Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Mina Song April 12, 2022

GENDER BIAS IN CHILDREN'S BOOKS:

how did the 19th century children's literature represent gender?

by

Mina Song

Lauren Klein
Adviser

Quantitative Theory and Methods

Lauren Klein

Adviser

Davide Fossati

Committee Member

Roberto Franzosi

Committee Member

2022

GENDER BIAS IN CHILDREN'S BOOKS:

how did the 19th century children's literature represent gender?

By

Mina Song

Lauren Klein

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2022

Abstract

GENDER BIAS IN CHILDREN'S BOOKS:
how did the 19th century children's literature represent gender?
By Mina Song

As children begin to develop their gender identity, one of the most influential factors is the book they read. However, gender bias is present in many children's books, which can be seen in the gendered depiction of characters. To explore this issue, this project investigates how gender is portrayed in the ChiLit corpus, which consists of 70 children's books published in the 19th century. Using methods of descriptive statistics, sentiment analysis, and word embedding models, I detect and document instances of gender bias in the corpus: namely, the male characters appear much more than female characters, and both genders are associated with stereotypical gender roles for the 19th century. These findings clearly show gender bias in the space of book characters and their dialogue in the 19th century children's literature.

GENDER BIAS IN CHILDREN'S BOOKS:

how did the 19th century children's literature represent gender?

By

Mina Song

Lauren Klein

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Quantitative Theory and Methods

2022

Table of Contents

Chapter 1: Introduction 1

Chapter 2: Literature Review 2

Chapter 3: Corpus 3

Chapter 4: Methods and Results 4

4.1 BookNLP and Descriptive Statistics 5

4.2 Sentiment Analysis 10

4.3 Non-contextualized Word Embeddings, Word2Vec 13

4.4 Contextualized Word Embeddings, BERT 15

Chapter 5: Conclusion 19

References 21

Appendix 24

I. Corpus List 24

II. Group Words 27

Chapter 1: Introduction

As children begin to develop their gender identity, one of the most influential factors is the book they read. By identifying and evaluating the characters and events in literature, children consider their own actions, beliefs, and emotions (Mendoza et al., 2001). However, gender bias is present in many children's books; female characters are underrepresented than male characters (Casey et al, 2021), and both genders are associated with stereotypical adjectives (Charlesworth et al., 2021). Such biases may perpetuate gender stereotypes and other gendered associations (Lewis et al. 2021).

In this project, I investigate how gender is represented in a corpus of the 19th century children's literature. I hypothesize that the gender roles pervasive during the time period will also be seen in the corpus. According to Welter (1996), the gender ideologies of the time in the United States and Britain, known as the Cult of Domesticity, describe women as the moral protector of home and family life. The four values of piety, purity, submissiveness, and domesticity are required for women in society; accordingly, women's actions were limited compared to men. Therefore, in addition to male characters appearing more than the female characters in the corpus, I would also expect females to be portrayed as pure, modest, passive, and supportive, and to be portrayed in religious or domestic settings.

Previous quantitative research on gender representation in literature has focused on methods of descriptive statistics and non-contextualized word embeddings. For example, Underwood et al. (2018) explored gender in fiction by analyzing the changes in the number of words describing female characters and the number of books written by female authors over time. Charlesworth et al. (2021) used word embeddings to quantify the association between gender groups and stereotypical traits. This project not only explores descriptive statistics and non-contextualized word embedding models, but also applies a contextualized word embedding model, a distilled version of Bidirectional Encoder Representations from Transformer

(DistilBERT) (Sanh et al., 2019). DistilBERT is relatively new but is increasingly utilized in many applications for classification.

The purpose of this project is to explore the gender bias implied in the 19th century children's literature using these and other methods in natural language processing (NLP) and machine learning. I compare the representation of male and female characters in children's books with four methods: descriptive statistics, SentiArt, Word2Vec, and BERT. The results of the four methods clearly show gender bias in the 19th century children's books. The male characters outnumber and speak more than the female characters, and also they are characterized with more words. Each gender is also associated with typical nineteenth-century gender roles. These results support and extend the findings of previous works and also suggest that the gender roles in society are implied in the children's books published during the periods.

Chapter 2: Literature Review

Much quantitative research has been conducted about gender bias in children's books; the results have revealed that female characters are underrepresented as compared to male characters, and that both genders are associated with stereotypical words. Most of this research has focused on 20th and 21st centuries children's books, with methods of analysis that include descriptive statistics and non-contextualized word embeddings. My project uses these methods, as well as several others, in order to extend the scope of children's literature to 19th century children's books.

More specifically, Underwood et al. (2018) study the changes in gender representation in fiction over time by examining 104,000 English-language fiction published from 1703 to 2009. Using the BookNLP pipeline, the same that I use in this project, they identify words that describe male and female characters (known as "words used in characterization"). Then, they compute and plot the proportion of words used in characterization that describe women as

compared to men; the proportion declined from the middle of the 19th century to the middle of the 20th.

Similarly, Casey et al. (2021) examined 3,280 children's books for 0-16 years published between 1960 and 2020 to provide an up-to-date estimate of the rates of gender representation in books. They identified the protagonist of each book, categorized its gender as male or female, and then computed the male-to-female ratio of protagonists across time with descriptive statistics. They found that female protagonists remain underrepresented in the most recently published books even though the proportion of male protagonists decreased between 1960 and 2020. They also explored the effect of author gender and found that female authors showed less male overrepresentation except when writing books featuring non-human central characters.

By using word embedding models, Lewis et al. (2021) measure gender bias in 247 popular, contemporary children's books for 0-5 years. They train word embedding models with their corpus, and then estimate the gender associations via cosine similarity between words. Their results show that females are associated with mental states and interactions with others while males are associated with physical events.

