



Cedarville University
DigitalCommons@Cedarville

Linguistics Senior Research Projects

Department of English, Literature, and Modern
Languages


2016

Linguistic Expression and Gender: A Function Word Analysis of Jane Austen's *Pride and Prejudice*

Erica Corbiere

Cedarville University, ecorbiere@cedarville.edu

Follow this and additional works at: http://digitalcommons.cedarville.edu/linguistics_senior_projects

 Part of the [Anthropological Linguistics and Sociolinguistics Commons](#), [Comparative and Historical Linguistics Commons](#), [Computational Linguistics Commons](#), [Literature in English, British Isles Commons](#), and the [Psycholinguistics and Neurolinguistics Commons](#)

Recommended Citation

Corbiere, Erica, "Linguistic Expression and Gender: A Function Word Analysis of Jane Austen's *Pride and Prejudice*" (2016). *Linguistics Senior Research Projects*. 5.
http://digitalcommons.cedarville.edu/linguistics_senior_projects/5

This Capstone Project is brought to you for free and open access by DigitalCommons@Cedarville, a service of the Centennial Library. It has been accepted for inclusion in Linguistics Senior Research Projects by an authorized administrator of DigitalCommons@Cedarville. For more information, please contact digitalcommons@cedarville.edu.



Linguistic Expression and Gender:

A Function Word Analysis of Jane Austen's *Pride and Prejudice*

Erica C. Corbiere

Cedarville University

Abstract

The current study investigates ten dimensions of female and male categories of speech, which focus on function words, as previously identified by Newman et al. (2008). Through the use of the Linguistic Inquiry and Word Count tool (using the LIWC2015 dictionary), these ten categories were analyzed in the dialogue of Jane Austen's *Pride and Prejudice*. Results were consistent with previous findings by Newman et al. (2008). Four of five previously identified categories as more often used by male speakers (numbers, words per sentence, prepositions, articles, and words greater than six letters) were used with an even greater difference between Austen's male and female characters. Four of five previously identified female categories of language (use of more negative emotion words, verbs, certainty words, negations, and personal pronouns) were also consistent with the Newman (2008) study and again revealed greater disparity between male and female usage. Results contribute to the idea that Jane Austen intentionally wrote character dialogue along gender language dimensions and to the argument that gendered language differences have long existed in the English language.

Keywords: gender, language, LIWC, Austen

Linguistic Expression and Gender:

A Function Word Analysis of Jane Austen's *Pride and Prejudice*

It is a truth not universally acknowledged that men and women use language differently. Various studies explain differences between language usage and gender in cultures and sociolinguistic communities around the world. In such situations where women have less mobility and education than men, a clearer disparity in language emerges. Other cultures reveal differences in language due to traditional social hierarchy or vestiges thereof which still operate in a society. In English-speaking Western societies, however, such differences may be seen as restrictive and unfair and are thus controversial. In the English language, then, we find research that both supports and negates the idea of gender language differences. The following study investigates the question of gender language as it applies to the work of Jane Austen: whether a distinction in gender language usage emerges from the dialogue of male and female characters in *Pride and Prejudice*.

Literature Review

Modern sociolinguistic study of gender and language in English began with the work of Robin Lakoff (1973). Her work on lexicon and syntactic differences claimed women's linguistic expression reflected a subjugated role in society. For example, she claimed that tag questions were a mark of deference to another's opinion, namely to that of a man. Thus, she concluded "the personal identity of women...is linguistically submerged" (Lakoff, 1973, p. 45). Her view, then, epitomized the idea that dominance is the large contributing factor to differences between male and female speech.

More recently, the popular view has shifted from dominance to simply differences to currently a more controversial debate on whether differences exist at all. Various studies have focused on syntax and lexical differences. While many studies report gender disparity, other results do not show significant differences (Cameron, 2005; Mulac & Lundell, 1986; Newman et al., 2008).

While such studies have focused on syntax and lexical differences, further research has turned to the role that function words, specifically, play in gender language. Function words are the semantic components of language that do not carry content or meaning, but rather serve a function in forming sentences and arranging meaning. James Pennebaker has delved into what function word usage reveals about various personal dimensions, including gender. This is the chief subject of his book *The Secret Life of Pronouns: What our Words Say about Us*, in which he highlights the significance of function words, including pronouns, prepositions, and articles, as the key to the person behind the words (2011).

Pennebaker argues for the validity of function words. Instead of dismissing them, he first points out that 55% of the average person's informal communication, both written and oral, is comprised of the 200 commonly used function words. Since function words reflect linguistic style rather than content and therefore, argues Pennebaker, can reveal a lot about the author. Through numerous correlative studies executed by both Pennebaker and other researchers, statistically significant data has been used to identify many characteristics of the author, including emotional state, gender, honesty, power dynamics and leadership position, and the closeness of relationship between two authors (Pennebaker, 2013).

Pennebaker and associate Martha E. Francis have developed a psycholinguistic research tool named *Linguistic Inquiry and Word Count*, or LIWC, which he has used for much of his research and conclusions (Pennebaker, 2011). Originally, this tool came from Pennebaker's desire to systematically analyze patients' trauma essays in psychology research for clues into their psychological states (Pennebaker, 2011). Currently, it is a computer program which analyzes texts of all different types (from self-narratives, texts, social media, letters, and other sources) along an identified list of 80 linguistically significant categories, including the function words like prepositions, articles, and pronouns. Using these dimensions to analyze texts, Pennebaker and numerous other researchers have found a plethora of correlative data between psychological traits and states and these categories. Thus the 80 categories are considered psychologically significant (Tausczik & Pennebaker, 2010).

In labeling correlating meaning to these categories, researchers have applied this technology in various beneficial ways. One of LIWC's uses is identification of author gender. Newman, Groom, Handelman, and Pennebaker (2008) have identified statistically significantly recurring word categories that display gender differences in both spoken language and written language of varying genres. On average, men use more "big words" (a term used by Newman et al.), nouns, prepositions, numbers, words per sentence (WPS), and swear words. Women, on the other hand, use more personal pronouns, verbs, negative emotion, negations, certainty words, and hedge phrases (such as "I think" or "I believe"). While this is statistical data, Pennebaker gives possible explanations for this, claiming that men's language is more concretely and analytically focused (as indicated by the presence of more prepositions to show relationship between ideas and nouns to talk about concepts or concrete objects), whereas

women speak more relationally (personal pronouns indicate talking about or to other people) (Pennebaker, 2011; Newman et al., 2008).

Such findings have been applied to authors' portrayal of gender in fiction as formed through function words. In a study of female and male playwrights and screenwriters, researcher Molly Ireland along with Pennebaker found that results varied widely, with female and male lead characters following typical gender patterns, both following male tendencies, or both following female. This varied whether or not the writers were female or male (Pennebaker, 2011).

Additional research has been done on characterization and gender language in fictional works through computerized literature analysis. One well-studied literary realm is the work of Jane Austen. J. F. Burrows' pioneering work *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* analyzes the dialogue of Jane Austen's major characters using thirty key function words including pronouns, prepositions, and common adverbs (1987). Burrows investigates questions ranging from change within characters, similarities between characters, to distinguishing characteristics. His book reports a large amount of data and offers some possible interpretations. For example, Mr. Collins' language is most estranged from any other character's dialogue in *Pride and Prejudice*, which is consistent with Austen's portrayal of him as an eccentric and socially awkward sort of personality. Throughout his intensive study, Burrows demonstrates the insight that his selected function words give toward Austen's choice of language in character development.

While Burrows did not focus on gender language specifically, his results allow for further analysis of Austen's male and female dialogue. His report on the peculiar use of the degree adverb *quite* throughout Austen's novels, for example, sparked the research by González-Díaz

(2014). Whereas Burrows cites various semantic meanings of *quite* in Jane Austen, González-Díaz's corpus-based analysis reveals a trend in usage of the various meanings of the word. She shows that female dialogue uses varied semantic range (including the adverb's role as maximizer, emphasizer, and scalar modifier) whereas the male characters only use one common semantic meaning of *quite* (for the maximizer function). Therefore, when Darcy responds to Elizabeth's rejection, he uses the typical masculine maximizer degree of "quite": "You have said quite enough, madam" (Austen, 1813, Chapter 34). However, female characters throughout Austen's work used more varied functions, as in Mrs. Bennet's usage of the emphasizer function: "I quite detest the man" (Austen, 1813, Chapter 3). González-Díaz argues that, even to the applied usage of particular functions of an adverb, Jane Austen intentionally gendered her characters' language. Literary critics have also claimed that Austen particularly formed language to challenge stereotypes of women's language (Michaelson, 2002, as cited in González-Díaz, 2014, p. 310).

Further, more comprehensive linguistic analysis of Jane Austen's work may reveal whether she wrote gendered language as intentionally as González-Díaz claims. Her work, then, will be the text analyzed in the present study by the LIWC tool in order to compare Newman's (2008) gendered linguistic dimensions with Austen's potentially intended gender writing. The present study analyzes the dialogue of her most prized work *Pride and Prejudice* to further understand these linguistic dimensions and Austen's gendered writing.

Methods

The basis of this study was the Linguistic Inquiry and Word Count (LIWC) computer program. This program analyzes a given text according to a possible total of 92 psychologically-

relevant categories and subcategories, as statistically revealed through previous research. Some of these categories are purely objective, such as the number of articles present in a text. Others are more subjective, such as frequency of negative emotion words, in which judges were used to compile a list of all words deemed relevant to the category. The sum total of all words chosen to be relevant for the 92 categories are placed in the LIWC dictionary. This study used LIWC2015. Upon entering a text, the program analyzes all selected dimensions and measures word frequency (the percentage of total words which represent the given category).

The text selected for analysis was the dialogue of individual characters. The following eleven characters were selected for their status as principal characters and because they have romantic counterparts through which the contrast of masculinity and femininity could be best understood: Caroline Bingley, Charlotte Lucas, Elizabeth Bennet, Jane Bennet, Lydia Bennet, Mrs. Bennet, Mr. Bennet, Mr. Bingley, Mr. Collins, Mr. Darcy, and Mr. Wickham. Thus, five couples were represented as well as Caroline Bingley to give another character in contrast to Elizabeth Bennet. These characters were also selected since they each have over 1,000 words of dialogue (with the exception of Charlotte Lucas, who has 900 words). The more words in a text analyzed by LIWC, the more accurate and significant the yielded data.

