6-2020

Genderlects in Social Media

Alina Korovatskaya
The Graduate Center, City University of New York

GENDERLECTS IN SOCIAL MEDIA

by

ALINA KOROVATSKAYA

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2020

Genderlects in Social Media

by

Alina Korovatskaya

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Kyle Gorman Thesis Advisor
Date	Gita Martohardjono Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Genderlects in Social Media

by

Alina Korovatskaya

Advisor: Kyle Gorman

Many studies have found significant differences in ways men and women use language; some argue that these differences occur as a result of culture differences, and others suggest that they are influenced by differences in social status and power between the genders. However, some of the major studies were concluded decades ago and do not reflect changes in gender relations in recent years. In this study, we analyze modern conversations using two social media platforms, Twitter and Reddit, to determine whether substantial differences between men and women's use of language were preserved between the genders.

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis advisor, Dr. Kyle Gorman, for his patience, endless support, and immense knowledge. I also would like to thank my parents for supporting and encouraging me throughout this journey.

Contents

Contents **vi**

List of Tables **vii**

1 Introduction **1**

2 Data **5**

2.1 Twitter 5

2.2 Reddit 6

3 Methodology **8**

3.1 Part-of-Speech Tagging 8

3.1.1 Twitter 8

3.1.2 Reddit 8

4 Results **9**

4.1 Word Frequencies 9

4.1.1 Adjectives and Adverbs 9

4.1.2 Pronouns 10

4.1.3 Conjunctions 12

4.1.4 Modal Verbs 13

4.2 Word Length 13

5 Discussion **15**

6 Conclusion **17**

Appendix A **18**

Appendix B **19**

References **20**

List of Tables

1 Number of posts and tokens per corpus and gender 7

2 Word frequencies for adjectives and adverbs 9

3 Word frequencies for personal pronouns (all) 10

4 Word frequencies for 1st person pronouns (singular and plural) 11

5 Word frequencies for 2nd person pronouns 11

6 Word frequencies for 3rd person pronouns (singular and plural) 12

7 Word frequencies for conjunctions 12

8 Word frequencies for modal verbs 13

9 Word frequencies for words in three length categories 14

10 Overview of Twitter tags 18

11 Overview of NLTK tags 19

1 Introduction

"Genderlects" is a term that is used to describe "differences in the speaking patterns of men and women" (Maltz and Borker 1982). Multiple studies, such as Robin Lakoff's (1973, 1975) work on language and gender, have been done on men and women's speech patterns among same- and cross-sex interactions (Herk 2012). These studies suggest that there are significant differences between male and female' speech. Maltz and Borker (1982) state in their article "A Cultural Approach to Male-Female Miscommunication" that women ask more questions, use more *you* and *we* pronouns, and use more positive minimal responses, like *uh-uh, yeah* or *mm*. Men, at the same time, are more direct in their speech, more likely to ignore other speaker's comments, and tend to interrupt their conversational partners. Maltz and Borker (1982) also propose that such variations among male and female speech are the result of the cultural differences between men and women because they "learn different ways of speaking". Other scholars, such as Deborah Tannen, argue that language differences occur as a result of the dominance differences between men and women. For instance, "interruption" is considered to be a feature of a male language, and Tannen (1993) states that "men dominate women by interrupting them in conversation". However, the question "Why such difference exists in the first place?" remains, and scholars continue to debate how large this difference is and whether it is influenced by power differences or by cultural differences (Herk 2012).

Early Work Two key early studies are Robin Lakoff's work (1973, 1975) on linguistic differences between male and female speech. In her works, Lakoff identifies linguistic differences between male and female speech, and some of the differences that she points out include vocabulary differences: commendatory adjectives (*adorable, lovely, fantastic*) and precise color terms (*cerise,*

mauve, magenta) are more likely to be used by women than men. Lakoff also argue that women are more likely to use tag questions (*It's hot out today, isn't it?*) and to adhere to the rules of standard English grammar (*Should it rain today, we would cancel the picnic.*). Men, on the other hand, often use non-standard grammatical constructions (*I ain't goin' with 'em.*) and are more likely to provide direct declaration of fact (*I don't give a damn what you think.*).

