# A Sentiment Analysis of Men's and Women's Speech in the BNC64

**Yong-hun Lee**
Chungnam National University
99 Daehak-ro, Yuseong-gu
Daejeon 34134, Korea
yleeuiuc@hanmail.net

**Ji-Hye Kim**
Hannam University
70 Hannamro, Daedeok-gu
Daejeon 34430, Korea
jihye_0612@naver.com

## Abstract

This study applied sentiment analysis to the corpus data and analyzes the men's and women's speech. For the analysis, the BNC64 corpus was used and three different types of analyses were employed: dictionary-based analysis, GRU-based analysis, and BERT-based analysis. When the data were analyzed with the dictionary-based analysis, there was no significant difference in the use of sentiment words between men and women. When the data were analyzed with the GRU-based and BERT-based analysis, it was observed that even though men and women used a similar proportion of sentiment words, women used more positive words. The tendency became much clearer in the BERT-based analysis. This study implied that the claims in previous studies were supported by the authentic corpus data and that it showed how the gender differences became clearer as the method developed from the dictionary-based method to the BERT-based analysis.

## 1 Introduction

It is known that men's language use is different from women's, and there have been a lot of studies on the gender differences in language use (Holmes and Meyrehoff, 2003; Baker, 2014). These studies investigated both similarities and differences between men and women.

As deep learning technology develops (Ian et al., 2016), there have been a few trials to apply the technology to examine the gender differences in language use, and one of the frequently-used methods is sentiment analysis. Since the advent of the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019), it is possible to use a pre-trained language model to conduct sentiment analysis, because it belongs to one of the sub-tasks that can be analyzed with the BERT.

This study employed one of the traditional methods in sentiment analysis (i.e., dictionary-based methods) and two deep-learning methods (GRU and BERT). The BNC64 corpus was used in the analysis, which was constructed by extracting the men's and women's speech from the British National Corpus (BNC). The data were analyzed with three different types of sentiment analyses, and their analysis results were compared.

This study has importance in that the claims in previous studies are supported by the authentic corpus data and that it demonstrates how the gender differences became clearer as the analysis method develops from the dictionary-based method to the BERT-based analysis. This study is also important in that it is one of a few studies which investigate gender differences with the BERT.

## 2 Previous Studies

### 2.1 Language and Gender

The study on language and gender started from the early academic approach which was called 'gender

difference paradigm'. Lakoff (1975) introduced a 'male dominance' theory of language use, which claimed that males used language to dominate females. Fisherman (1977) developed the theory further and proposed that women engaged in the so-called 'interactional shitwork', which involved using questions and hedges to force responses from men in order to facilitate conversation. On the other hand, Tannen (1990) based on interactional sociolinguistics and mentioned that the different language use by men and women was originated from 'gender differences', rather than from 'male dominance'. This study had a position that males and females had distinct and separate 'genderlects' which result in 'cross-cultural miscommunications'.

While some scholars claimed that there were clear differences between men and women (Loke, 2011), others mentioned that the differences were not so big (Hofland and Johansson, 1982; Cameron, 2008). The former groups of scholars claimed that the differences in language use by men and women could be attributed to essential biological differences which included chemicals in the brain, different reproductive systems, and body size and musculature. All of these factors could impact on how men and women came to see themselves and were viewed by others. In addition, the differences were possibly related to the ways that society socialized males and females differently and different expectations on appropriate language behaviors. However, Butler (1990) said that gender was just performative. That is, gender was determined by a form of *doing*, not by a form of *being*. She mentioned that people did not speak a certain way because they are males or females but that they used the language such a way to perform a male or female identity according to current social conventions about how males and females should behave.

Since the 1990s, there has been a paradigm shift from the studies which forced all men and women to go into separate categories for comparison to the studies which explored differences within the same gender groups (i.e., differences among men or differences among women). Recent studies also focused on the ways that the *gender* factor interacted with other factors (Eckert and McConnnell-Ginet, 1992).