Chapter 3: Corpus

The corpus used in this project is the new GLARE 19th Century Children's Literature (ChiLit). It originally consisted of 71 books by 38 authors, but I excluded one book titled 'Stalky&Co.' due to Unicode encoding error [Appendix I]. The corpus is considered to be a representative sample of the "Golden Age" of children's literature in English, guided by Peter Hunt's principles (2001) which define the characteristics of representativeness for children's literature using three principles. The first principle aims to cover "a reasonable representation of what was written for and read widely by English-speaking children in the 19th century". The second principle aims

at choosing “historically significant, or good examples of their kind” (particularly in terms of the newly emerging genres, e.g. fantasy or school story). The third principle ensures that the books selected are “readable today” with the “heaviest emphasis … on books that ‘entertain’ rather than instruct” (Hunt 2001: xv-xvi); therefore, all the books selected for ChiLit have been also recently reprinted at least once after 2010. However, unlike Hunt’s principles, ChiLit excluded translations (or retelling myths, legends, and other folklore texts), nursery rhymes, classic fairy tales, poetry, and books that were not written by British authors (or those with British background) (Čermáková, 2017). Also, the gender ratio of authors is balanced in ChiLit; 35 books were written by male writers and 35 by female writers. Collecting the corpus was quite simple because data was directly observable as raw text. There is no required consent for using the data because all the U.S. copyrights of books are expired.

Chapter 4: Methods and Results

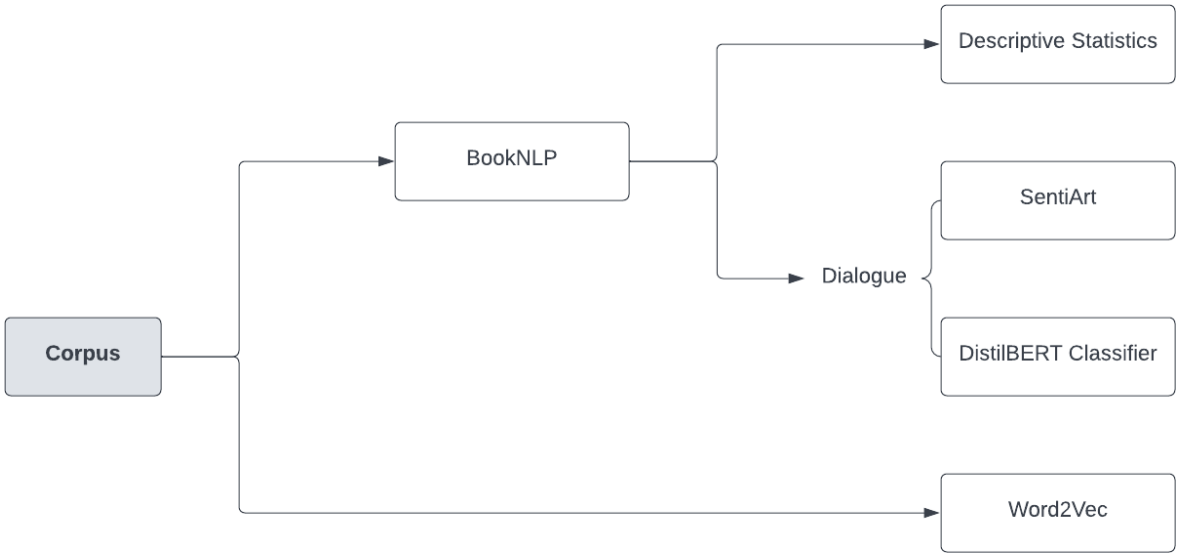

Figure 1. a diagram of entire project pipeline

To compare gender representation in the corpus, I use the four methods displayed on the right side of the diagram above: descriptive statistics, SentiArt, DistilBERT classifier, and Word2Vec. The Word2Vec method takes raw text of the corpus as its input, but the other methods require specific parts of each text; therefore, I ran a BookNLP (Bamman et al., 2014) pipeline before implementing the three other methods. The BookNLP pipeline identifies the characters in each book and their gender, the dialogue associated with each character, and the parts of speech for each associated word in each line of dialogue (verbs, objects, and modifiers). In the descriptive statistics method, all of the elements identified by BookNLP are used. Both the SentiArt and DistilBERT methods take only the identified lines of dialogue and the speaker as their input. These methods will be further explained in the sections below.

4.1 BookNLP and Descriptive Statistics

Method

As a first step, I used a natural language processing pipeline called BookNLP (Bamman et al., 2014). BookNLP provides various functions including part-of-speech tagging, dependency parsing, entity recognition, character name clustering and coreference resolution, quotation speaker identification, supersense tagging, event tagging, and referential gender inference. It takes raw text as input and produces JSON files storing the results of its functions. In this project, BookNLP first annotates entities and then clusters character names to a group—for example, "Oliver" and "James Oliver" are grouped into a single entity. Then, it infers the gender of characters by associating them with pronouns and honorifics. It reasonably allocates gender to each character in a book; women are identified with 94.7 precision and men are identified with 91.3% precision, with a collection of 104,000 works of fiction (Underwood et al., 2018). Also, BookNLP traces the grammatical dependencies to annotate the characters' actions, objects, and modifiers by tagging the associated verbs, nouns, and adjectives. Lastly, it detects lines of dialogue along with their speaker. For example, in the sentences from Alone in London,

“ ‘Ah! it's like another world!’ said the old woman, shaking her head slowly ”, BookNLP identifies 'the old woman' as a character and inferences its gender as female based on the pronoun 'her'. Then, it annotates the connected verbs ('said', 'shaking'), noun ('head'), and modifiers ('old', 'slowly’). Finally, it detects the phrase ‘Ah! it's like another world!’ as the dialogue spoken by 'old woman'.