Some of the most common male-female vocabulary differences described in prior literature include:

1. *Color words*

   Women are more likely to use precise color terms, such as *lavender, azure* or *aquamarine;* men do not use such terms.

2. *Adjectives*

   Women use more adjectives in their daily life, such as *lovely, gorgeous, heavenly*. Women are also more likely to say, "This house is *splendid*"; men would just say, "This house is *nice*".

3. *Adverbs*

   Women tend to use such adverbs as *quite, so, terribly;* men are more likely to use *very, really.*

4. *Swear words and Expletives*

   Unlike men, women usually avoid swear words like *hell* or *fuck,* instead they are more likely to use *oh, dear* or *my god.*

5. *Diminutives*

   Women like to use words that have the meaning of "small", such as *bookie, hanky;* they also show politeness by using such words as *please* and *thanks.*

6. *Pronouns*

   Women show a greater tendency to use pronouns *you* and *we*.

Other studies of gender variations suggest that women use more intensive adverbs, more connectors, and more modal auxiliary verbs that turn statements into questions. Men, on the other hand, use longer words and a greater number of articles (McMillan et al. 1977; Mulac at al. 2001; Mulac and Lundell 1986; Melh and Pennebaker 2003).

Gender classification Because of these "patterns" that men and women display in their conversations, modern researchers are developing tools for automatic gender predictions. These tools utilize specific linguistics "markers" that characterize each group based solely on gender and rely heavily on predefined word classes, such as part-of-speech (Bamman et al. 2012). Argamon et al. (2007) uses 19,320 English blogs to build a predictive model using words with the higher information gain; Rao et al. (2010) uses Twitter posts of 1,000 users to build a predictive model that combines n-grams collected from Twitter with traditional word and phrase classes. Despite obtaining high accuracy, such gender classifications might become problematic because they assign stereotypical labels that are associated with each gender and don't take into account social identity of the speaker. For example, Nguyen et al. (2014) reports that more than 10 percent of the Twitter users of not employ language that crowd associates with their biological sex.

For years the traditional concept of social roles has dominated people, and men were seen as the possessors of power and status (Erickson et al. 1978). Different roles in society resulted in the variation of men and women's language. But with the "era of feminism" that began in the late 1960s, the women's movement towards equality and with the development of education level, linguistics differences between genders are bound to decrease and linguistics similarities between men and women's language are bound to increase (Flotow 2004). In our modern society men and women have more equal rights; women are no longer associated with "stay-at-home" roles

anymore. Kreider and Elliott (2010) report that the proportion of stay-at-home mothers declined significantly from 44 percent in 1969 to 26 percent in 2009. As Flotow (2004) writes, “The women’s movement of the late 1960s and early 1970s tried to show how women’s difference from men was in many ways due to the artificial behavioral stereotypes that come with gender conditioning”. Insofar as use of language is connected with social systems, when those systems change, one may expect language to change as well. In addition to changes in social systems, there are also new ways for people to interact, like social media. In the age of computers and Internet, people are able to interact with one another not only in person or via phone, but by sending e-mails or posting messages on online forums. As one’s language is constructed by personal interactions and the culture that surrounds them (Lakoff 1973, 1975), the Internet, an important new medium of human communication, is bound to have important long-term effects on language use. By analyzing two social media platforms, Twitter and Reddit, we aim to answer whether linguistics differences between male and female speech, established nearly 50 years ago, still hold true in our modern society.

2 Data

Data for this experiment is manually collected from two widely used social media sources: Twitter and Reddit. Both data sets are manually annotated and then analyzed according to the criteria described in the previous section. Our task is not a gender-prediction task; therefore, we select posts whose authors provide self-identification. Larson (2017) suggested that "participant self-identification should be the gold standard for ascribing gender categories" to avoid binary classification. However, we acknowledge that some people might identify themselves differently from their biological sex or to have different conception of gender, its meaning, or its relation to sex. We are also aware that users of these two platforms may report a gender that is different than their preferred identity, for a variety of reasons.