Additionally, only an individual character's dialogue was included. Five other types of expression were not included so as to eliminate any confounding variables: letters, internal thoughts, descriptions of dialogue, quoted dialogue, and speech in unison. Any one of these types of expression, although verbal and personal in nature, could reflect either unreliable word usage, inconsistent reflections of each character's expression, or a lack of individual expression.

First, internal thoughts and descriptions of dialogue, used by Austen frequently to replace dialogue and often quite similar to dialogue, could reflect unrepresentative word usage. For example, Austen describes the following encounter between Mr. Darcy, Caroline Bingley, and Elizabeth Bennet:

He was directly invited to join their party, but *he* declined it, observing that *he* could imagine but two motives for *their* choosing to walk up and down the room together, with either of which motives *his* joining *them* would interfere. “What could he mean? *She* was dying to know what could be his meaning?”—and asked Elizabeth whether *she* could at all understand him? [emphasis added] (Austen, 1813, Chapter 11).

Here, “he,” “his,” “their,” and “you” are used instead of “I,” “my,” “your,” and “you” respectively as part of Darcy’s described speech, and “she” is used instead of “I” and “you” in Caroline’s described speech. Since this change of pronoun also affects verbs and possible other word choices, such samples were not analyzed (as verbs and pronouns were both types of Pennebaker’s dimensions that were chosen to be analyzed).

Letters were not analyzed since characters may express themselves differently in this language form (a question not being tested within the confines of this study). Since not every one of the eleven characters writes a letter included in the book nor are the lengths of the letters the same (Collins and Darcy write long letters, yet Bingley never writes any), any letters present were left unanalyzed, in case letters proved to be a significantly different form of verbal expression.

Quoted dialogue and speech in unison lacked individual expression. At one point in the book, Mr. Darcy quotes Elizabeth’s words back to her, saying “Your reproof, so well applied, I

shall never forget: ‘*had you behaved in a more gentlemanlike manner* [emphasis added].’ Those were your words” (Austen, 1813, Chapter 58). Since this phrase is not his personal expression, it too was omitted. Second, Kitty’s and Lydia Bennet’s speech sometimes occur in unison, as in: “‘*Good Heaven! What is to become of us? What are we to do?*’ [emphasis added] would they often exclaim in the bitterness of woe. ‘*How can you be smiling so, Lizzy?*’” [emphasis added] (Austen, 1813, Chapter 41). Here, it is not possible that both characters uttered the same words individually. The expression was omitted from analysis since it could not be accurately attributed to Kitty or Lydia.

With these omissions in mind, all dialogue was collected from the Project Gutenberg online edition of *Pride and Prejudice*. Each character’s dialogue was placed in a separate text document. Since this step was done manually, the entire book’s dialogue was reviewed twice to separate out each character’s individual speech correctly. Once this was done, each character’s finished text document, a collection of his or her entire dialogue, was analyzed once more for any errors, especially to make sure that extra-dialogue expressions such as “...said Elizabeth...” were excluded. Therefore, each speech sample was reviewed three times before being analyzed by the LIWC tool.

LIWC then analyzed each of the 11 texts of dialogue along the 92 categories and subcategories using the LIWC2015 dictionary and computer program. Pennebaker lists the following dimensions as most statistically significant to gender language (more frequently used by either males or females), according to his and fellow researchers’ previous research (Newman et al., 2008; Pennebaker, 2011, p. 40-43): articles, big words, nouns, prepositions, numbers, words per sentence, and swear words for male categories, and personal pronouns, negative

emotion words, cognitive words, social words, verbs, negations, certainty words, and hedge phrases for female categories (see Table 1). Of these dimensions, five were usable for male speech and seven for female speech. “Nouns” and “hedge phrases” did not coincide well with LIWC2015 categories and therefore were not considered. “Swear words” were judged to not be an applicable category to Jane Austen’s writing. In order to have an equal analysis, five of the seven female dimensions were chosen at random. Thus, ten dimensions were used for analysis. According to Newman (2008), five have been statistically used more by men (big words, articles, prepositions, numbers, and words per sentence) and five by women (personal pronouns, verbs, negative emotion words, negations, and certainty words) (see Table 2).

After analyzing the eleven text documents of dialogue with the LIWC program, the data for each dimension was organized by first comparing male and female speech patterns and then by comparing the findings to those of Newman et al. (2008). The comparison between studies was to indicate whether Jane Austen portrayed gender in dialogue similarly according to modern statistics of male and female usage of these ten language dimensions.

To gain insight into popular perception of gender language, a questionnaire was also incorporated into the study. The questionnaire was distributed to members of a university academic community. It sought out general opinion on how the characters’ speech reflects gender language. The two-page form asked various questions to shed light on this issue (See Appendix A). The questionnaire also asked for the participant’s exposure to *Pride and Prejudice* in case a certain movie portrayal versus the book were to skew the audiences’ perceptions of gender speech. Questionnaires were sent out to participants and returned anonymously. A small

sample was returned. Commonalities in answers were then found and compared to the LIWC data on the ten dimensions.

Data and Analysis

Current Study

Previously Male Dimensions. The LIWC2015 analysis of *Pride and Prejudice* dialogue according to the ten aforementioned linguistic categories yielded the following results. First, the five categories which Pennebaker (2011) reports as indicative of masculine dimensions of speech were analyzed (see Figure 1). Averages reported for all data indicate an average percentage of words identified by the LIWC tool. For example, as Figure 1 indicates, articles accounted for 4.72% of male characters' speech, on average. All following data will also indicate average percentage.

Regarding words per sentence (WPS), Jane Austen's writing reveals a higher average for the five male characters selected (17.41) as compared to her female characters (14.04). Additionally, male characters (16.51) used more "big words" (defined by Pennebaker as words with six letters or more) than the female characters (14.03). Additionally, these categories appear correlated, as characters who had high averages for WPS also had high averages for big words and those who scored lower had consistent results in both categories (see Figure 2). For example, Collins scored highest in both categories. On the other hand, female characters such as Mrs. Bennet scored low in both of these dimensions. This could be due, as Pennebaker has stated, to men's more organized and categorical way of thinking, resulting in more words and bigger words.

Articles also reveal a greater use by male characters. The five men use articles on average at 4.72% of total words whereas the women's average is 4.16%. Here, however, Caroline Bingley at 5.63% reveals the most usage of articles:

“A woman must have *a* thorough knowledge of music, singing, drawing, dancing, and *the* modern languages, to deserve *the* word; and besides all this, she must possess *a* certain something in her air and manner of walking, *the* tone of her voice, her address and expressions, or *the* word will be but half-deserved” [emphasis added] (Austen, 1813, Chapter 8).

Pennebaker (2011) claims that such a use of articles reflects concrete language use. According to his analysis, it is typical of men to talk about things and the world, whereas women are more likely to emphasize relationships and speak in a less concrete manner. Here, however, we see Caroline incorporating a concrete, analytic speech. Despite Caroline's dialogue revealing such a high average, Wickham (5.18), Collins (5.16), Mr. Bennet (4.65), and Darcy (4.61) scored the next highest. Thus, the men still statistically reflect this concrete language as evidenced by article usage.

Pennebaker (2011) also claims that preposition usage is responsible for constructing more concrete language typical of male gender speech. Again, the male characters (12.70) in *Pride and Prejudice* had a higher average than the female characters (12.22). Here, Darcy and Collins had the highest averages. From Darcy's first utterance, we see his propensity for concrete language using prepositions:

"I certainly shall not. You know how I detest it, unless I am particularly acquainted *with* my partner. *At* such an assembly *as* this it would be insupportable. Your sisters are

engaged, and there is not another woman *in* the room whom it would not be a punishment *to me to stand up with*" [emphasis added] (Austen, 1813, Chapter 3).

His language, at the country ball, reveals an analysis of the situation, explaining concrete terms and relationships between them.

The last category, unlike the previous four, did not follow previous research. On average, according to Pennebaker (2011), men use more numbers, to go along with their categorical, concrete thinking. However, in *Pride and Prejudice* women used numbers at an average of 1.07, whereas the male characters' average was 0.80. This category revealed controversial data. The numbers were low to begin with, lower than the Newman et al. (2008) reported data as well as those of other categories. The data was especially skewed by the outlier Lydia, as shown in Figure 3. In the book, she often references numbers for time (in months), for the amount of people, and other instances as in "there were two or three much uglier in the shop" (Austen, 1813, Chapter 39), when talking about hats. It may be that use of numbers reflects a focus on easily-countable concrete objects, found throughout Lydia's speech. Without Lydia's data, the disparity in averages lessened but the female speech still was higher than male speech (see Figure 4).

Previously Female Dimensions. Just as four of the five previously male dimensions were consistent with male language use, four of the five previously female dimensions were also consistent with Austen's writing for her female characters (see Figure 5). As previously found, Austen's female characters used more verbs, negative emotions, negation, and certainty than her male characters. However, the male characters used more personal pronouns, which was inconsistent with previous findings.

Use of personal pronouns was the inconsistent dimension. Pennebaker (2011) breaks down personal pronouns into first person singular (I, me, my), first person plural (we, us, our), second person (you, your), and third person pronouns (singular and plural). Drawing from his research in Newman et al. (2008), he states statistically women use first person singular more than men, but men and women use first person plural the same. Men, however, use second person more than women with a small effect. (Each individual category of personal pronouns is shown in Figure 6 for comparison.)

The present data reveals Austen's male characters used more first person singular and second person than her female characters. The second person subcategory was consistent with Newman et al. (2008) findings. However, the greatest disparity was in the first person singular category. Male characters (6.85) on average were much higher than female characters (5.47). The individual characters' averages are shown in Figure 7. Bingley, who on average across the ten dimensions used the least male language, had the highest average (8.59) in this category. The various examples of his first person singular speech included the following quote about his book collection: "And *I* wish *my* collection were larger for your benefit and *my* own credit; but *I* am an idle fellow, and though *I* have not many, *I* have more than *I* ever looked into" [emphasis added] (Austen, 1813, Chapter 8). Though Bingley and Collins are quite different characters, Austen incorporated this same feature into Collin's speech (6.62):

"*My* reasons for marrying are, first, that *I* think it a right thing for every clergyman in easy circumstances (like *myself*) to set the example of matrimony in his parish; secondly, that *I* am convinced that it will add very greatly to *my* happiness; and thirdly—which perhaps *I* ought to have mentioned earlier, that it is the particular advice and

recommendation of the very noble lady whom *I* have the honour of calling patroness”

[emphasis added] (Austen, 1813, Chapter 19).