By running the bookNLP pipeline on each children’s book in the corpus, I annotated the book characters with their gender, the lines of dialogue associated with them, and any words used in characterization. "Words used in characterization," as defined by Underwood et al. (2018), includes verbs that a character governed, nouns they possessed, as well as adjectives that modify characters. Then, in order to compare the extent to which gender is represented as a book character, I calculated the sum of (1) the number of characters, (2) the average number of lines of dialogue, and (3) the number of words used in characterization, according to the gender of book characters and authors.

Result

Figure 2 shows the number and percentage of male and female characters according to the author’s gender. It clearly shows that male characters (N=10187; 38.1%) appear much more than female characters (N=4097; 15.3%) in total. Specifically, in books written by male and female writers, there are more male characters than female characters. There are 4305 male characters (39.3%, Mean = 168), 1756 female characters (11.7%) in books by male writers; there are 5882 male characters (36.5%), 2341 female characters (19.9%) in books by female writers.

Then, I divide the number of lines of dialogue spoken by each character by the number of characters in each book in order to calculate the average number of lines of dialogue spoken by characters of each gender. In Figure 3, we can see that male characters talk more in books written by male authors, while female characters talk more in books written by female authors.

In books by male writers, each male character speaks 6.07 (SD = 3.20) times and each female character speaks 3.97 times (SD = 7.38) on average (p = 0.26); in books by female writers, each male character speaks 4.53 times (SD = 2.53) and each female character speaks 6.69 times (SD = 6.40) on average (p = 0.02).

Lastly, I plot the number and percentage of words used in characterization that describe each gender in Figure 4. Both male and female authors use more words to describe male characters than female characters. In total, 295,555 words (60.2%) are used to describe male characters but only 124,295 words (25.3%) are used to describe female characters. In male-authored books, 11,759 words (49.7%) are used for male characters and 91643 words (38.8%) are used for female characters; in female-authored' books, 178,296 words (70.0%) are used for male characters and 32,652 words (12.8%) are used for female characters.

These results will be discussed in more detail in the following analysis part, but it is clear that these descriptive statistics already begin to show gender imbalances consistent with prior research.

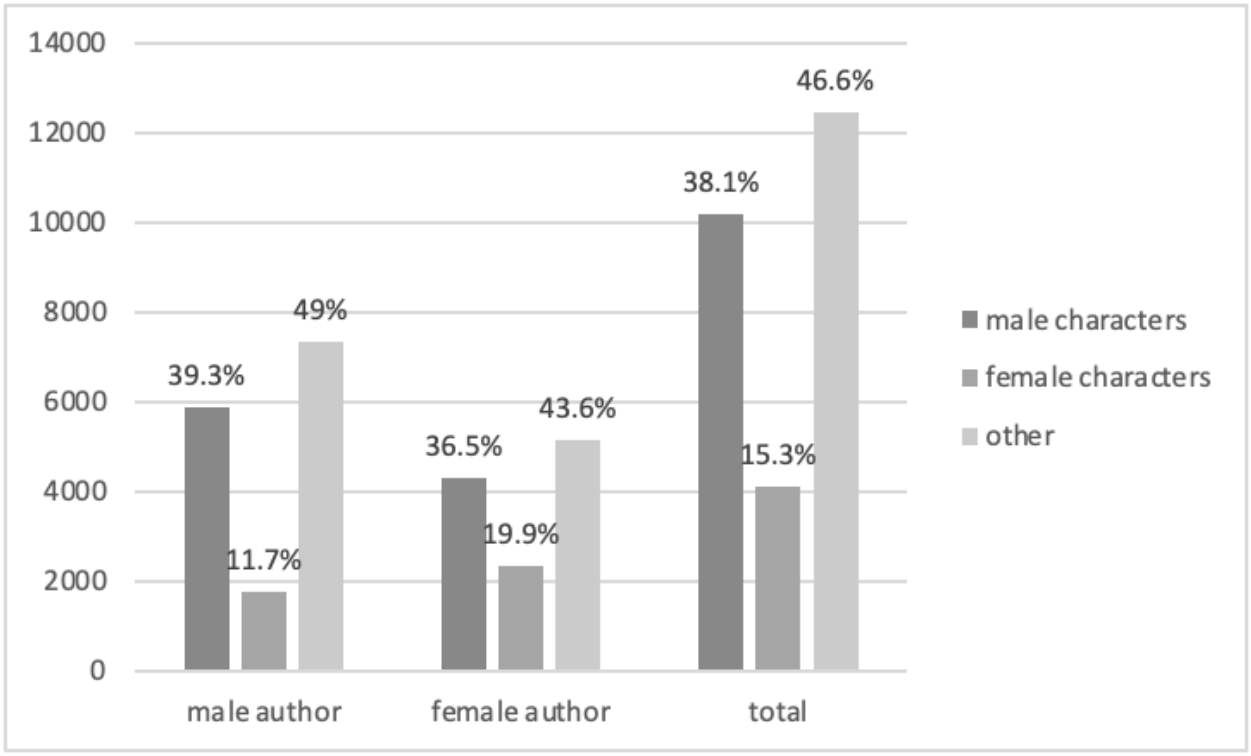

Figure 2. the number and percentage of male characters and female characters by author gender