2.1 Twitter

Twitter is a microblogging and social networking platform on which users can post messages known as "tweets" and interact with other people, known as *followers*. Twitter data is publicly available for all users of the Internet, unless the owner of the accounts decides to make their tweets "private" or available only to their followers. However, many users do not indicate their gender in their profiles. And judging author's gender based on their name is problematic because it can lead to miscategorizations. Even if user states their gender, there is a possibility that gender that the user indicates is different from their biological sex. To avoid these issues we used a data set that consists of tweets by Hilary Clinton and Donald Trump, the two major-party presidential nominees in the 2016 US Presidential Election, available on Kaggle.[1] This data set provides 6,434 unique tweets: 3,167 tweets by Hilary Clinton and 3,167 tweets by Donald Trump, that were created

[1] https://www.kaggle.com/benhamner/clinton-trump-tweets

between January and September 2016. One limitation to selecting this particular data set is that posts by public figures may not be written by said public figures but may be drafted by others at times. Another important thing to note is that by studying public figures there are minimal concerns about privacy.

Preprocessing Each entry in the data set provides a wide range of information, including: ID, handle (*@realDonalTrump*), text, time, whether the post is a retweet or not, and other information that can be used for a quantitative analysis. For the purpose of our study we are interested in only two items: 1) handler or author; and 2) text or tweet. Other elements of metadata were not analyzed. Hilary Clinton's posts are labeled "female" and Donald Trump's posts are labeled "male". We assign gender labels to these authors because they are well-known public figures; gender-labeling of other authors is not attempted.

2.2 Reddit

Reddit is a social news aggregation, web content rating, and discussion website. Its members can post various content and discuss a variety of topics. Discussions on Reddit are organized into subcategories of interests, called "subreddits". When posting on subreddits, authors normally provide some form of self-identification which might include age and gender. This information is useful because researchers don't need to predict author's gender based on the author's name, like it was pursued by Hovy (2015) and Garimella et al. (2019). For our specific task we decide to focus on a subreddit called "r/relationships", where people post questions and discussions about their relationships, romantic and non-romantic. Archives of Reddit data collected from various months are available on the website Pushshift.io.[2] For our data set we collect posts that were submitted to Reddit in November 2018. Originally we obtain a total of 1912 posts: 956 post by men and 956

[2] https://files.pushshift.io/reddit/submissions/

posts by women, but we control for the total number of tokens by downsampling the more frequent category to achieve balance between the two categories. After downsampling, our data consists of 956 posts by men and 873 posts by women. Unlike Twitter data, Reddit data is mostly anonymous.

Preprocessing In the vast majority of posts to r/relationships subreddit authors indicate their age and gender in the title of their post. Author of *How should I [25M] propose to my girlfriend [23F]?* self-identifies as a 25-years-old male and indicates that his girlfriend is a 23-years-old female. This format is important because we want to avoid the presupposition of gender binary based on author's user name. Using these demographic tokens we split our data set into posts made by male users and posts made by female users.

Statistics of Twitter and Reddit text data, such as number of posts and tokens, are summarized in Table 1.

Corpus	Demographic	Total # of Posts	Tokens	
			Unique	Total
Twitter	Male	3,167	7,295	53,478
	Female	3,167	7,060	54,177
Reddit	Male	956	19,146	547,402
	Female	873	19,769	548,987

Table 1: Number of posts and tokens per corpus and gender.

3 Methodology

3.1 Part-Of-Speech Tagging

3.1.1. Twitter

For part-of-speech tagging for Twitter data we use fast and robust Java-based tokenizer and part-of-speech tagger created by Owoputi et al. (2013). We use this particular tagger as opposed to others because it is made specifically for Twitter data which has some unique characteristics, such as hashtags, abbreviations, and improper word spelling. This is an example of an actual Twitter post: “ikr smh he asked fir yo last name so he can add u on fb lololol”. This tool, however, uses a relatively small tag set: the part-of-speech tagger in the NLTK library consists of 35 tags, and Owoputi et al. tagger consists of only 25 tags. And this tagger does not differentiate between singular and plural forms (“N” – common noun) or between various forms of verbs (“V” – verb, including auxiliaries). A full list of tags and their corresponding part-of-speech labels can be found in Appendix A.