Even at a moment such as this, his proposal to Elizabeth, Collins has a high usage of personal references, finding a way to explain all of his reasons for marrying Elizabeth without a single reference to her. With such an explicit example as this, it could reflect Austen’s general intentionality in her pronoun usage. While her usage of first person singular pronouns was not consistent, her second person pronouns were. Overall, her average usage of personal pronouns were lower for women, which was inconsistent with previous findings on gender language usage.

The other four dimensions - verbs, negative emotions, negation, and certainty words - were consistently female dimensions for both Austen and in Newman’s (2008) findings. It should be noted that Pennebaker (2011) claims that women have statistically higher usage of verbs and auxiliary verbs. Since this study used the LIWC2007 dictionary whereas the current study used the LIWC2015 study, some categories and subcategories were different. The data reported here, then, only includes the verb category and not the categories of verb and auxiliary verbs. (See Appendix B with raw data comparing studies.) With only the verb category, Austen’s female characters (20.23) scored higher on average than the male characters (17.99). Lydia, Mrs. Bennet, and Jane respectively scored highest of all the characters. (As the following section suggests, they have the most total feminine language usage of all female characters, see Figure 8). This verb usage may reflect a focus on description of events instead of on concrete objects (which male speech statistically reflects). By way of example, we see verb usage (and auxiliary verb usage) in Lydia’s speech, revealing relationships between objects and people:

"And we *mean* to *treat* you all, but you must *lend* us the money, for we have just *spent* ours at the shop out there. Look here, I have *bought* this bonnet. I do not *think* it is very pretty; but I *thought* I might as well *buy* it as not. I shall *pull* it to pieces as soon as I *get* home, and *see* if I can *make* it up any better" [emphasis added] (Austen, 1813, Chapter 39).

Interestingly, the verb category is the only one in which Lydia ranked highest in feminine language use. In the following three dimensions, she scored much lower (see Appendix C).

The last three of the ten categories were negation of verbs, negative emotion, and certainty words and phrases. The female characters scored higher on average than the male characters. These three dimensions were the feminine dimensions for which Elizabeth's speech ranked highest (see Appendix C). In fact, she uses the words "cannot" and "not" 258 times, often preferring to negate what she does not mean instead of directly stating what she does mean. There are many examples of this when she converses with Lady Catherine, as in the following conversation at Rosings:

"Yes, my youngest is *not* sixteen. Perhaps *she* is full young to be much in company. But really, ma'am, I think it would be very hard upon younger sisters, that they should *not* have their share of society and amusement, because the elder may *not* have the means or inclination to marry early. The last-born has as good a right to the pleasures of youth as the first. And to be kept back on *such* a motive! I think it would *not* be very likely to promote sisterly affection or delicacy of mind" [emphasis added] (Austen, 1813, Chapter 29).

This is one of the four negated responses she gives during one conversation with Lady Catherine. In this instance and throughout the book, her speech is riddled with negations. Her sister Jane ranked even higher in this category, further validating this language dimension as female. This may also be related to the high description of negative emotion as a necessary expression of feeling and to the certainty category as ways to make one's point known.

Mrs. Bennet scored the highest in the Certainty category. Generally, this category is more common in female speech in addition to hedge phrases, a combination which allows women to convey a stance but also allow room for the respect of other's opinions, as women are psychologically shown to be more relationally-oriented (Pennebaker, 2011). This is applied to Mrs. Bennet's speech in a more distinct way. She has been doubting Mr. Bennet, frustrated with him for not visiting Bingley. Then upon hearing his declaration that he has just visited Bingley, Mrs. Bennet immediately responds, "How good it was in you, my dear Mr. Bennet! But I *knew* I should persuade you at last. I was *sure* you loved your girls too well to neglect such an acquaintance" [emphasis added] (Austen, 1813, Chapter 2). She is quick to reverse her opinion and convey her longstanding certainty on a new position. Thus, this dimension of language may manifest itself this way in Mrs. Bennet's speech.

In all, the four of the five female dimensions of speech were confirmed in this study: verbs, negation, negative emotion, and certainty. Conversely, the personal pronouns dimension was not consistent with previous research, but instead revealed a complex usage of each category of pronoun (by person and plurality) by the two genders.

Trends in Individual Character Findings

Having seen the above examples and how individual characters' dialogues incorporated the ten dimensions of this study, the following explains the overall masculine/feminine gender language usage for each character (see Figure 8). Ranking characters in each dimension from 1 (most female) to 11 (most male), and then averaging the ten rankings together, the characters appear in the following order from most feminine language to most masculine: Jane (smallest bubble in Figure 8), Mrs. Bennet, Bingley, Elizabeth, Lydia, Mr. Bennet, Darcy, Caroline, Wickham, Charlotte, and Collins. Male and female characters, therefore, are not grouped distinctly from one another when examined on an individual level. Rather, Bingley, for example, ranked as more feminine in language usage than four of the six female characters. There is, therefore, individual variation along the dimensions.

While the averages of Austen's male and female characters were consistent for four of the five masculine dimensions and four out of the five feminine dimensions, each individual character still shows a variety of usage. For example, Collins had the most masculine language tendencies in WPS, big words, prepositions, negations, and verbs. However, for certainty and negative emotions he ranked in the middle. For the numbers dimension, he even ranked most feminine. Therefore, his general tendency was more masculine language use, yet he, like all of the other characters, reveal a complex language usage that did not often follow a predictable pattern. (See Appendix C for all rankings and raw data.)

As an additional example, Lydia proved to be the least consistent in ranking, and for this reason places close to the overall median. Lydia used female language in the WPS, big words, prepositions, and verb dimensions yet also ranked masculine in negative emotions, certainty, and

numbers. Possibly more significant is that in the categories where Lydia scored most feminine, Collins scored most masculine, and for the dimensions where Lydia scored most masculine, Collins conversely scored most feminine. Such a contrast may reflect Austen's intentional character portrayal, as Collins and Lydia are opposite characters in the story. The one is defined by her impropriety, lack of respect, and flirtatiousness, and the other by his strict moral code, undying allegiance to others (namely his esteemed patroness, Lady Catherine), and most awkward social interactions. This may not be a matter of gender language *per se*, yet Austen may have used these language patterns to point out inconsistencies in expression between these contrasting characters of opposite gender.

That being said, more masculine or more feminine traits in language usage do not predict character similarity or likeability either. Surprisingly, the closest in average language usage to Jane is Mrs. Bennet. Their characters, personalities, and self-expression contrast greatly. Yet Austen portrays them both as the most feminine of characters along the ten dimensions. Interestingly, the female characters pair up according to feminine language: Jane and Mrs. Bennet having most feminine language, Lydia and Elizabeth in the middle, and Caroline and Charlotte with the least feminine language. Thus, Austen does not use more feminine language in female characters to portray certain personality or character traits. Rather, she may have paired female characters according to linguistic expression to show both the negative and positive portrayal of that expression. That is, if Jane and Mrs. Bennet speak the most similarly, then more feminine language can come from the polite and demur Jane or the bold and domineering Mrs. Bennet just the same. Therefore, female characters are not defined by their feminine language use. Rather, a respected female in Austen's contemporary society could

express herself along the spectrum of female language. Conversely, those more controversial characters (Mrs. Bennet, Lydia, and Caroline), could also use similar language but with opposite effect. As is typical with Austen's work, it shows that women's roles need not be so confined, nor in this case, their language.

By way of contrast, male characters revealed a more one-dimensional, predictable depiction according to language use. The most masculine of the male characters are the most controversial in popular reception of the books and in Austen's portrayal of them. Bingley and Mr. Bennet are typically considered the warmer, more reliable and likable characters. They, interestingly, are the most feminine in their language use. The more controversial male characters, namely Collins and Wickham, scored the most masculine. Mr. Darcy, though controversial in his own right, undergoes the most character development, starting as a colder personality and ending as a warmer, more likeable character. It may be for these reasons that he falls in the middle of the spectrum of male characters. (Future research study examining his speech chronologically may better indicate whether his speech becomes more feminine in nature as his character develops.) It could suggest that Austen intentionally portrays more of the male characters in a more predictable manner (more likeable meaning more feminine usage), and yet makes the females the true center of her dialogue and story and therefore more complex in their portrayal. (The female characters have twice the average dialogue as the male characters: see Figure 13.)

While the characters are individually complex, when paired together another interesting trend emerges. When the female characters are ranked by most feminine to least feminine and the male characters least masculine to most masculine (or most feminine to least feminine), the

rankings reveals that the characters' masculinity and femininity also go in pairs (see Figure 9). Bingley is the most feminine (or least masculine) in his language speech in comparison to the other male characters. Jane is the same for the female characters. The trend continues and we find that each the each character in a romantic couple is ranked the same. (The exception lies with Charlotte and Caroline since there is an uneven number of female and male characters, though maybe this would suggest that Caroline would do best with a Mr. Collins.) Elizabeth and Darcy fall in the middle, as they are the couple given the most attention, complexity of character change, and amount of dialogue through the course of the book (see Figure 10). It should be noted that these are not the rankings of most similar characters. Lydia, for example, is actually closer in overall feminine language use to Mr. Bennet or Darcy before Wickham (See Figure 8). Yet comparatively, she is most like Wickham according to ranking. Linguistic Style Matching (LSM), a popular linguistic theory, may explain the similarities between characters. This theory posits that the closer the relationship between any two people becomes, the more similar their language (Ireland et al., 2011). (Future research comparing character progression over time and how they speak with one another would be more conclusive.)

In all, the LIWC2015 data reveals a consistent pattern in gender language use when comparing the results from Newman et al. (2008) to present data. Analysis of the data shows various trends that emerge for male, female, and paired language use.

Comparison with Questionnaires

Along with collecting the data through LIWC analysis, questionnaires were sent out to a small pool of participants who had varying exposure to *Pride and Prejudice* and different levels of literary training. The questionnaire asked participants to rank male and female characters

according to masculine and female gender language (see Appendix A for sample questionnaire). Results varied widely. Figure 11 gives the average results. Participants believed Jane, Caroline, and Charlotte to have the most feminine language, whereas Elizabeth, Lydia, and Mrs. Bennet had the least feminine language. Reasoning for these responses varied as much as the responses themselves. It is possible that participants, who were all female, could not separate the controversial personalities and character traits of Lydia and Mrs. Bennet from their gender language. Thus, they were claimed to be least feminine for their lack of propriety. This is unlike the statistical data from LIWC, which shows Austen as separating female gender language from likability or propriety.