There have also been increasing tendencies to study the differences from the discourse contexts, rather than from the biological differences between men and women. For example, Foucault (1972) said that the terms like *social convention* and/or *expectations* could be related to the concept of *discourse* and defined them as 'practices which systematically form the objects of which they speak'. Burr (1995) pointed out that discourse could be defined as the production of 'meanings, metaphors, representations, images, stories, statements, and so on which produced a particular version of events'. Gill (1993) mentioned that language had increasingly become important across the social sciences, due to the 'influence of post-structuralist ideas which stressed the thoroughly discursive, textual nature of social life'. Cameron (1998) said that in fact, this 'linguistic' turn was mainly a turn to discourse analysis. Livia and Hall (1997) added that "[…] it is discourse that produces the speaker, and not the other way round, because the performance will be intelligible, only if it 'emerges in the context of binding conventions'".

The claim was also supported that the language differences had to be analyzed with the environmental and discourse contexts rather than biological differences. Schmid (2003) created a set of topics and some linguistic features which could be stereotypical to male or female speeches (such as relationships, work, questions, minimal responses). Then, he examined the sex-tagged spoken data in the British National Corpus (BNC) to see whether there were statistically significant differences for sex. The study observed that the female speakers were statistically more likely to use the words such as *dinner*, *tea*, *lunch*, *eggs*, *wine*, *milk*, and *steak*; although males were more likely to say the words including *pint*, *pizza*, and *beer*. The study pointed out that males and females were using language in stereotypically gendered ways. The study analyzed the differences in their environments and primary concerns. Males were more likely to use a lexicon associated with public affairs, abstract concepts, and sport; while females employed more words referencing clothing, colors, and the home.

On the other hand, Romaine (2003) took a sociolinguistic approach and studied some sociolingustic patterns among social classes, style, and gender differences. She found strong correlations between patterns of social stratification and gender and said that: "One of these sociolinguistic patterns is that women, regardless of other social characteristics such as class, age, etc., tended to use more standard forms than men."

## 2.2 Sentiment Analysis

Sentiment analysis is a study which systematically identifies, extracts, quantifies, and studies affective states and subjective information from the use of natural languages (Liu, 2020). It is also known as opinion mining or emotion AI, and it is related to various areas including computational linguistics, natural language processing (NLP), text analysis, and biometrics.

There are roughly four different types of methods in sentiment analysis: dictionary/lexicon-based, corpus-based, machine-learning-based, and deep-learning-based. In the dictionary/lexicon-based analysis, there are some dictionaries where both *positive* and *negative* words are stored. The analysis was conducted based on the frequency of *positive* and *negative* words in the given texts. The corpus-based analysis utilizes a large size of corpora where both *positive* and *negative* annotations are included. In the machine-learning approach, various kinds of machine learning techniques are employed such as Semantic Latent Analysis (SLA), Support Vector Machine (SVM), and so on. In the deep-learning approach, more advanced learning techniques are adopted including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long-Short Tem Memory (LSTM), Gated Recurrent Unit (GRU), and so on.

There were a few studies where sentiment analysis was employed in gender studies. Keith (2017) used a corpus of Wikipedia film summaries to build a word embedding model using the `wordVectors` library in R. The study found that language about women tends to be consistently positive but that words about men skew slightly toward negative. In Park and Woo (2019), the corpus of health-related web forums was used to build a gender detection model. The study observed that senti-words give better results with SVM and that the deep learning algorithm overcame the drawbacks of previous models. In Sun et al. (2020), they used the text data from the Python Technology Community, and they analyzed them with the Latent Dirichlet Allocation (LDA), sentiment analysis, and regression analysis. This study revealed (i) that male and female users mostly expressed *positive* emotions, but female users expressed *positive* emotions more frequently and that (ii) different emotional tendencies of male and female users under different topics had different effects on their activity in the community.

# 3 Research Method

## 3.1 The BNC64 Corpus

The corpus which was used in this study was the BNC64 corpus (Brezina, 2013; 1.5 million word tokens). The BNC64 is a socially-balanced corpus of informal British speech, which was constructed by extracting the texts from the *demogrpaphic* section of the British National Corpus (BNC).