3.1.2 Reddit

For part-of speech tagging for Reddit data we use part-of-speech tagger in the NLTK library. And because Reddit posts often consist of multiple sentences, we also use sentence tokenizer to split paragraphs into sentences, and a word tokenizer to split sentences into words. As mentioned above, NLTK part-of-speech tagger consists of 35 tags and does differentiate between singular and plural forms (“NN” – singular noun, “NNS” – plural noun), as well as between different forms of verbs (“VBD” – verb, past tense; “VBZ” – verb, 3rd person sing. present). However, in this study we combine all nouns into one category “noun”, all verbs into one category “verb” etc. to make evaluation of Reddit data consistent with the evaluation of Twitter data. A full list of tags and their corresponding part-of-speech labels can be found in Appendix B.

4 Results

We analyze Twitter and Reddit data sets with respect to some of the most common linguistics patterns associated with each gender: word frequencies for adjectives, adverbs, pronouns, conjunctions and modal verbs; and word length.

4.1 Word Frequencies

4.1.1 Adjectives and Adverbs

Previous studies suggest that women use more adjectives and adverbs than men do (Lakoff 1973, 1975; McMillan et al. 1977; Mulac at al. 2001; Melh and Pennebaker 2003). But our analysis of our Twitter corpus shows that the absolute differences between men and women's use of adjectives and adverbs are quite small. And in contrast to the earlier findings, our results suggest that men use more adjectives and adverbs than women: men's use of adjectives and adverbs is about 10 percent higher than women's. However, in our Reddit corpus, women use adjectives and adverbs more frequently than men, but only by 4 percent for adjectives and 6 percent for adverbs.

Corpus	Demographic	Adjectives		Adverbs	
		Unique	Total	Unique	Total
Twitter	Male	669	**4,407**	**244**	**3,008**
	Female	**765**	3,681	201	2,605
Reddit	Male	3,414	**36,089**	1,000	47,339
	Female	**3,549**	36,031	**1,029**	**48,927**

Table 2: Word frequencies for adjectives and adverbs.

4.1.2 Pronouns

Maltz and Borker (1982) suggest that women show a greater tendency to use the pronouns *you* and *we*, which explicitly acknowledge the existence of the other speaker. Our data shows that use of pronouns are not particularly frequent in our two corpora of social media data. Our corpus of posts by two US politicians on Twitter suggests that women use *we* pronoun more often than men, but men use *you* pronoun more often than women. Our Reddit data show the opposite result: women use *you* more frequently than men, and men use *we* more frequently than women. Another interesting result is that our male Twitter users use first person singular personal pronoun *I* almost 2.5 times more often than our female Twitter users, and first person singular object pronoun *me* almost 4 times more often.

Corpus	Demographic	Pronouns	
		Unique	Total
Twitter	Male	35	**3,785**
	Female	**41**	3,620
Reddit	Male	77	60,168
	Female	77	**60,685**

Table 3: Word frequencies for pronouns (all).

We also observe an interesting, though not unexpected, correlation between uses of third person singular pronouns among our Reddit authors. Our data is collected from a space where users seek advice about interpersonal relationships, romantic or non-romantic. We observe that women use male third person pronouns (*he, him, his*) 5 to 7 times more and men, while men use female third person pronouns (*she, her, hers*) 5 to 9 times more often than women. Although it is evident

that the majority of Reddit authors in our corpus solicit advice about interpersonal relationships with the participants of the opposite gender, we acknowledge that this is not always the case. Different corpora might display alternative results because people seek advice about relationships with the participants of the same gender as well.

Corpus	Demographic	1st Person Singular			1st Person Plural		
		I	*Me*	*Mine*	*We*	*Us*	*Ours*
Twitter	Male	**835**	**288**	2	288	64	0
	Female	339	74	**3**	**688**	**173**	**5**
Reddit	Male	23,258	6,440	**103**	**5,680**	624	9
	Female	**24,758**	**6,930**	82	5,119	**631**	**13**

Table 4: Word frequencies for 1st person pronouns (singular and plural).