Participants also ranked male characters according to gender language use. With varied results (though less varied than those for female characters), Darcy ranked as having the most masculine language, followed by Mr. Bennet, Bingley, Wickham, and lastly Collins with the least masculine language (see Figure 12). Results are almost the opposite of LIWC findings, with the exception of Darcy who instead placed in the middle. Participants again may not have separated likeability from masculine language usage. While Austen seems not to have done this either (portraying male characters by likeability), she, conversely, portrayed the least likable characters as the most masculine in their language. Thus, the questionnaire results, though sampling public perception of gender language use, did not coincide with trends in Austen's dialogue.

Comparison with Newman et al. (2008) Findings

The Newman et al. (2008) study analyzed 14,000 text samples of all genres largely from modern American English. Even with the differences in the sample texts between this study and

Austen's 1813 dialogue, many similarities emerge. Eight of the ten total dimensions studied were consistent with the gender dominance for each dimension. That is, four of the five dimensions which Newman et al. found to be characteristic of male speech were also more common in the male characters' speech of *Pride and Prejudice*, just as four of the five dimensions identified by Newman et al. as female language characteristics were also consistent in Austen's female characters (2008).

The reliability of this data can be compared with the LIWC tool and Newman et al. (2008) findings. The dialogue language samples used were on average larger than those of the Newman study, and the larger the language sample the more reliable the LIWC data (Newman et al., 2008; Pennebaker, 2011). (See Figures 13 and 14 to compare language sample for female and male characters.)

In addition to word count, LIWC states that a reliable text analysis occurs when a minimum of 80% of the words in the text sample correlate to the words in the LIWC2015 dictionary (Tool category: 'dic'). Figure 15 shows that the dialogues of all the eleven characters analyzed was above 85%. Thus in word count and dictionary correlation, the analysis of the samples was statistically sound.

Lastly, the current study revealed an even higher percentage difference between male and female speech for a given dimension. This occurred in all eight dimensions in which *Pride and Prejudice* results were consistent with Newman et al. findings. As shown in Figure 16, Austen's writing for her dialogue showed an even greater percentage difference between her male and female characters than Newman et al. findings. The exception lies in the numbers category, which conversely showed more usage by female characters than male. Additionally, Austen's

writing revealed a higher difference in all of the female dimensions for the four categories which were consistent with previous findings (see Figure 17). Thus, all of the results for the eight consistent categories were stronger than those of the previous study. (See Appendix B for raw data score comparison between the studies.) Thus, when comparing the two studies, we see a greater disparity between male and female speakers in the dialogue of *Pride and Prejudice* than in the 14,000 generally modern samples of English.

Austen's Intentional Use of Gender Speech

Other findings and comparisons reveal more about Austen's intentional construction of dialogue which may yield further insight into her use of gender languages.

Numerical observations reflect intentional dialogue construction. Simple dialogue analysis of Jane Austen yields data that indicates her overall intentionality with dialogue. Even after parsing out dialogue descriptions, internal thoughts, and the other three aforementioned categories of character expression (thus leaving only explicit dialogue), much insight can be gained that reflects Austen's literary tactics. For example, the following observations can be made. First, Jane and Bingley never converse in dialogue with one another throughout the course of the book. Even in the proposal scene, the only words uttered by Bingley are as a question to Elizabeth, "Where is your sister?" (Austen, 1813, Chapter 55). This may reflect the true focus of the story lies in Elizabeth and her conquests (which word count alone can reveal, as she has about four times the dialogue of the average character). This may be related to the second observation: Wickham only utters words to Elizabeth in the entire novel, never to his arch nemesis Darcy, nor to his own wife Lydia. A third interesting observation reflects the personalities and dynamics of characters as well: Mrs. Bennet almost entirely initiates

conversations with outsiders, be it with Collins, Bingley, Darcy, or otherwise. Mr. Bennet rarely does, thus revealing Bennet family dynamics from numbers alone. Even from statistical analysis of dialogue in this way, then, one can learn of Jane Austen's character development, and more poignant here, her intentional construction of dialogue.

Intentional Gender Dialogue. Pennebaker previously conducted a similar study analyzing fifteen playwrights and screenwriters and their depiction of female and male characters (Pennebaker, 2011). He found that female authors sometimes wrote more feminine dialogue for both male and female characters. Sometimes both were more masculine. The same was true for male writers. One third of the writers wrote female gender dialogue for female characters and male gender dialogue for male characters. Thus, it is not a foregone conclusion that authors write dialogue according to how men and women are statistically proven to speak. Authors may purposefully change the dialogue as a literary technique. Therefore, whether explicitly intentional or naturally so, Jane Austen accurately wrote her characters' dialogues in accordance with statistically proven dimensions of gender speech.

Conclusions and Future Study

The ten dimensions of male and female language previously studied by Newman et al. (2008) and further interpreted by Pennebaker (2011) reveal a consistent pattern of gender language usage in the characters of Jane Austen's *Pride and Prejudice*. Of the five dimensions statistically shown to be common to male speech (more words per sentence, "big words," articles, prepositions, and numbers), only one dimension (numbers) was inconsistent between the previous study and current study. Of the five dimensions of female speech (the use of more personal pronouns, negation, verbs, negative emotions, and certainty), only the personal pronoun

dimension yielded mixed results in Austen's dialogue. Additionally, Austen's dialogue revealed a greater disparity between male and female language in all eight of the consistent dimensions when compared with Newman et al. (2008) findings.

While individual characters show variety and complexity in gender dimensions, Austen's dialogue reveals interesting trends in gender language along these ten dimensions. First, male gender speech of the characters Collins, Wickham, Darcy, Mr. Bennet, and Bingley painted a distinguishable pattern. The generally more-liked characters Mr. Bennet and Bingley had the most feminine (least masculine) language usage. On the other hand, the more controversial characters (Collins and Wickham) had the most male language of any of the male characters. Mr. Darcy, the most developed character, falls in the middle. This may be from his transformation from a more reserved to a more open and expressive persona through character development. Future research on possible change in his speech patterns may reveal whether he slowly incorporates more feminine language usage (to become more likable) as the story develops. If this is the case, Austen may have made feminine expression her basis for the portrayal of male characters.

Austen's female characters, on the other hand, do not follow as easily-interpretable of a pattern, as the female characters most similar to one another according to gender language were in the following pairs: Jane and Mrs. Bennet, Lydia and Elizabeth, and Charlotte and Caroline. As other linguistic and literary studies have indicated, Austen makes her female characters the main focus of her story (as word count averages alone would indicate) and intentionally diversifies feminine language to depict and contrast female roles in her contemporary society.

The last of the principal trends in gender language portrayal lies in Austen's romantic pairing, intentional or not, of characters. Ranking the female characters from most feminine to least feminine and then male characters from least masculine to most masculine (or most feminine to least feminine) places Jane and Bingley as the most feminine in their language usage, followed by Mr. and Mrs. Bennet, then Elizabeth and Darcy, Lydia and Wickham, Collins and Charlotte, and Caroline. Romantic couples, therefore, have the same ranking on the gender language spectrum. This could be explained through Linguistic Style Matching or Austen's own pairing of characters according to their language expression.

While these findings revealed more about the character portrayal and dialogue in *Pride and Prejudice* in addition to overall consistency with the findings of Newman et al. (2008), limitations to the study may also be factors in interpreting the data, which should be accounted for in future study. The LIWC dictionary version used in the 2008 study was LIWC2007 whereas the present study used the LIWC2015 dictionary. The work of Pennebaker et al. (2015) explains the differences between the two dictionaries to compensate for any affected data. This may account for some differences in the data.

The questionnaire data was limited since only a small sample of responses was collected. Additionally, all participants were female and had varying levels of interaction with *Pride and Prejudice* (ranging from the movie-goer to academics trained in literary criticism). These factors may all limit the results of the questionnaire. Future survey of public opinion would therefore be more statistically significant with a larger pool of participants.

Additional factors regarding setting and society may be involved. Austen lived in and wrote for a society of two hundred years ago. There may be social differences in her gender

language conventions that would affect interpretations when compared with modern writers (though the sample texts from the Newman et al. study were not necessarily exclusively modern texts). Next, gender may not be the only interpretation for some of these dimensions. Future research grouping Austen's characters by socioeconomic levels, for example, may indicate to what degree social class affects these ten categories of language.

Despite all of these potential caveats, the gender language in Jane Austen's dialogue was largely consistent with that of even modern, younger writers of previous studies. This may suggest a long-lasting trend in gender language across time periods and genres. Only future linguistic study can tell us more.