The basic information on the BNC64 corpus is as follows.

| Group | File | Token | Type |
|---|---|---|---|
| Male | 32 | 642,942 | 15,453 |
| Female | 32 | 967,571 | 16,423 |

**Table 1:** BNC64 Corpus

As you can observe in this table, even though the male and female group contained the same number of files, the number of total word tokens in the female group was 1.5 times as many as that in the male group. On the other hand, the number of total word types in the female group was slightly more than that in the male group.

The reasons why the BNC64 corpus was chosen were (i) that the data were extracted from the *representative* and *balanced* corpora (i.e., the BNC corpus) and (ii) that the gender information could be identified clearly. Even though previous studies (Keith, 2017; Park and Woo, 2019; Sun et al., 2020) also used the corpus data, their corpora were hard to be said that they were *representative* and are carefully designed. Their corpora were *specific* corpora rather than *general* corpora. On the other hand, the BNC was a *representative*, *balanced*, and *general*(*-purpose*) corpus. It was carefully designed not only in genre but also in age, gender, socio-economic classes, and so on. Because the BNC64 corpus was constructed by taking the samples from such a *representative* and *balanced* corpus, it was possible to examine the gender differences without any bias in sociological factors. Even though it is also useful to investigate the gender differences in the large scale domain-specific corpora, it is also important to study the gender differences in the *representative*, *balanced*, and *general* corpus. In addition, if the findings of previous studies were also observed in the BNC64 corpus, the study will be another piece of evidence that supports the claims of previous studies.

### 3.2 Dictionary-based Sentiment Analysis

This study started from the dictionary-based (or lexicon-based) sentiment analysis, and the results became the baseline of the comparisons for the other types of analyses. In this paper, the sentiment dictionary in Hu and Liu (2004) was used, which contained about 6,800 words for *positive* and *negative* sentiment. All the words which were not included are classified to be *neutral*.

Three sorts of statistics were extracted from this dictionary-based sentiment analysis: the sentiment score (SS), the ratio of sentiment words (SR), and the ratio of positive words among the sentiment words (PR). They were calculated as follows.

$$\text{Sentiment score} = \frac{\text{\# of positive - \# of negative}}{\text{\# of positive + \# of negative}}$$

**Figure 1:** Calculation of SS

$$\text{Sentiment ratio} = \frac{\text{\# of positive + \# of negative}}{\text{total \# of word tokens}}$$

**Figure 2:** Calculation of SR

$$\text{Positive ratio} = \frac{\text{\# of positive}}{\text{\# of positive + \# of negative}}$$

**Figure 3:** Calculation of PR

Here, SR and PR are converted into the percentage, and they range from 0% to 100%. These three statistics represent different aspects of the given text.

First of all, SS ranges from -1 to 1, and the sign of SS indicates the overall sentiment tendency of the text. If the sign is +, it indicates that the text contains more *positive* words. If the sign is -, on the other hand, it indicates that the text contains more *negative* words.

SR indicates how many sentiment words are included in the given text. Because all the words in the text can be classified into three groups (*positive*, *neutral*, and *negative*) and most of the words are *neutral*, SR tells us how much the speaker use the sentiment words in the given text.

PR indicates how many positive words are included (among the sentiment words) in the given text. Though two different authors may use the same percentage of sentiment words, one may use more *positive* words and the other may employ more *negative* words. Accordingly, PR tells us the speaker's overall sentiment tendency (*positive* or *negative*) in the given text.

Among these statistics, SS and SR are calculated just once based on the sentiment dictionary, and PR will be updated in the GRU-based and BERT-based analysis.

### 3.3 GRU-based Sentiment Analysis

The second type of analysis that this paper took was a deep-learning-based analysis.[1] Among many different deep-learning architectures, this paper took the GRU method (Cho et al., 2014), because the GRU was one of the advanced methods of RNN which could be used for sequential data such as natural languages.

The IMBD movie review dataset (Maas et al., 2011) was used for the training of GRU. IMDB (an acronym for Internet Movie Database) was an online database of movies, television programs, home videos, video games, and streaming content online. The database not only included the information on cast, production crew, personal biographies, plot summaries, trivia, ratings, and fan but also critical reviews on the movies (either *positive* or *negative*). IMDB database contained approximately 7.5 million titles, 10.4 million personalities, and 83 million registered users.