Corpus	Demographic	2nd Person	
		You	*Yours*
Twitter	Male	**785**	1
	Female	502	**4**
Reddit	Male	**1,135**	4
	Female	1,102	4

Table 5: Word frequencies for 2nd person pronouns.[3]

[3] 2nd person pronouns have identical singular and plural forms; we do not differentiate between 2nd person singular and plural pronouns in this study

Corpus	Demographic	3rd Person Singular						3rd Person Plural		
		He	*She*	*Him*	*Her*	*His*	*Hers*	*They*	*Them*	*Their*
Twitter	Male	**258**	**146**	**76**	**146**	125	0	**168**	63	69
	Female	229	113	70	143	**257**	**2**	118	**95**	**127**
Reddit	Male	1,500	**12,959**	996	**10,491**	443	**78**	**1,000**	622	**281**
	Female	**13,537**	1,785	**5,916**	1,817	**3,563**	15	958	**742**	272

Table 6: Word frequencies for 3rd person pronouns (singular and plural).

4.1.3 Conjunctions

Scholars suggest that women use more conjunctions then men (McMillan et al. 1977; Mulac at al. 2001; Melh and Pennebaker 2003). Our data shows that on Reddit female users do use more conjunctions, but on Twitter conjunctions are used more often by male users. These results, however, are based on a limited amount of data and might change with an additional data.

Corpus	Demographic	Conjunctions	
		Unique	Total
Twitter	Male	9	**1,376**
	Female	9	1,324
Reddit	Male	41	22,946
	Female	**44**	**23,925**

Table 7: Word frequencies for conjunctions.

4.1.4 Modal Verbs

Numerous studies suggest that women use more modal auxiliary verbs, such as *could, should,* or *may* (McMillan et al. 1977; Mulac at al. 2001; Melh and Pennebaker 2003). Our Twitter corpus shows that women use almost all of the modal verbs more frequently than men. Modal verb *must* is the only exception. One can suggest that *must* is associated with a statement of an obligation or an order which are more likely to be expressed by men, not women (Erickson et al. 1978). Our Reddit corpus, on the other hand, does not display any patterns in how modal verbs are used by the genders: *can, could* and *might* are used more frequently by women, and *may, should, must* are more frequent by men. Although our corpora do not show consistencies in men and women's use of modal verbs, adding more text data might reveal patterns that are not evident with a smaller set of data.

Corpus	Demographic	*Can*	*Could*	*May*	*Might*	*Should*	*Must*
Twitter	Male	95	22	11	2	70	**41**
	Female	**224**	**49**	**17**	**10**	**130**	23
Reddit	Male	**1,123**	**715**	**185**	183	**901**	**35**
	Female	1,171	666	145	183	805	29

Table 8: Word frequencies for modal verbs.

4.2 Word Length

Men have been reported to use longer words (Mulac at al. 2001; Melh and Pennebaker 2003). To test this theory, we split words into three categories based on their length: a) words that have up to

4 characters; b) words that have between 5 and 9 characters; c) words that are 10 characters or longer. To compute these calculations we also remove any punctuation marks, URLs, hashtags, and user names because those are not actual words and could skew results drastically. Hashtags, for example, are often made up of multiple words not separated by spaces, and we want to avoid counting *#HappyMonday* as a 12-letter word. According to our data, women use longer words on both social media websites, which contradicts findings in previous studies.

Corpus	Demographic	Number of characters		
		Up to 4	5 to 9	10+
Twitter	Male	**1,226**	3,726	721
	Female	1,200	**3,947**	**864**
Reddit	Male	3,747	12,019	**3,252**
	Female	**4,035**	**12,373**	**3,214**

Table 9: Word frequencies for words in three length categories.