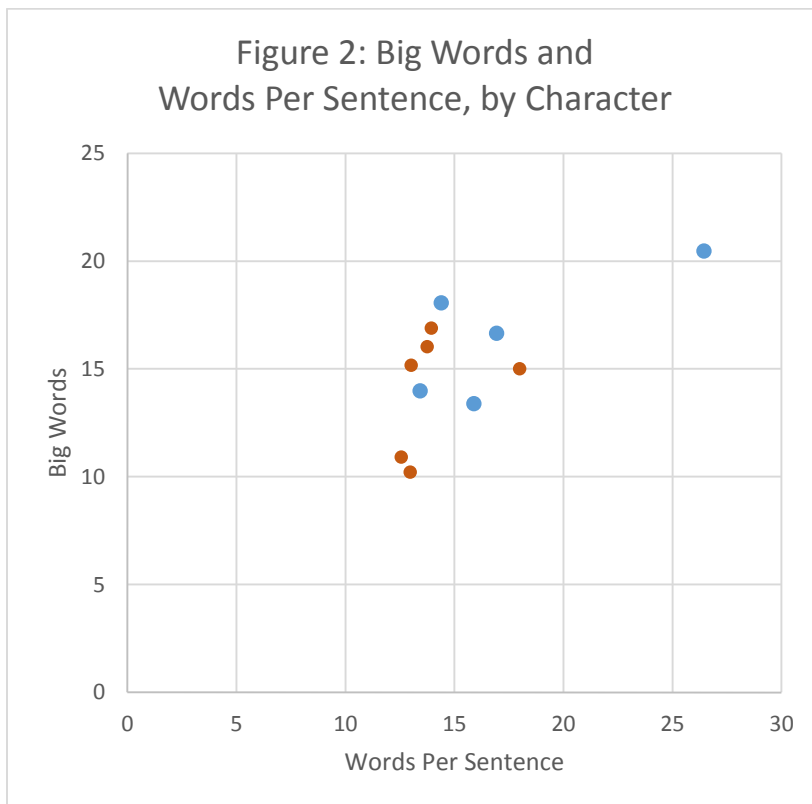
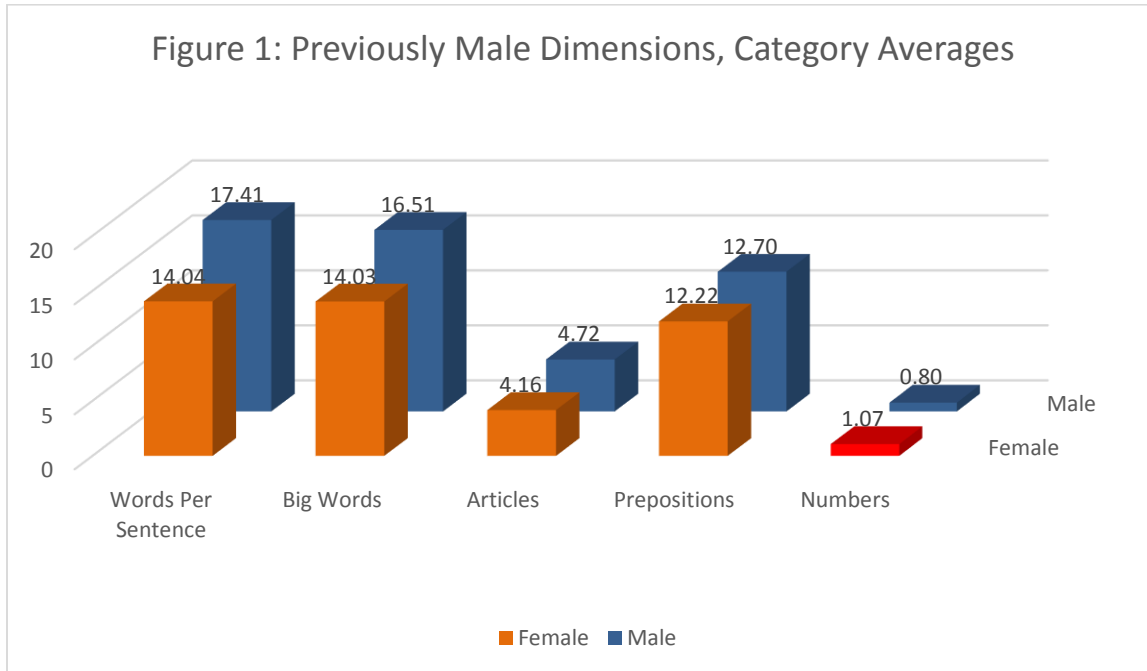
References

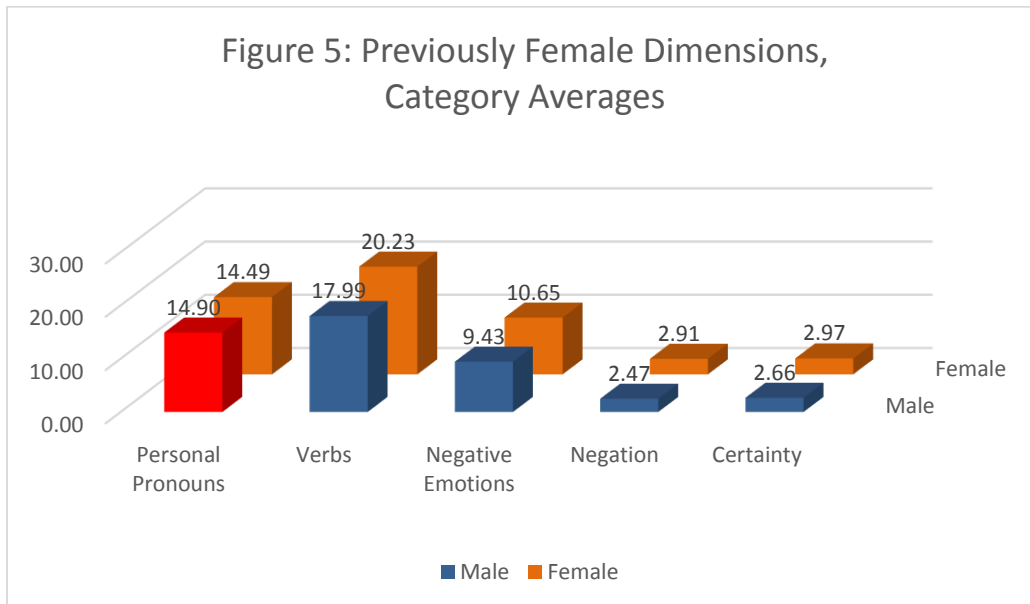
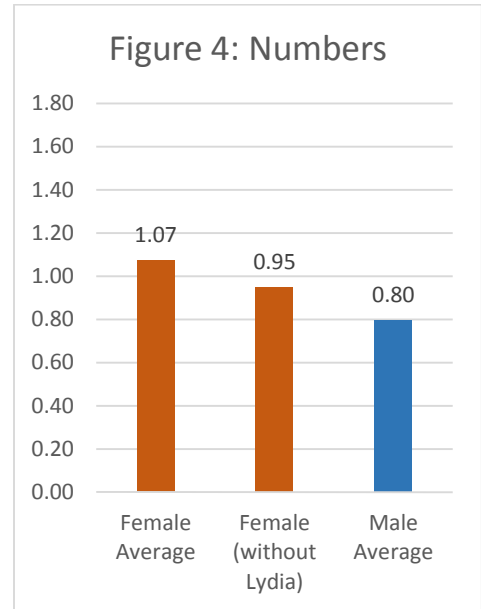
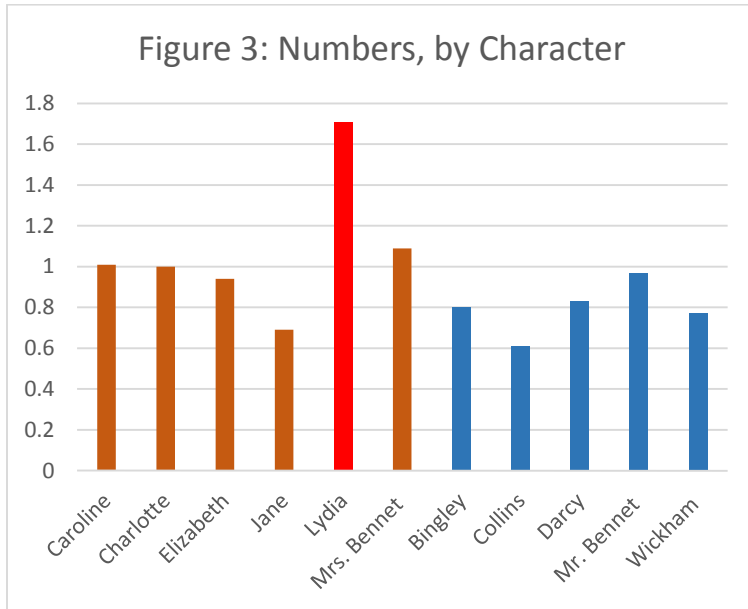
- Austen, J. (1813). *Pride and Prejudice* [Project Gutenberg eBook]. Retrieved from <https://www.gutenberg.org/files/1342/1342-h/1342-h.htm>
- Burrows, J. F. (1987). *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. New York, NY: Oxford University Press.
- Cameron, D. (2005). Language, gender, and sexuality: Current issues and new directions. *Applied Linguistics*, 26(4), 482-502.
- González-Díaz, V. (2014). 'I quite detest the man': Degree adverbs, female language and Jane Austen. *Language and Literature*, 23(4), 310-330. doi:10.1177/0963947014534123
- Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1), 39-44. doi:10.1177/0956797610392928
- Lakoff, R. (1973). Language and woman's place. *Language in Society*, 2(1), 45-80.
- Michaelson, P. H. (2002). *Speaking volumes: Women, reading, and speech in the age of Austen*. Stanford, CA: Stanford University Press.
- Mulac, A., & Lundell, T. L. (1986). Linguistic contributors to the gender-linked language effect. *Journal of Language & Social Psychology*, 5(2), 81.
- Newman et al. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45, 211–236.
- Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York, NY: Bloomsbury Press.

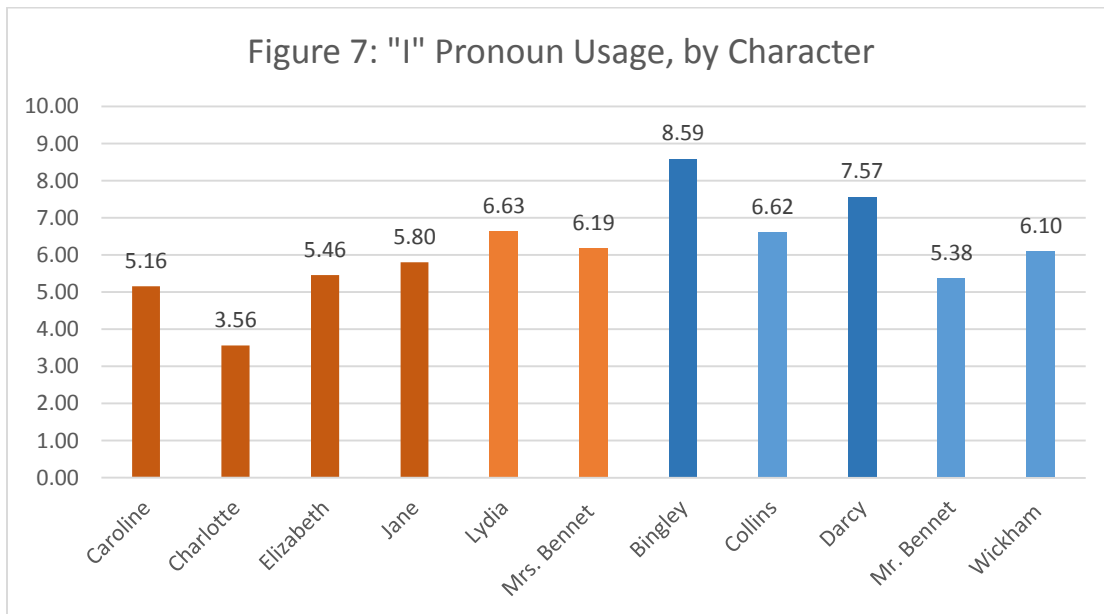
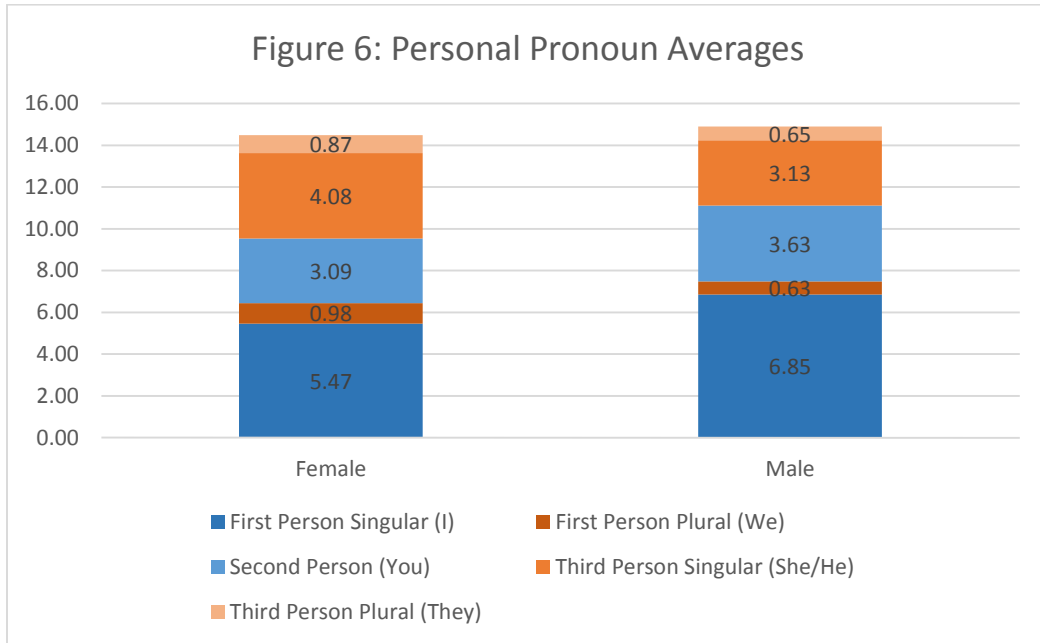
- (2013). How authors' forgettable words reveal their personality and social behaviors. *Information Design Journal (IDJ)*, 20(3), 267-272. doi:10.1075/idj.20.3.06.2
- Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.
- Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24-54.