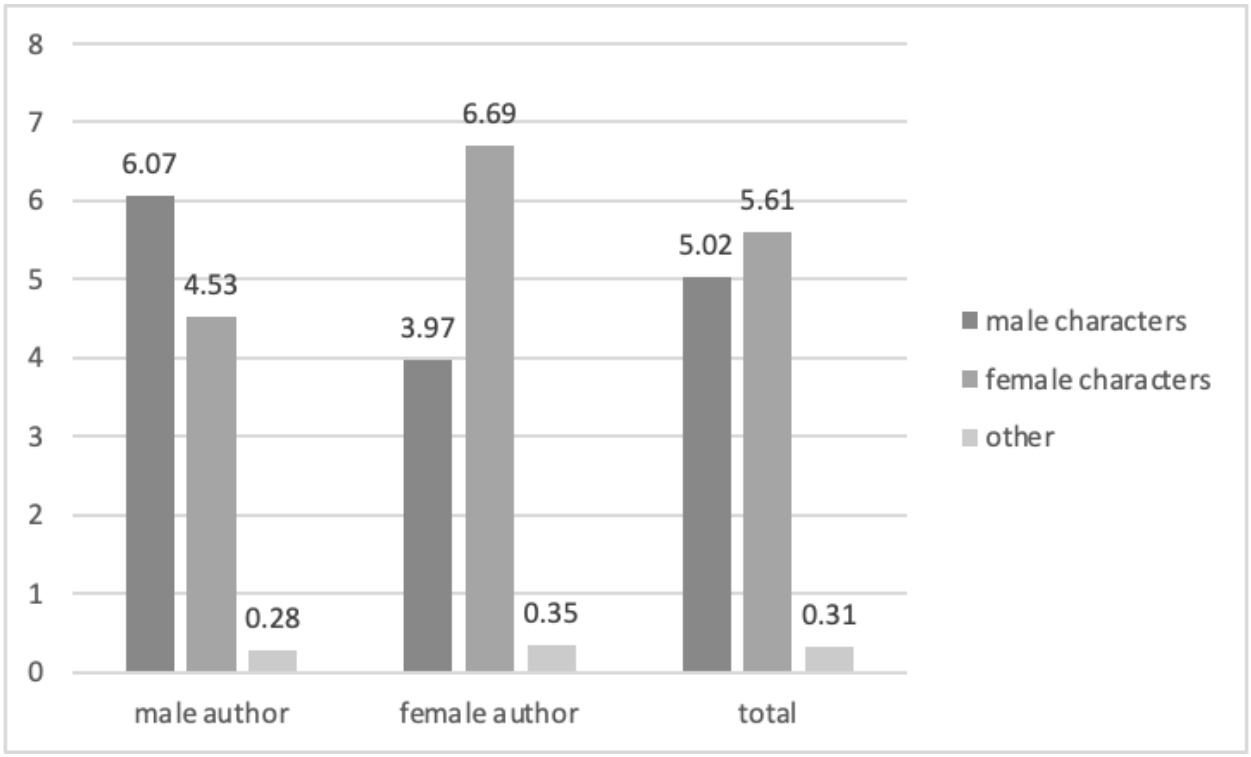

Figure 3. the average number of dialogue of male characters and females by author gender

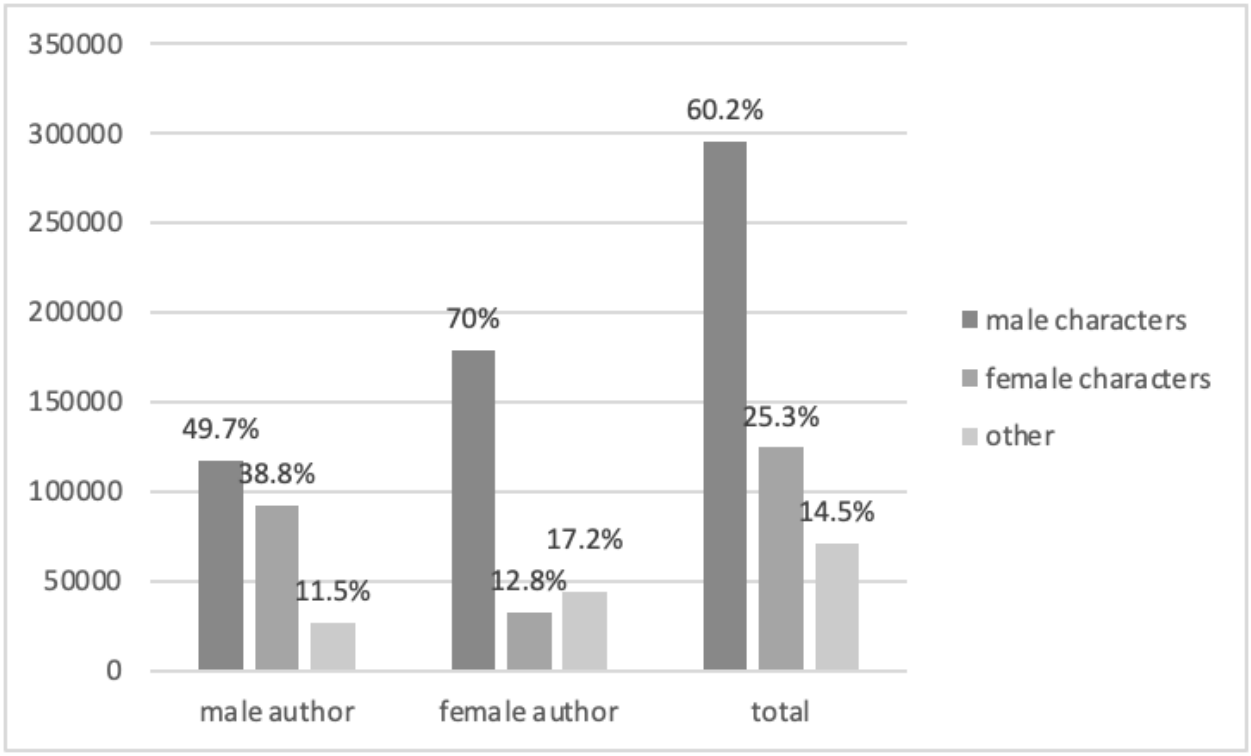

Figure 4. the number and percentage of words used in characterization that describe characters by author gender

Analysis

The results above demonstrate the gender imbalance in 19th century children's literature. In the corpus, male characters outnumber female characters, and male characters dominate in terms of the number of dialogue and characterization words. It supports and extends the findings of other works; most previous works are focused on exploring gender imbalance of children's literature in the 20th and 21st centuries, but this research shows that gender bias is present in earlier time periods as well.