5 Discussion

After analyzing 6,434 posts collected from Twitter and 1,829 posts collected from Reddit, we can conclude that some of our findings mirror prior research claims, but many contradict them. For instance, we discover that on Twitter men use more adjectives, more adverbs, more conjunctions, and shorter words. These results do not align with Lakoff's findings that claim that those are the features of women's language. At the same time, some analyses of Reddit data confirm results of the earlier works. We hypothesize that men use modal verb *must* more than women because it associates with social power and status, and that women use more modal verbs like *can, should,* and *could* to offer suggestions rather than giving an order (Lakoff, 1973, 1975). Some of our findings about the use of pronouns are also consistent with prior research (Maltz and Borker 1982). We also theorize that women use personal pronoun *we* to a greater extent than men to show inclusiveness, that the conversation is not only about them, but about their conversational partner as well, whereas men display a greater use of personal pronoun *I.* But the claim that women generally use more pronouns (Rao et al. 2010; Bamman et al. 2014) is not confirmed by our study: total number of pronouns used by men is 3,785 and by women is 3,620.

It is also important to note that numbers, obtained from two social media websites, display differences on several measures. One possible explanation of this phenomenon is that our Twitter corpus consists of posts by just two public figures, whereas our Reddit corpus includes posts by many users. Another potential reason that can justify these differences is that Twitter and Reddit are very different platforms, although both are used for online communication and interaction with other people. First, Twitter allows its users to post messages that are limited to 280 characters. Often, to fit their message within these limitations, users use very informal language and a lot of abbreviations (Owoputi et al. 2013). Reddit does not have such limitation; therefore, users are free

to “properly” state their thoughts. Second, users might interact with different social circles on Twitter and Reddit, and, depending on the context, they might accommodate their own styles to those of other people (Nguyen et al. 2014). This study does not examine language variations *within* subjects, but it would be interesting to analyze whether or not people employ different linguistics markers depending on the social media platform they are using, which brings us to our last point.

The majority of NLP research focusing on predicting gender has approached this variable as *biological*, rather than *social* and ignores the fact that language use is related to the social identity of speakers, which might be different from their biological identity (Nguyen et al. 2014). Our analysis on language variation between subjects indicates that many users do not perform the stereotypical language associated with their biological gender, which means that these stereotypical language markers are no longer valid and needed to be revised; or researches should instead focus on sociolinguistics factors that can shape people’s use of language.

6 Conclusion

In this study, we demonstrate that on social media platforms men and women do display some of the stereotypical linguistics patterns associated with both genders; however, many opposite patterns are displayed as well. We find that on some social media websites men use as many modal verbs as women do, that women use longer words and fewer adjectives and adverbs, while on others they display opposite patters. These findings suggest that a deeper evaluation of each platform is needed to understand how they influence people's use of language, and that researchers should re-evaluate how gender is described and studied. Larson (2017) suggests avoiding using gender as a variable in NLP research to avoid bias towards gender binary. And Nguyen et al. (2014) proposes that language variations should be analyzed as a social variable to allow for richer analyses not only *between* individuals, but also *within* individuals. We suggest that future research should not overly rely on older work's description of gender differences. It's not that such approaches are invalid, just merely that they may be sensitive to change over time, and also to genre.

Appendix A

Table 10: Overview of Twitter tags.[4]

Tag	Part of Speech
N	Common Noun
O	Pronoun (Personal/WH; not possessive)
^	Proper Noun
S	Nominal + Possessive
Z	Proper Noun + Possessive
V	Verb incl. Copula, Auxiliaries
A	Adjective
R	Adverb
!	Interjection
D	Determiner
P	Pre- or Postposition, or Subordinating Conjunction
&	Coordinating Conjunction
T	Verb particle
X	Existential *there*, Predeterminers
#	Hashtag
@	At-mention
~	Discourse Marker
U	URL or Email Address
E	Emoticon
$	Numeral
,	Punctuation
G	Other Abbreviations, Foreign Words, Possessive Endings, Symbols
L	Nominal + Verbal, Verbal + Nominal
M	Proper Noun + Verbal
Y	`X` + Verbal

[4] https://github.com/brendano/ark-tweet-nlp/blob/master/docs/annot_guidelines.pdf

Appendix B

Table 11: Overview of NLTK tags.[5]