Table 1	
<i>Possible LIWC Categories for Analysis (Adapted from Newman et al., 2008; Pennebaker, 2011)</i>	
Dimensions for Male Gender Speech	Dimensions for Female Gender Speech
<u>Five Usable Categories</u>	<u>Seven Usable Categories</u>
Articles	Personal pronouns
Big words	Negative emotion words
Nouns [unusable]*	Cognitive words
Prepositions	Social words
Numbers	Verbs
Words per sentence	Negations
Swear words [not applicable]**	Certainty words
	Hedge phrases [unusable]*
<p>*These were unusable since the LIWC2015 categories did not coincide with these categories **This category was judged to be not applicable to Jane Austen's writing</p>	

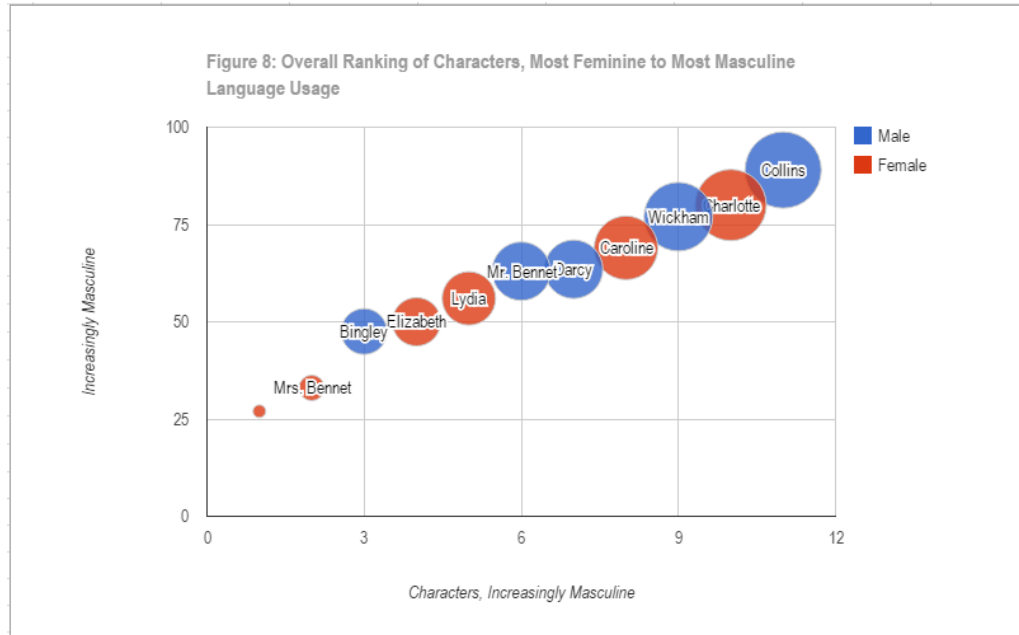
Table 2			
<i>Dimensions for Gender Language Analyzed in Current Study</i>			
Dimensions for Male Gender Speech		Dimensions for Female Gender Speech	
<u>Dimension</u>	<u>LIWC Category</u>	<u>Dimension</u>	<u>LIWC Category</u>
Big words	Sixltr	Personal Pronouns	ppron
Articles	article	Verbs	verb
Prepositions	prep	Negative Emotion	negemo
Numbers	number	Negations	negate
Words Per Sentence	WPS	Certainty Words	certain



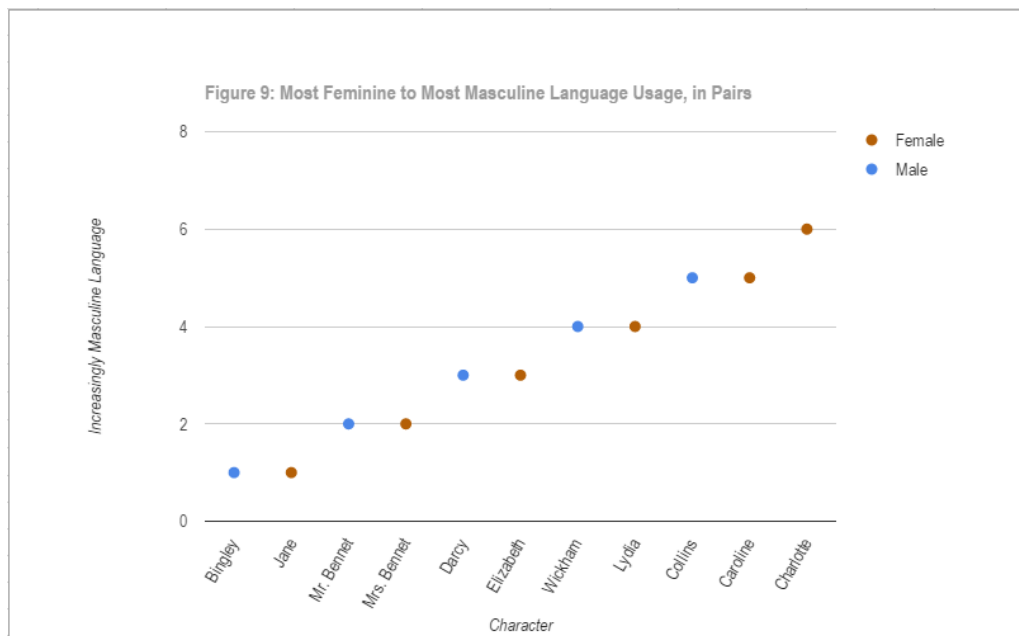


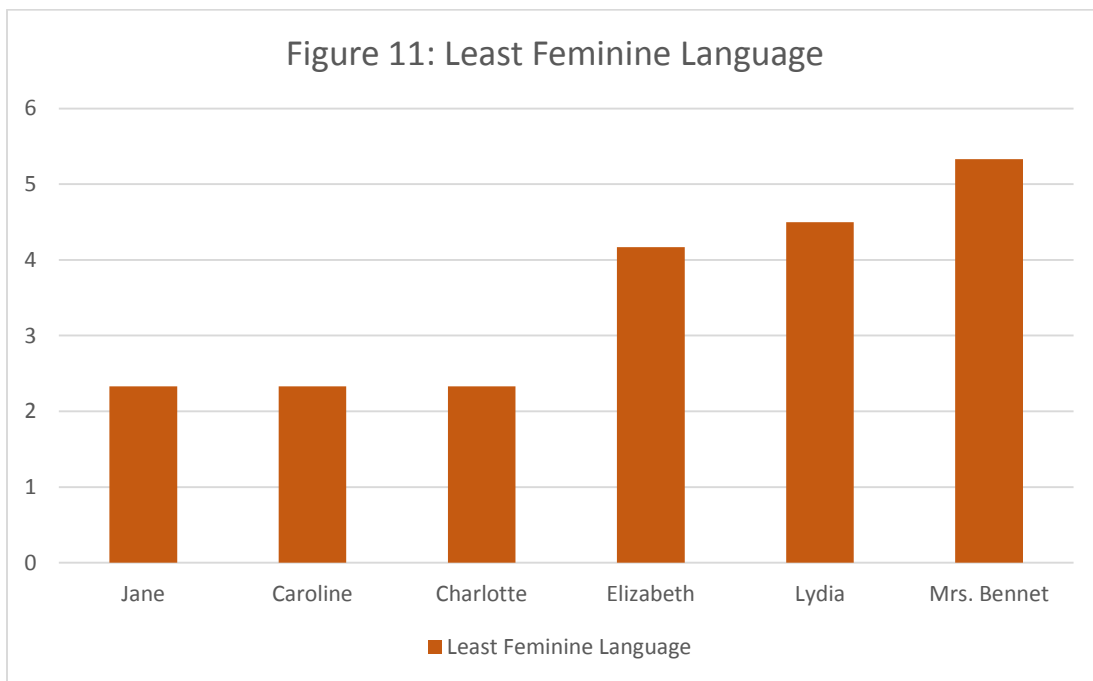
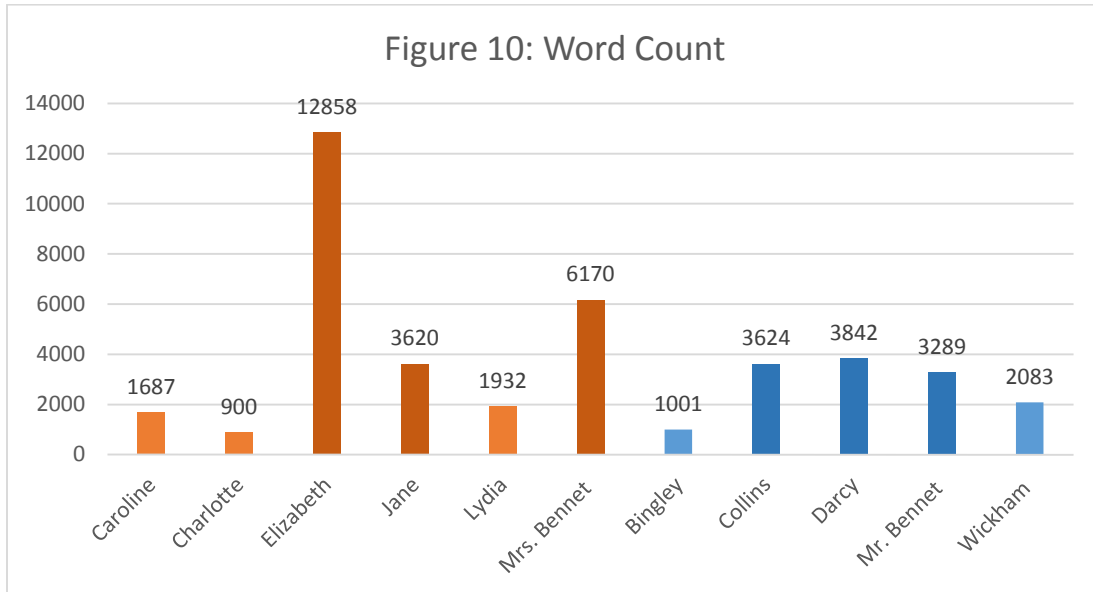