For example, Casey et al. (2021) found that the proportion of male protagonists remains overrepresented in 3,280 children's books published between 1960-2020. An observer study (Ferguson, 2018) found that male characters are given far more speaking parts than female characters in the 100 most popular children's picture books of 2017. The findings, especially Figure 2 and Figure 3, suggest that male characters outnumber female characters in the 19th

century as well. Also, Underwood (2018) found that the proportion of words used in characterization that describe women is lower than 50 percent in the Chicago Novel Corpus published between 1800-2000, which declines from the middle of the 19th-century to the middle of the twentieth. It is consistent with the finding, especially Figure 4.

I also found the differences in gender representation between male and female writers. In Figure 3, I observe that female writers make female characters speak more on average, which may suggest that the gender imbalance is less represented in books by female writers than those by male writers. This pattern is also shown in the findings of Casey et.al (2021); female writers significantly decreased the overrepresentation of male human characters from 1960 to 2020.

4.2 Sentiment Analysis

Method

I use a simple sentiment analysis tool, the SentiArt (Jacobs, 2019), to quantify the sentiment in each line of dialogue spoken by male and female characters. SentiArt computes the affective-aesthetic potential (AAP) score of sentences by using publicly vector space models (VSMs). The English VSM (skip-gram, 500d) is based on the Gutenberg Literary English Corpus (GLEC), containing 250 million words (Jacobs, 2018). SentiArt locates each word of the sentence in a two-dimensional space and computes the semantic relatedness via the cosine similarity. Specifically, SentiArt computes AAP score by the average semantic relatedness between each word in sentences and 120 sentiment labels; there are 60 positive labels (affection, amuse, …, unity) and 60 negative labels (abominable, …, ugly) in total, so the AAP score is the average cosine similarity between each word in sentences and each positive label minus that of each negative label (Jacobs, 2018). Therefore, higher AAP score indicates higher potential for evoking positive affective responses. According to Jacobs (2019), SentiArt outperforms two standard

sentiment analysis tools, VADER (Hutto and Gilbert, 2014) and HU-LIU (Hu and Liu, 2004). I also employed VADER to calculate the sentiment of each line of dialogue in the corpus, but found little of interest. Therefore, in order to best compare the sentiment of lines of dialogue by gender, I computed the AAP score of each line of dialogue extracted by BookNLP. Then, I plotted the average AAP score by the gender of book characters and authors.

Result

Figure 5 shows that female characters speak more positively than male characters on average. In total, the average APP score of male characters is -0.25 (SD = 0.51) and that of the APP score of female characters is -0.23 (SD = 0.53) ($p = 1.75E-14$). Specifically, in books by male writers, the average APP score of male characters is -0.27 (SD = 0.50) and that of female characters is -0.24 (SD = 0.53) ($p = 0.003$); in books by female writers, the average APP score of male characters is -0.23 (SD = 0.52) and that of female characters is -0.21 (SD = 0.53) ($p = 0.004$). This might be because there is a higher ratio of emotional sentences with exclamation marks in the dialogue of female characters, such as “How beautiful!” and “It’s lovely!”. In the dialogues spoken by female characters, there are 1118 out of 27001 sentences (4.14%) that contain the exclamation mark with a positive AAP score; however, in the dialogues spoken by male characters, there are 1908 out of 55406 sentences (3.44%).

