.

SERVICE

- Member, Women In Computing (WIC), School of Informatics, Computing, and Engineering, 2017 - present.
- Fellow, Center for Computer-Mediated Communication, Indiana University Bloomington, 2015 – present.
- Grader, North American Computational Linguistics Olympiad (NACLO), 2018.
- Co-organizer, Poster Committee for InWIC 2017, *Indianapolis, October 27-28, 2017.*
- Research Judge, SoIC Projects and Research Symposium 2017, School of Informatics, Computing, and Engineering, *April 20, 2017*.
- Reviewer, *Corpus Linguistics Fest* (CLiF 2016), 2016.
- Chair, ILS Doctoral Student Association, Indiana University Bloomington, November 2014 – 2016.
- Reviewer, *Language@Internet*, 2015.
- Student Representative, School of Informatics, Computing, and Engineering, Luddy Hall Groundbreaking Ceremony, *October 2, 2015*.
- Reviewer, *SAGE Open*, 2015.
- Reviewer, *Pragmatics and Society*, 2015.
- Organizer, Half-Day Doctoral Student Seminar, Indiana University Bloomington, *April 24, 2015*.
- Session Chair, Spring 2015 CCMC Symposia, *April 4th 2015*, Bloomington, IN.
- Co-organizer, 11th ILS Doctoral Research Forum, October 12th 2013, Indiana University Bloomington.
- Co-chair, ILS Doctoral Student Association, Indiana University Bloomington, November 2013 – 2014.

FELLOWSHIPS, AWARDS AND GRANTS

- Scholarship for the Grace Hopper Celebration of Women in Computing Conference 2018, AnitaB.org, 2018.

- Rob Kling Social Informatics Fellowship, Rob Kling Center for Social Informatics, 2017-2018.
- Full sponsorship for the Grace Hopper Celebration of Women in Computing Conference 2017, School of Informatics, Computing, and Engineering, Indiana University Bloomington, 2017.
- Best presentation (2nd place), *The 14th ILS Annual Doctoral Research Forum*, Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington, 2016.
- Dean's Fellowship, Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington, 2012-2016.
- Doctoral Student Travel Grants, Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington, 2013-2014 / 2015-2016 /2016-2017.
- Best presentation (1st place), *The 11th ILS Annual Doctoral Research Forum*, Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington, 2013.
- Greek National Scholarships Foundation Award for Academic Excellence, 2003.