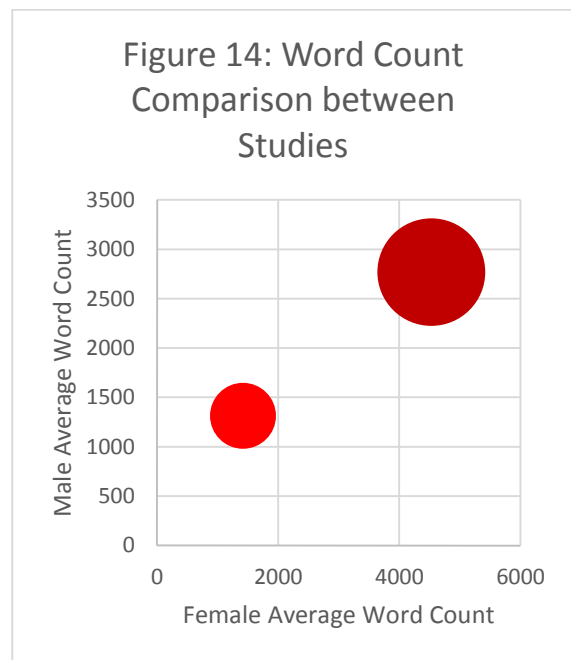
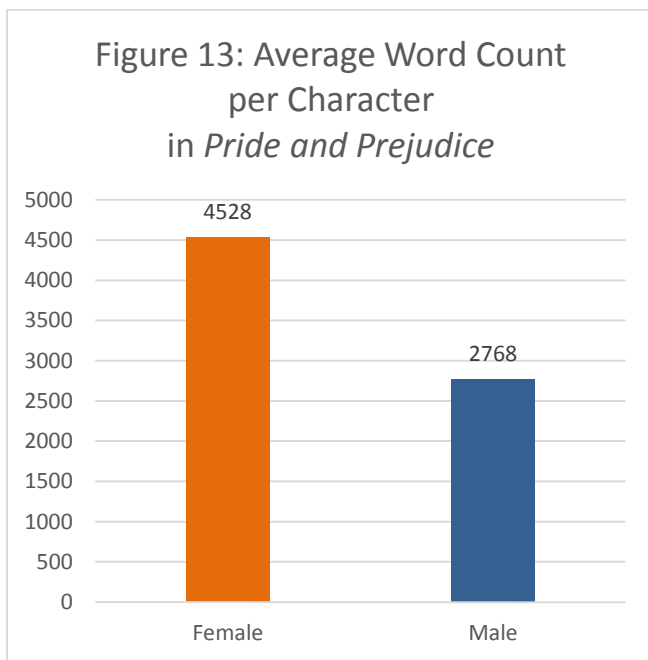
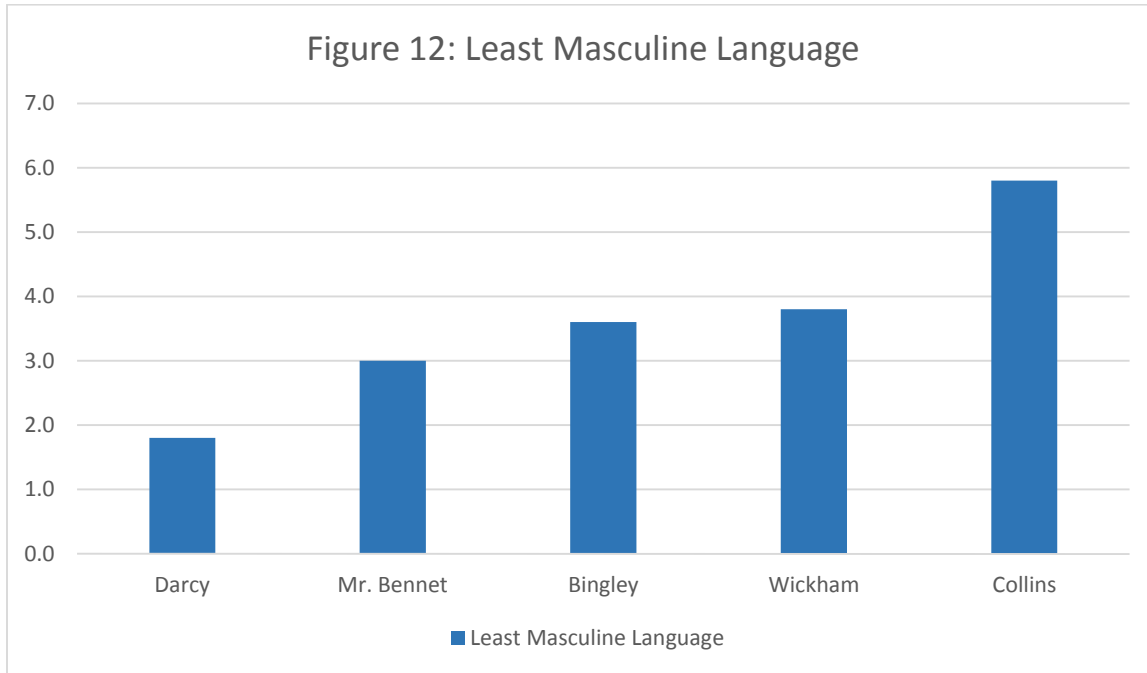
Note: Lydia and Mrs. Bennet have scores higher still than Collins, Mr. Bennet, and Wickham.

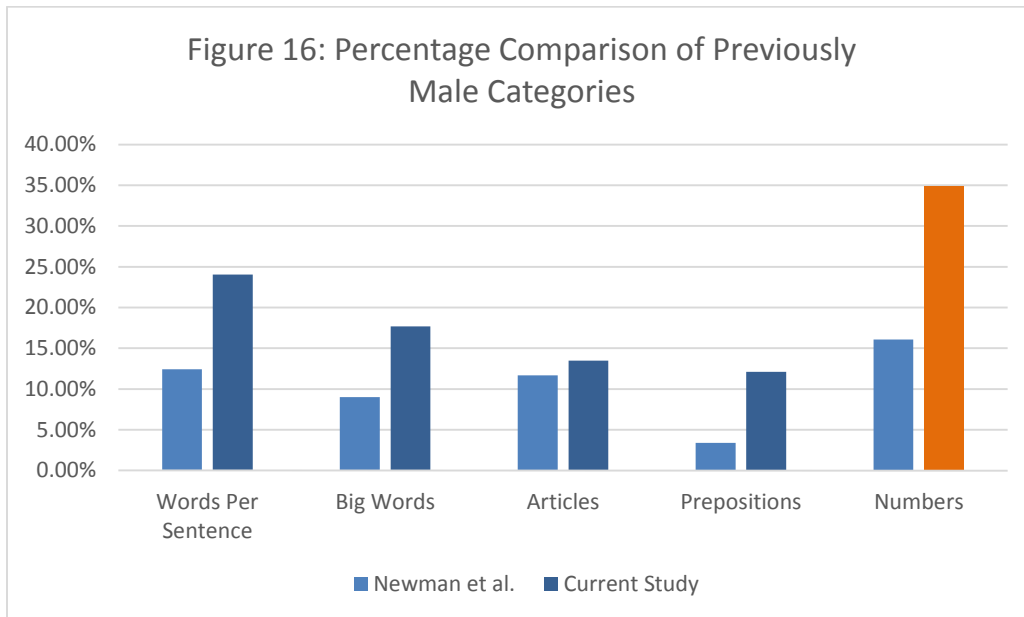
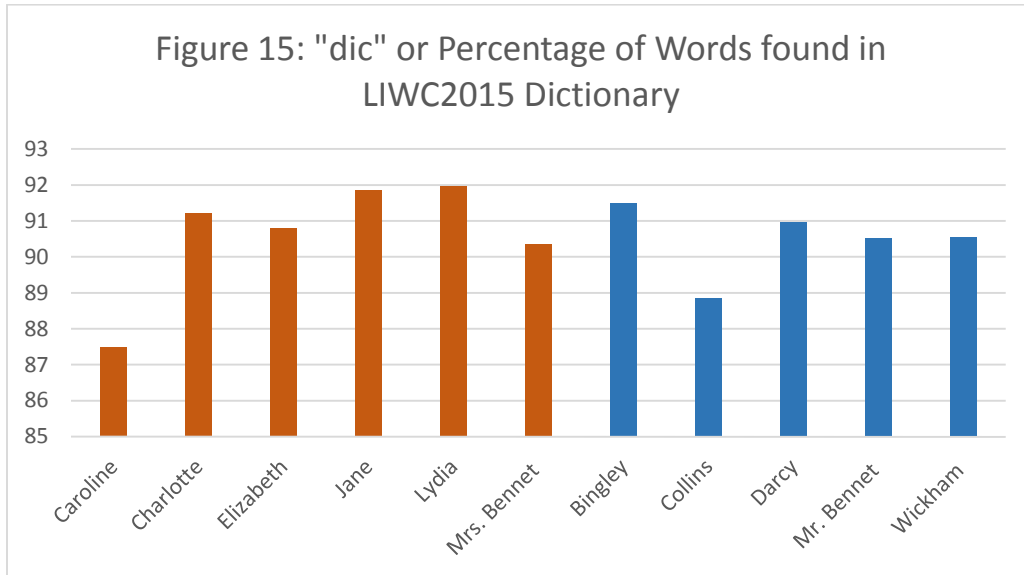