A GRU model was constructed, and the model was trained and tested with the IMDB dataset. It was found that the accuracy was over 90% with 10% of the test dataset. After the GRU model was trained with the IMDB dataset, the sentiment of all the texts in the BNC64 corpus was analyzed with the GRU model. Then, the PR statistics were extracted and the values were plotted against the SR statistics.

---

[1] The following sentences demonstrate the problem of the dictionary/lexicon-based sentiment analysis.

(i) a. To violate a law is not good.
b. Not to violate a law is good.

Because both (ia) and (ib) contain the identical words (i.e., the same seven words), all of the statistics in Section 3.2 will be identical. Notwithstanding, we can say that (ib) is slightly more *positive* than (ia). Traditional dictionary/lexicon-based sentiment analysis cannot capture these differentces, and that is why other kinds of analyses are necessary, including machine-learning-based or deep-learning-based analysis.

### 3.4 BERT-based Sentiment Analysis

After the Transformer-based model was introduced in the deep-learning architecture (Vaswani et al., 2017), there were a few deep learning models which were pre-trained with a huge amount of data. Such models included ELMo (Embeddings from Language Model; Peters et al., 2018), GPT (Generative Pre-trained Transformer; Radford, et al., 2018), and BERT (Devlin et al., 2019).

According to Devlin et al. (2019), the original English BERT had two models: (i) the $\text{BERT}_{\text{BASE}}$: 12 Encoders with 12 bidirectional self-attention heads, and (ii) the $\text{BERT}_{\text{LARGE}}$: 24 Encoders with 16 bidirectional self-attention heads. Both models were pre-trained from unlabeled data extracted from the BooksCorpus (Zhu et al., 2015) with 800M words and English Wikipedia with 2,500M words (Annamoradnejad and Zoghi, 2020).

Among these two models, the $\text{BERT}_{\text{LARGE}}$ was taken. A $\text{BERT}_{\text{LARGE}}$ model was constructed, and the model was trained and tested with the IMDB dataset. It was found that the accuracy was over 90% with 10% of the test dataset. After the $\text{BERT}_{\text{LARGE}}$ model was trained with the IMDB dataset, the sentiment of all the texts in the BNC64 corpus was analyzed with the $\text{BERT}_{\text{LARGE}}$ model. Then, the PR statistics were extracted and the values were plotted against the SR statistics.

## 4 Analysis Results

### 4.1 Sentiment Score

The following bar plot shows us the distributions of SS all over the 64 files (32 files for males and 32 files for females)

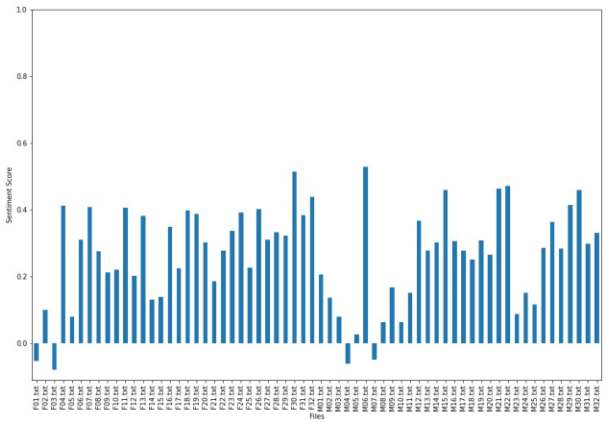


**Figure 4:** Distribution of SS

It was observed that most of the files had a positive SS score, which implied that most of the files contained more positive words than negative words. Only 4 files (2 for female and 2 for male) showed the opposite tendency.

To examine whether there were significant differences between the male group and the female group, a Mann-Whitney test was conducted. The result was that there was no significant difference between the two groups ($W$=591.5, $p$=0.289).[2]

### 4.2 Dictionary-based Sentiment Analysis

The following scatter plot shows the distributions of SR ($x$-axis) and PR ($y$-axis) of the dictionary-based analysis.