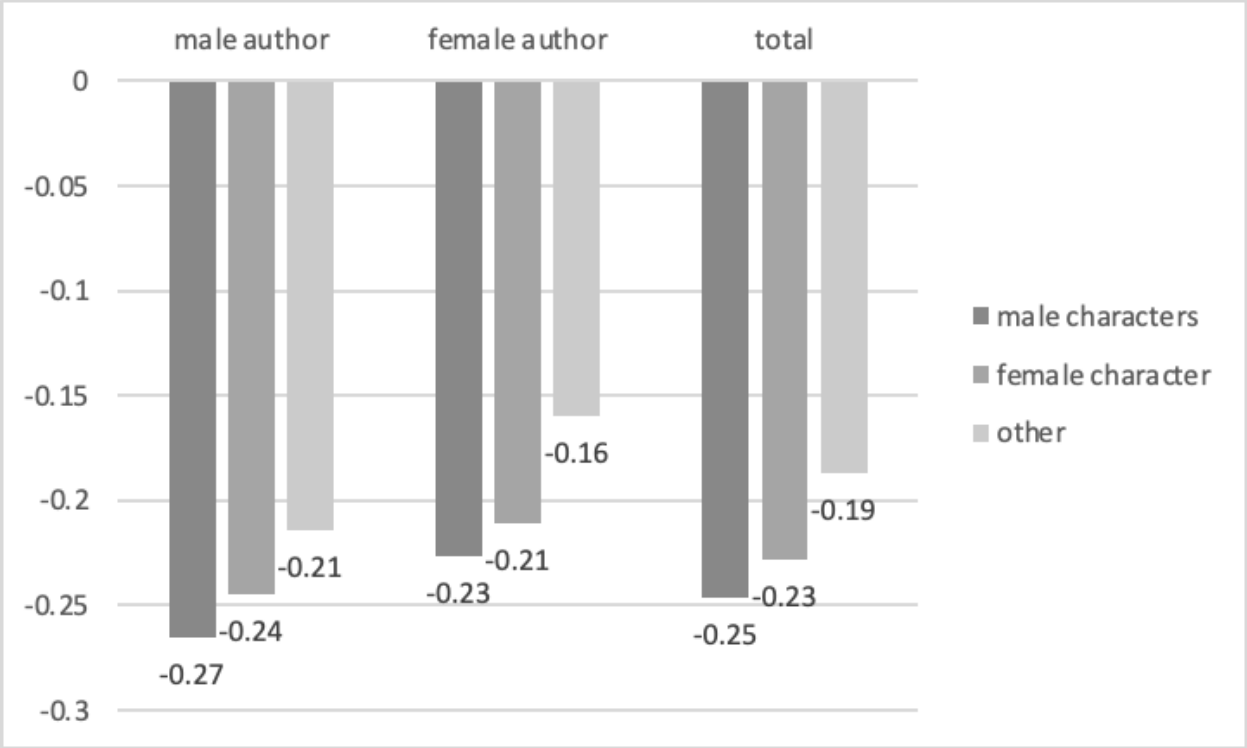

Figure 5. the average APP score of dialogue spoken by book characters

Analysis

The results above suggest that the book characters' way of speaking corresponds to typical gender roles in the 19th century. Looking at Figure 5, female characters speak more positively than male characters on average. It is consistent with the gender roles expected to women in the 19th century in that female characters tend to react or respond without presenting their opinions; in the 19th century, women were instructed to keep quiet and obey their husbands (Welter, 1996).

4.3 Non-contextualized Word Embeddings, Word2Vec

Method

Word embedding models assign high-dimensional vectors to the words in texts, and these vectors in geometry capture semantic relations between the words; words being closer to each other are more similar (Collobert et al., 2011). In much research, word embedding models are used to capture gender biases implied in texts. For example, Garg et.al (2018) examined the changes in gender stereotypes over time using the non-contextualized word embedding model, Word2Vec (Mikolov et al., 2013). They first constructed groups of words ("man" words and "woman" words) representing each gender [Appendix II]. Then, they calculated the relative distance between the group words and a list of neutral words (adjectives and occupations). In order to arrive at the neutral words, Rozado (2020) proposes the usage of larger sentiment lexicons such as the Harvard General Inquirer lexicon (HGI) (Stone et al., 2007) on Word2Vec. The HGI consists of 3,623 words with positive/negative labels. Following the approach described in this work, I trained a Word2Vec model on the raw texts of the corpus and then calculated the distances as measured by cosine similarity between group words(man words, woman words) and the HGI lexicon.

Result

From the 3,624 words in the HGI lexicon, I measured the cosine distance between each group word and 1,132 positive and 1,344 negative words. Then, I calculated the top 10 words in my corpus that were most similar to the "man" words and "woman" words as defined by Garg et al. The words 'hero, loyalty, chum, and vengeance' appear in the top 10 words of "man" words, but not in the top 10"woman" words. This shows that male characters are associated with the traits of strength and adventure. In contrast, 'lover, nurse, beloved, displeasure, and sweetheart' appear in the female group vectors but not in the male group vectors. This shows

that females are associated with emotional and nurturing traits, which is consistent with typical gender roles in the 19th century (Welter, 1996).

Man Group

words	similarity	sentiment
companion	0.4303766	positive
rival	0.4288319	negative
counsel	0.41165955	positive
sincerity	0.40921308	positive
hero	0.40676507	positive
loyalty	0.40610562	positive
friend	0.4054358	positive
chum	0.39936404	positive
friendship	0.3982691	positive
vengeance	0.38858377	negative

Table 1. the top 10 words most similar to the group of men words

Woman Group

words	similarity	sentiment
companion	0.46064833	positive
lover	0.44694729	negative
nurse	0.42629921	positive
counsel	0.42608758	positive
sincerity	0.4137254	positive
beloved	0.39994105	positive
rival	0.39655607	negative
displeasure	0.39374768	negative

friend	0.3930437	positive
sweetheart	0.39109411	positive

Table 2. the top 10 words most similar to the group of women words

Analysis

Looking at Table 1 and Table 2, male and female characters are associated with different words: male words with strong and adventure-related words such as hero, loyalty, chum, and vengeance, and female words with emotional and nurturing related words such as lover, nurse, beloved, and sweetheart.

This is consistent with the gender roles pervasive in the 19th century. As previously discussed, women were expected to be pious, pure, submissive, and domestic. Accordingly, their sphere of influence is limited to home and family. In contrast, men belonged to the public sphere because they were expected to be powerful, logical, and independent. This seems to carry over into the children's literature of the time as well.

4.4 Contextualized Word Embeddings, BERT

Method

Without gender bias, one would expect little difference between the dialogue spoken by men and women. It would follow, then, that a machine learning model would not be able to pick up on statistical differences in gendered language because such differences would not exist (Dinan et al., 2020). In order to test this hypothesis, I fine-tuned a classifier based on the DistilBERT model (Sanh et al., 2019) using the lines of dialogue labeled by the gender of book characters.

BERT (Devlin et al., 2019) is a multi-layer bidirectional transformer encoder that utilizes the self-attention mechanism. As its name would suggest, BERT consists of the layers of transformer blocks, the hidden state, and the self-attention heads, which helps to incorporate context from both directions (forwards and backwards). To implement the BERT model, two steps are required: pre-training and fine-tuning. Because DistilBERT is a general-purpose pre-trained version of BERT, we only need to pre-train the model for classification tasks. According to Sanh et al. (2019), DistilBERT is 40% smaller but 60% faster. It also retains 97% of its language understanding capabilities.