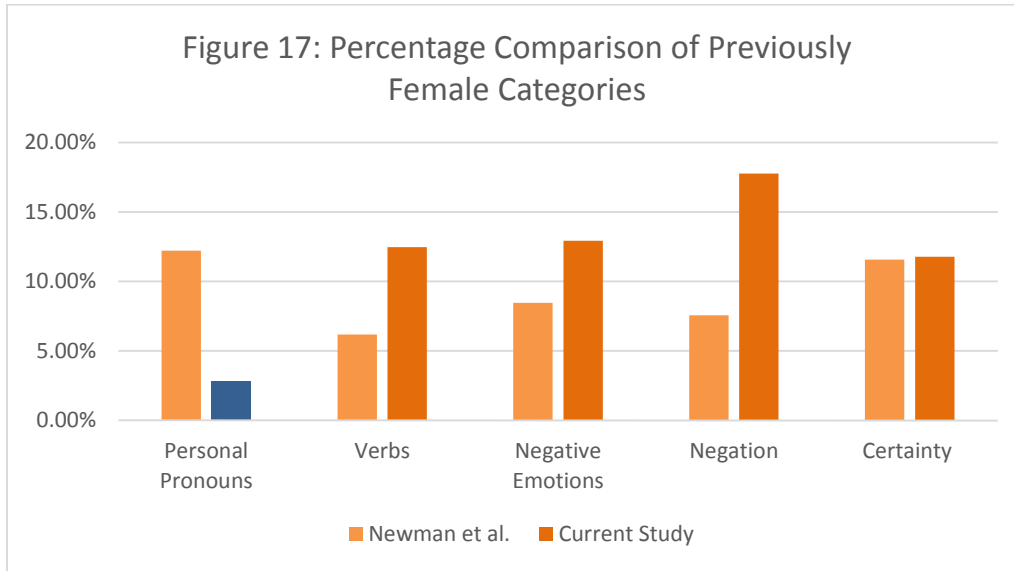
Note: The above chart takes into account an average ranking in each category. The closest to 0 is most characteristic of female language usage while the furthest is characteristic of male language usage.











Appendix A
Questionnaire Questions

***Pride and Prejudice* Questions:**

In *Pride and Prejudice*, how do characters express their masculinity and femininity?

How does Jane Austen use character's speech to express their gender?

List the following characters in order of their use of feminine language (first being most feminine): Caroline Bingley, Charlotte Lucas, Elizabeth Bennet, Jane Bennet, Mrs. Bennet

Explain your reasoning for the characters you chose to have the *most* and *least* feminine language.

List the following characters in order of male gender language (first being most masculine): Mr. Bennet, Mr. Bingley, Mr. Collins, Mr. Darcy, Mr. Wickham

Explain your reasoning for the characters you chose to have the *most* and *least* masculine language.

Personal Questions:

What is your exposure to *Pride and Prejudice*? (Check all that apply:)

- 2005, Keira Knightley movie version of the book
- 1995, BBC version of the book
- Read the book
- Heard from friends
- Other: _____

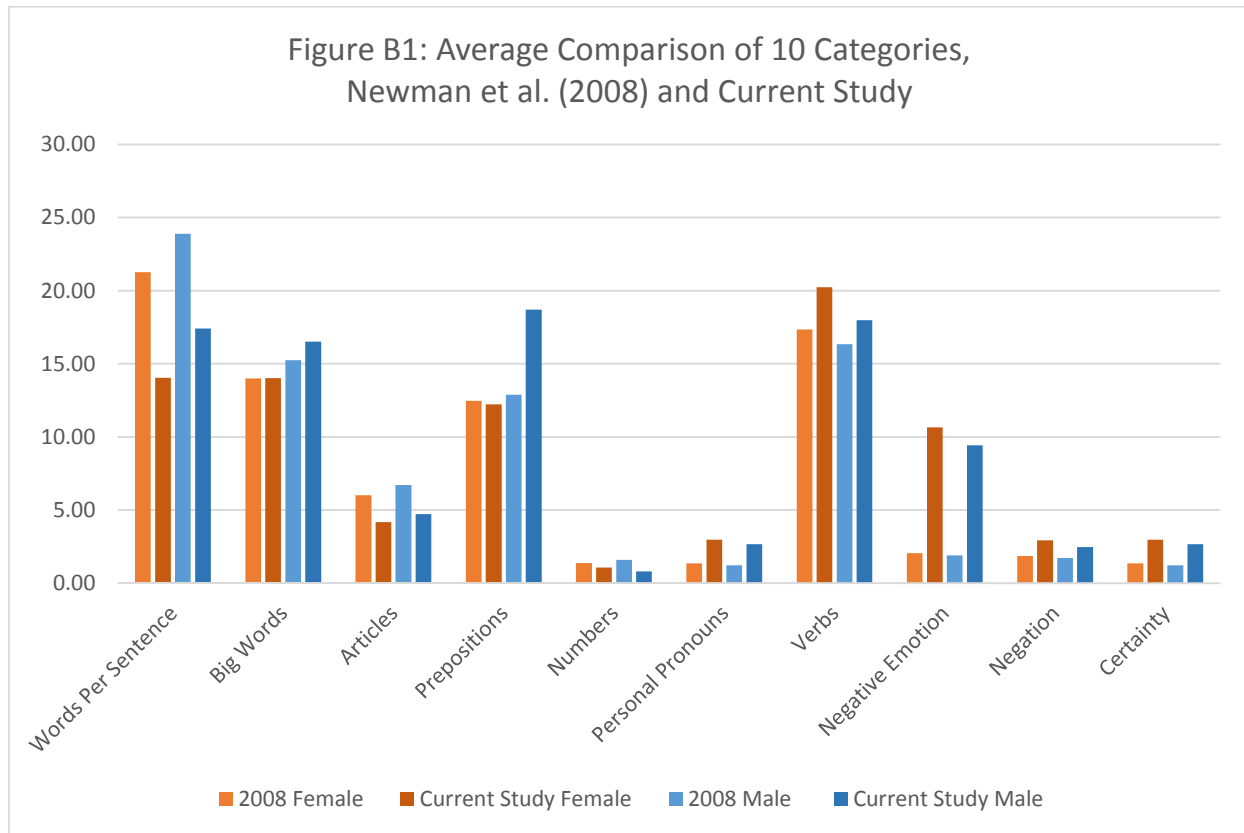
What is your exposure to Jane Austen in general? (List any other movies or books you have watched or read).

Sex: M / F

Occupation: Student Professor Other

Appendix B
Comparison of Data between Newman et al. (2008) and Current Study

Table B1					
<i>Male and Female Mean Scores, Newman et al. (2008) and Current Study (2016)</i>					
<u>Dimension</u>	Female Mean		Male Mean		<u>Notes:</u>
	<u>(2008)</u>	<u>(2016)</u>	<u>(2008)</u>	<u>(2016)</u>	
WPS	21.26	14.04	23.90	17.42	<i>Newman et al. report small effect for their data</i>
Big Word	13.99	14.03	15.25	16.51	
Numbers	1.37	1.07	1.59	0.80	
Negation	1.85	2.91	1.72	2.47	
Article	6.00	4.16	6.70	4.72	
Preposition	12.46	12.22	12.88	13.70	
Negative Emotion	2.05	10.65	1.89	9.43	
Certainty	1.35	2.97	1.21	2.66	
Personal pronouns	14.24	14.49	12.69	14.90	
Verbs	17.34	20.24	16.33	17.99	
<i>Note: Numbers in bold mark the dimensions with inconsistent results between the two studies</i>					



Appendix C
Character Rankings by Male and Female Dimensions of Speech

<u>Male</u>										
<u>Dimensions</u>	<u>WPS</u>	<u>11</u>	<u>Sixltr</u>	<u>11</u>	<u>article</u>	<u>11</u>	<u>prep</u>	<u>11</u>	<u>number*</u>	<u>11</u>
Bingley	15.89	8	13.39	3	4.00	3	12.89	6	0.80	4
Caroline	13.94	6	16.89	9	5.63	11	11.74	3	1.01	9
Charlotte	18.00	10	15.00	5	4.44	6	13.67	9	1.00	8
Collins	26.45	11	20.47	11	5.16	9	15.67	11	0.61	1
Darcy	14.39	7	18.06	10	4.61	7	13.69	10	0.83	5
Elizabeth	13.75	5	16.02	7	4.28	5	13.23	7	0.94	6
Jane	13.02	3	15.17	6	2.82	1	11.91	4	0.69	2
Lydia	12.97	2	10.20	1	4.09	4	11.44	2	1.71	11
Mr. Bennet	13.42	4	13.99	4	4.65	8	13.44	8	0.97	7
Mrs. Bennet	12.57	1	10.91	2	3.70	2	11.35	1	1.09	10
Wickham	16.93	9	16.66	8	5.18	10	12.82	5	0.77	3
<u>Female</u>										
<u>Dimensions</u>	<u>ppron*</u>	<u>1</u>	<u>negate</u>	<u>1</u>	<u>verb</u>	<u>1</u>	<u>negemo</u>	<u>1</u>	<u>certain</u>	<u>1</u>
Bingley	15.18	4	2.80	5	19.98	4	1.60	7.5	3.20	3
Caroline	14.05	9	2.37	9	17.90	9	2.13	4	2.96	5
Charlotte	12.89	11	2.44	7	19.56	5	1.11	11	2.67	8
Collins	13.96	10	1.85	11	15.15	11	1.6	7.5	2.73	6.5
Darcy	15.04	5	2.81	4	18.64	8	2.58	1	2.73	6.5
Elizabeth	14.43	7	3.46	2	19.54	6	2.22	3	3.30	2
Jane	15.69	1	3.62	1	20.55	3	2.51	2	2.98	4
Lydia	14.39	8	2.43	8	22.83	1	1.29	10	2.48	9
Mr. Bennet	15.32	3	2.65	6	18.88	7	1.73	6	2.34	10
Mrs. Bennet	15.49	2	3.16	3	20.99	2	1.39	9	3.45	1
Wickham	14.98	6	2.26	10	17.28	10	1.92	5	2.30	11
<i>Note: 1 is most feminine, 11 is most masculine</i>										
<i>*Inconsistent dimension between Newman et al. (2008) findings and the current study</i>										