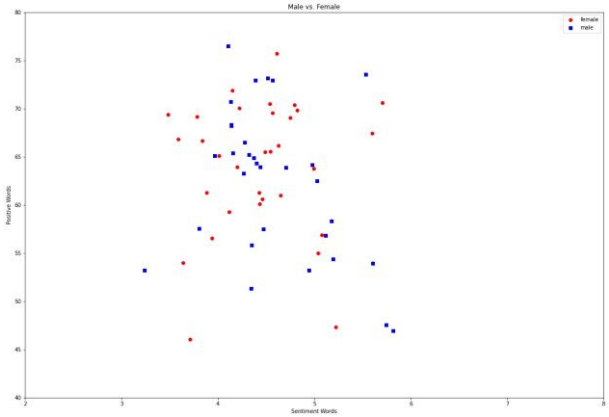


**Figure 5:** Distribution of SR and PR

It was observed that SR ranges from 3% to 6% and that PR ranges from 40% to 80%. It was also found that the distributions of the values in the male groups were mixed with those of the female groups.

To examine if there are significant differences in SR between the male group and the female group, an independent sample $t$-test was conducted, since both groups follow the normal distribution. The result was that there is no significant difference between the two groups ($t$(62)=-1.153, $p$=.253).

[2] When the normality tests (Shapiro-Wilks tests) were conducted, the female group didn't follow the normal distribution ($p$=0.043), while the male group followed the normal distribution ($p$=0.499). That's why a Mann-Whitney test (the non-parametric version of independent sample $t$-test) was conducted. The independent sample $t$-test was also conducted, but there was no significant difference between the two groups ($t$(62)=0.913, $p$=.365).

Likewise, to examine if there were significant differences in PR between the male group and the female group, a Mann-Whitney test was conducted. The result was that there was no significant difference between the two groups ($W$=591.0, $p$=0.294).[3]

### 4.3 GRU-based Sentiment Analysis

The following scatter plot shows the distributions of SR ($x$-axis) and PR ($y$-axis) of the GRU-based analysis.

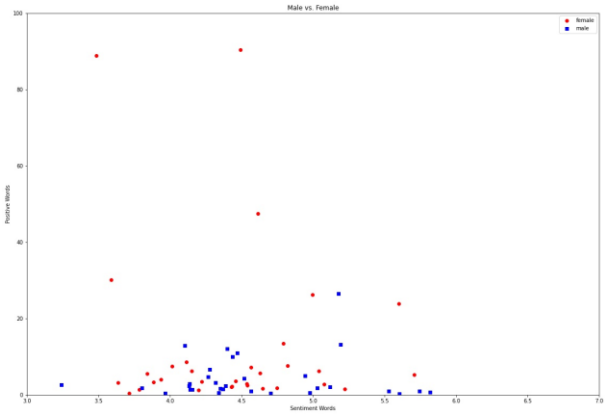


**Figure 6:** GRU-based Sentiment Analysis

It was observed that PR ranges from 0% to 100%, which was different from the PR values of the dictionary-based analysis.

In order to examine if there were significant differences in PR between the male group and the female group, a Mann-Whitney test was conducted. The result was that there was a significant difference between the two groups ($W$=741.0, $p$=0.002).[4] The median of the female group was 59.0 and that of the male group was 1.664. Thus, we can say that the female group speakers say more positively than the male group speakers.

### 4.4 BERT-based Sentiment Analysis

The following scatter plot shows the distributions of SR ($x$-axis) and PR ($y$-axis) of the BERT-based analysis.

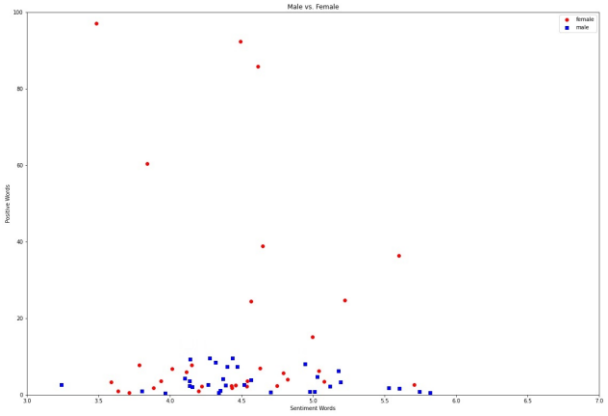


**Figure 7:** BERT-based Sentiment Analysis

It was also observed that PR ranges from 0% to 100%, which was similar to the PR values of the GRU-based analysis.