Tag	Part of Speech
CC	Coordinating Conjunction
CD	Cardinal Digit
DT	Determiner
EX	Existential There
FW	Foreign Word
IN	Preposition/Subordinating Conjunction
JJ	Adjective
JJR	Adjective, Comparative
JJS	Adjective, Superlative
LS	List Marker
MD	Modal
NN	Noun
NNS	Noun Plural
NNP	Proper Noun
NNPS	Proper Noun, Plural
PDT	Predeterminer
POS	Possessive Ending
PRP	Personal Pronoun
PRP$	Possessive Pronoun
RB	Adverb
RBR	Adverb, Comparative
RBS	Adverb, Superlative
RP	Particle
TO	To go ‘to’
UH	Interjection
VB	Verb
VBD	Verb, Past Tense
VBG	Verb, Gerund/Present Participle
VBN	Verb, Past Participle
VBP	Verb, Sing. Present
VBZ	Verb, 3rd Person Sing. Present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive Wh-pronoun
WRB	Wh-adverb

[5] https://medium.com/@muddaprince456/categorizing-and-pos-tagging-with-nltk-python-28f2bc9312c3

References

Argamon, S., Koppel, M., Pennebaker, J., and Schler, J. (2007). Mining the blogosphere: Age, gender, and the varieties of self-expression. In *First Monday*, 12(9).

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2012). Gender in Twitter: Styles, Stances, and Social Networks. Presentation at NWAV 41, Indiana University, Bloomington.

Erickson, B., Lind, A.E., Johnson, B.C., and O'Barr, W.M. (1978). Speech Style and Impression Formation in a Court Setting: The Effects of "Powerful" and "Powerless" Speech. *Journal of Experimental Social Psychology,* 14, pages 266-279.

Flotow, L. V. (2004). *Translation and Gender*. Shanghai: Shanghai Foreign Language Education Press.

Garimella, A., Banea, C., Hovy, D., and Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493-3498, Florence, Italy. Association for Computational Linguistics.

Herk, G. V. (2012). Gender and Identity. *What Is Sociolinguistics?*, pages 85-103, Malden, MA: Wiley-Blackwell.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 752-762, Beijing, China. Association for Computational Linguistics.

Kreider, R.M., and Elliott, D.B. (2010). Historical Changes in Stay-at-Home Mothers: 1969 to 2009. Presented at the American Sociological Association 2010 annual meetings, Atlanta, GA

Lakoff, R. T. (1973). Language and Women's Place. *Language in Society*, 1(2), pages 45-80.

Lakoff , R. T. (1975). *Language and Women's Place*. New York: Harper & Row.

Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1-11, Valencia, Spain. Association for Computational Linguistics.

Maltz, D. N. and Borker, R. A. (1982). A Cultural Approach to Male-Female Miscommunication. Gumperz, J. J., editor, *Language and Social Identity*, pages 196-216, Cambridge, UK: Cambridge University Press.

Mehl, M. R. and Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality & Social Psychology*, 84(4), pages 857-870.

McMillan, J. R., Clifton, A. K., McGrath, D., and Gale, W. S. (1977). Women's language: Uncertainty or interpersonal sensitivity and emotionality? *Sex Roles,* 3(6), pages 545-559.

Mulac, A., Bradac, J. J., and Gibbons, P. (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research,* 27(1), pages 121-152.

Mulac, A. and Lundell, T. L. (1986). Linguistic contributors to the gender-linked language effect. *Journal of Language & Social Psychology,* 5(2), pages 81-101.

Nguyen, D-P., Trieschnigg, R. B., Dogruoz, A. S., Gravel, R., Theune, M., Meder, T., and de Jong, F. M. G. (2014). Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014,* pages 1950-1961, Association for Computational Linguistics (ACL).

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pages 380-390, Atlanta. Association for Computational Linguistics.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in Twitter. In *Proceedings of the 2nd international Workshop on Search and Mining User-Generated Contents*, pages 37-44.

Tannen, D. (1993). The Relativity of Linguistic Strategies: Rethinking Power and Solidarity in Gender and Dominance. Tannen, D., editor, *Gender and Conversational Interaction*, pages 165-188, Oxford University Press.