To fine-tune the DistilBERT model, I split the quotes spoken by the character's gender into training (N=59752; 38820 for male characters, 18872 for female characters, and 2060 for others) and test sets (N=25608; 16586 for male characters, 8129 for female characters, and 893 for others), respectively. Then, I fine-tuned the model with the training set and examined the performance of the model through a confusion matrix.

Result

Table 3 shows the performance of the DistilBERT classifier that is fine-tuned with the lines of dialogue of book characters in the corpus; the classifier works well in classifying the dialogue spoken by male characters (sensitivity: 83.3%), while it fails to classify those spoken female characters (specificity: 49.2%). It suggests the way that male characters speak is more distinct.

Then, I computed the top five quotes by each gender, which are correctly classified with the highest probabilities; Table 4 shows the quotes of male characters, and Table 6 shows those of female characters. Looking at Table 4, the first and fourth quotes are from Alone in London, the second one is from Vice Versa or A Lesson to Fathers, the third one is from Kidnapped, and the last one is from King Solomon's Mines. All of the five quotes clearly express the speakers' thoughts; all the speakers evaluate other characters and convey their opinions.

Next, looking at Table 5, the first and fifth quotes are from A World of Girls: The Story of a School, the second and third ones are from Leila at Home. A continuation of Leila in England, and the fourth one is from The Carved Lions. The five quotes are about the speakers' family or their emotions. Specifically, the third and fourth quotes contain information about the speaker's family; in the third quote, the speaker describes her conversation with her family, and in the fourth quote, the speaker talks about her mother and brother. In the other quotes, the speakers talk about their feelings about the listeners. Also, all of the speakers in the five quotes call the names of listers, which may suggest that their relationship is close to each other.

	Male Character	Female Character	Other
Male Character	13826	4118	35
Female Character	2725	3996	15
Other	35	227	68

Table 3. The confusion matrix of the fine-tuned DistilBERT classifier

Male Characters

3.27615595	cause the old master grows worse and worse for forgetting , and I must mind shop for him now as well as I can . He 's not off his head , as you may say ; he 's sharp enough sometimes ; but there 's no trusting to him being sharp always . He talks to Dolly as if she was here , and could hear him , till I ca n't hardly bear it . But I 'm very fond of him,--fonder of him than anythink else , ' cept my little Dolly ; and I 've made up my mind as his Master shall be my master , and he 's always ready to tell me all he knows about him . I 'm no ways afeared of not getting along .
3.26610494	do n't be in such a nurry now . You tell me what you want to know straightforward , and I do n't mean to say as I wo n't help you so far as I can . Do n't be afraid of my telling no tales . I 've bin a schoolboy myself in my time , bless your ' art . I should n't wonder now if I could n't make a pretty good guess without telling at what you 're after . You 've bin a catchin ' of it hot , and you want to make a clean bolt of it . I ai n't very far off , now , am I ?

3.26260495	but the point is , Would I go ? Now I will tell you what I am thinking . I am thinking that it is here upon this doorstep that we must confer upon this business ; and it shall be here or nowhere at all whatever ; for I would have you to understand that I am as stiffnecked as yoursel ’ , and a gentleman of better family .
3.25905442	she 's very ill , and you could cure her here , and take better care of her than Tony and me , and I thought that was enough . I never thought of getting any recommendation , and I do n't know where I could get one .
3.2562809	thou hast fought enough , and if aught befell thee at his hands it would cut my heart in twain .

Table 4. the top 5 correctly classified quotes of male characters with the highest probabilities

Female Characters

3.43617988	I 'm not going to be afraid of you . I have no more silver to give you . If you like , you may go up to the house and tell what you have seen . I am very unhappy , and whether you tell or not can make very little difference to me now . Good - night ; I am not the least afraid of you -- you can do just as you please about telling Mrs Willis .
3.43348098	Now there is a dear one ; you will be good , you will be patient , and say it is all quite , quite right . You know , Matilda it must be so .
3.42865682	Because when I went into the drawing - room with one of the parcels , (mamma ’s blotting - book , you know ,) I heard Uncle Howard say , ‘ Yes , I certainly do see the advantage of having a governess ; _ but_---- ’ and then I put down the parcel very slowly , that I might hear more ; but mamma said , ‘ Matilda , do n’t linger in the room , for we are engaged at present , and wish to be alone . ’ So , you know , I was obliged to be off very quick ; do you think you will like it Selina?--to be sure , it wo n’t be so bad for you , but it will be bad enough for poor me , with all my scrapes ; and yet I should like to see what sort of a face she has got , though I am quite sure I shall not like it .
3.42553449	It was n't only mamma 's going away ; I know Haddie -- that 's my brother -- loves her as much as I do , but he 's not very unhappy , because he likes his school . Oh , Myra , what _ shall _ I do when I have to go back to school ? I 'd rather be ill always . Do you think I 'll have to go back to - morrow ?
3.41572952	Susan , you are looking pale and cold , walk up and down that path half - a - dozen times , and then go into the house . Phyllis and Nora , you can come with me as far as the lodge . I want to take a message from Mrs Willis to Mary Martin about the fowl for to - morrow 's dinner . "

Table 5. the top 5 correctly classified quotes of female characters with the highest probabilities

Analysis

The results above also demonstrate that characters in 19th century children's literature adhere to the typical gender roles of the time. Table 3 shows the dialogue of male characters are classified relatively accurately, unlike that of female characters. It may suggest the way that male characters speak is more distinct. Looking at the instances of correctly classified male dialogue (Table 3), we can see that the male characters evaluate and present their opinions clearly. It means that there is more information that can be used to infer who is the speaker in the dialogue of male characters. In contrast, the female characters convey information about their family or emotions about the listeners.