In order to examine if there were significant differences in PR between the male group and the female group, a Mann-Whitney test was conducted. The result was that there was a significant difference between the two groups ($W$=790.0, $p$<0.001).[5] The median of the female group was 59.069 and that of the male group was 1.661. Thus, we can say that the female group speakers say more positively than the male group speakers.

## 5 Discussion

In this paper, sentiment analysis was used to examine the differences between men and women. This study demonstrated how men and women used sentiment words differently.

In the dictionary-based analysis, the sentiment words in Hu and Liu (2004) were used to calculate SS, SR, and PR. In SS, it was found that most speakers, whether they are male or female, use more positive words than negative words. Though there may be a few speakers which used more negative words, they could be thought an individual preference that did not affect the overall tendency. The finding that there was no significant difference

[3] When the normality tests were conducted, the female group didn't follow the normal distribution ($p$=0.042), while the male group followed the normal distribution ($p$=0.495). The independent sample $t$-test was also conducted, but there was no significant difference between the two groups ($t$(62)=0.913, $p$=.365).

[4] Because there were many ourtliers in the PR values ($y$-axis), only non-parametric test was conducted (Gries, 2013).

[5] Likewise, because there were many ourtliers in the PR values ($y$-axis), only non-parametric test was conducted. (Gries, 2013).

between the two groups implied that both groups of speakers showed a similar tendency.

It was also found that the percentage of SR values ranged from 3% to 6%. It implied that most of the words in the human speech were sentimentally *neutral*. The finding that there was no significant difference between the male group and the female group in SR implied that both groups of speakers used a similar percentage of sentiment words.

In the dictionary-based sentiment analysis, it was observed that PR ranges from 40% to 80%. It implied that people preferred to use more *positive* words among the sentiment words, and the tendency was also observed in the analysis of SS.

In the GRU-based sentiment analysis, it was found that the range of PR values extended to the extreme, i.e., nearly from 0% up to (close to) 90%. It implied that there were many sentences which had to be reanalyzed with the deep-learning method and that people used to express their sentiment not directly but metaphorically or euphemistically.

In the GRU-based sentiment analysis, it was observed that there was a significant difference between the two groups and that the female group speakers said more positively than the male group speakers. It implied that, even though both groups of speakers used a similar ratio of sentiment words, the female group used more *positive* words than the male group, which indicated that women had a more *positive* stance toward the world. These findings were accordant with the observations of Keith (2018) and Sun et al. (2020), and it implied that their observations could be generalized to the populations although their studies were conducted with domain-specific corpora.

Technically, the differences between the analysis results of dictionary-based methods and those of the GRU-based methods demonstrated the shortcomings of dictionary-based methods and why more advanced machine-learning-based or deep-learning-based approaches are necessary for sentiment analysis, as Park & Woo (2019) pointed out. In addition, the clear differences between the two groups of speakers indicated that sentiment analysis could be used when we classified the speech of the male group from the female group, as in Keith (2017) and Park & Woo (2019).

In the BERT-based sentiment analysis, it was observed that there was a significant difference between the two groups and that the tendency was intensified more, compared with the analysis results of the GRU-based methods. It could be said that the BERT-based method revealed the gender differences more clearly.

However, more studies are necessary to examine why the two groups of speakers use the sentiment words differently. The difference can be originated from biological differences (Butler, 1990; Lakoff, 1995; Fisherman, 1997; Tannen, 1990) or from discourse contexts (Foucault, 1972; Burr, 1995; Gill, 1993; Livia and Hall, 1997; Cameron, 1998). There is also a possibility that *gender* interacts with other linguistic factors (Romaine, 2003; Schmid, 2003; Eckert and McConnnell-Ginet, 1992).

## 6 Conclusion

This paper examined the gender differences with sentiment analysis. Three different types of analysis methods were adopted: dictionary-based, GRU-based, and BERT-based analysis method.

In the dictionary-based analysis, it was observed that there was no significant difference between the male group and the female group and that both groups of speakers showed a similar tendency. In the GRU-based and BERT-based analysis, it was found that there was a significant difference between the male group and the female group. It was also observed that the tendency was intensified in the BERT-based analysis.

The analysis in this study demonstrated that why more advanced machine-learning-based or deep-learning-based approaches are necessary and that sentiment analysis can effectively be used for gender classification.

## References


Issa Annamoradnejad and Gohar Zoghi. 2020. ColBERT: Using BERT Sentence Embedding for Humor Detection. arXiv preprint arXiv:2004.12765.

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

Vaclav Brezina. 2013. BNC64. http://corpora.lancs.ac.uk/bnc64

Vivien Burr. 1995. An Introduction to Social Constructionism. Routledge, London.

Judith Butler. 1990. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.

Deborah Cameron. 1995. Gender, Language and Discourse: A Review Essay. Signs: Journal of Women in Culture and Society, 23(4):945-73.

Deborah Cameron. 2007. The Myth of Mars and Venus. Oxford University Press, Oxford.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Minneapolis, MN), 1:4171-4186.

Penelope Eckert and Sally McConnell-Ginet. 1992. Think Practically and Look Locally: Language and Gender as Community-based Practice. Annual Review of Anthropology, 21:461-90.

Pamela Fishman. 1977. Interactional Shitwork. Heresies, 2: 99-101.

Michel Foucault. 1972. The Archaeology of Knowledge. Tavistock, London.

Rosalind Gill. 1993. Justifying Justice: Broadcasters' Accounts of Inequality in Radio. in Erika Burman and Ian Parker (eds), Discourse Analytic Research, 75-93. Routledge, London.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT Press, Cambridge, MA.

Stefan Gries. 2013. Statistics for Linguistics with R.. Mouton, Berlin.

Knut Hofland and Stig Johansson. 1982. Word Frequencies in British and American English. The Norwegian Computing Centre for the Humanities, Bergen, Norway.

Janet Holmes and Miriam Meyrehoff. 2003. The Handbook of Language and Gender. 2nd Edition. Blackwell, Oxford.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004), 168-177. Seattle, Washington, USA, Aug 22-25, 2004.

Ellyn Keith. 2017. A Sentiment Analysis of Language & Gender Using Word Embedding Models. MA Thesis. City University of New York.

Robin Lakoff. 1975. Language and Woman's Place. Harper and Row, New York.

Bing Liu. 2020. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. 2nd Edition. Cambridge University Press, Cambridge.

Anna Livia and Kira Hall. 1997. Queerly Phrased: Language, Gender and Sexuality. Oxford University Press, Oxford.

John Locke. 2011. Duels and Duets. Why Men and Women Talk So Differently. Cambridge University Press, Cambridge, MA.

Andrew Maas, Raymond Daly, Peter Pham, Dan Huang, Andrew Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011), 142-150.

Sunghee Park and Jiyoung Woo. 2019. Gender Classification Using Sentiment Analysis and Deep Learning in a Health Web Forum. Applied Science 9, Article 1249.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. arXiv preprint arXiv:1802.05365.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Suzanne Romaine. 2003. Variation in Language and Gender. In Holmes, Janet and Miriam Meyrehoff. 2003. The Handbook of Language and Gender. 98-118. Blackwell, Oxford.

Hans Schmid. 2003. Do Men and Women Really Live in Different Cultures? Evidence from the BNC. in A.ndrew Wilson, Paul. Rayson and Tony McEnery (eds), Corpus Linguistics by the Lune. Lódź Studies in Language 8, 185-221. Peter Lang, Frankfurt.

Bing Sun, Hongying Mao, and Chengshun Yin. 2020. Male and Female Users' Differences in Online Technology Community Based on Text Mining. Frontiers in Psychology 11, Article 806.

Deborah Tannen. 1990. You Just Don't Understand: Women and Men in Conversation. Virago, London.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. arXiv preprint arXiv:1506.06724.