Chapter 5: Conclusion

In this paper, I explored the representation of gender in 70 children's books published in the 19th century to see whether gender stereotypes are reflected in the books. Overall, the findings of this project confirm the existence of gender bias in the books; there are differences in the representation of male characters and female characters; we observe that the male characters appear more than female characters; and we observe typical (and stereotypical) gender roles in the corpus that are consistent with the attitudes of the time. This project provides a quantitative analysis of how gender is represented in the 19th century children's literature and reveals the gender imbalance pervasive during the time period.It shows that gender bias is present in children's literature published in earlier periods than the 20th century, and also it suggests that the expected gender roles in society are reflected well in children's literature

In this project, I only focused on the representation of male and female characters even though there are many book characters inferred as 'other' gender. I would like to analyze 'other' gender characters as well, with close reading in-depth. Also, I would like to analyze the trends

of gender bias in children's literature by increasing the number of books in the corpus and expanding their publication years.

References

Bamman, D., Underwood T., & Smith, N. (2014). A Bayesian Mixed Effects Model of Literary Character. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Casey, K., Novick, K., & Lourenco, S. F. (2021). Sixty years of gender representation in children's books: Conditions associated with overrepresentation of male versus female protagonists. PloS one, 16(12), e0260566. https://doi.org/10.1371/journal.pone.0260566

Čermáková, A. (2017). The GLARE 19th Century Children's Literature Corpus in CLiC [Blog post]. Retrieved from https://blog.bham.ac.uk/clic-dickens/2017/11/28/the-glare-19th-century-childrens-literature-corpus-in-clic/

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. Psychological Science, 32(2), 218–240. https://doi.org/10.1177/0956797620963619

Collobert, R., Weston, J., Bottou, Lé., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res., 999888, 2493--2537.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams A. (2020). Multi-Dimensional Gender Bias Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 314–331, Online. Association for Computational Linguistics.

Ferguson, D. (2018). Must monsters always be male? Huge gender bias revealed in children's books. The Observer.

Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS 201720347 doi:10.1073/pnas.1720347115

Hu, M., and Liu, B. (2004). "Mining and summarizing customer reviews," in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, eds W. Kim and R. Kohavi (Washington, DC: ACM Press), 168–177. doi: 10.1145/1014052.1014073

Hunt, P. (2001). Children's Literature: An Anthology, 1801–1902. Malden: Blackwell Publishers.

Hutto, C. J., and Gilbert, E. E. (2014). "VADER: a parsimonious rule-based model for sentiment analysis of social media rext," in Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.

Jacobs, A. M. (2018). The gutenberg english poetry corpus: exemplary quantitative narrative analyses. Front. Digit. Humanit. 5:5. doi: 10.3389/fdigh.2018.00005

Jacobs, A. M. (2019). Sentiment analysis for words and fiction characters from the perspective of computational (Neuro-)Poetics. Front. Robot. 6:53. doi: 10.3389/ frobt.2019.00053

Jacobs, A. M., and Kinder, A. (2019). Computing the affective-aesthetic potential of literary texts. Artif. Intell. 1, 11–27. doi: 10.3390/ai1010002

Lewis, M., Cooper Borkenhagen, M., Converse, E., Lupyan, G., & Seidenberg, M. S. (2022). What Might Books Be Teaching Young Children About Gender? Psychological Science, 33(1), 33–47. https://doi.org/10.1177/09567976211024643

Mendoza, J., & Reese, D. (2001). Examining Multicultural Picture Books for the Early Childhood Classroom: Possibilities and Pitfalls. Early Childhood Research & Practice, 3, 1-38.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv: 1301.3781.

Rozado, D. (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. PLoS One, 15(4), e0231189. https://doi.org/10.1371/journal.pone.0231189

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.2.pdf

Stone, P., Bales, R., & Namenwirth, J. & Ogilvie, Daniel. (2007). The General Inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. Behavioral Science. 7. 484 - 498. 10.1002/bs.3830070412.

Underwood, T., Bamman, D., & Lee, S. (2018). The Transformation of Gender in English-Language Fiction. Journal of Cultural Analytics. doi.org/10.22148/16.019

Welter, B. (1966). The Cult of True Womanhood: 1820-1860. American Quarterly, The Johns Hopkins University Press, www.csun.edu/~sa54649/355/Womanhood.pdf. Vol. 18, No. 2, Part 1, pp. 151-174

Appendix

I. Corpus List

Author	Gender	Title	Published
Meade, L. T.	Female	A World of Girls: The Story of a School	1886
Farrow, G. E.	Male	Adventures in Wallypug-Land	1898
Carroll, L.	Male	Alice's Adventures in Wonderland	1865
Haggard, H. R.	Male	Allan Quatermain	1887
Stretton, H.	Female	Alone In London	1869
MacDonald, G.	Male	At the Back of the North Wind	1871
Sewell, A.	Female	Black Beauty. The Autobiography of a Horse	1877
Grahame, K.	Male	Dream Days	1898
Farrar, F. W.	Male	Eric, Or, Little by Little, A Tale of Roslyn School	1858
Martineau, H.	Female	Feats on the Fiord	1841
Nesbit, E.	Female	Five Children and It	1906
Sinclair, C.	Female	Holiday House: A Series of Tales	1839
Ewing, J. H.	Female	Jackanapes	1883
Stretton, H.	Female	Jessica's First Prayer — Jessica's Mother	1867
Stevenson, R. L.	Male	Kidnapped	1886
Haggard, H. R.	Male	King Solomon's Mines	1885
Tytler, A. F.	Female	Leila at Home. A continuation of Leila in England	1870
Stretton, H.	Female	Little Meg's Children	1868
Kingsley, C.	Male	Madam How and Lady Why. Or, First Lessons in Earth Lore for Children	1870

Marryat, F.	Male	Masterman Ready. The Wreck of the “Pacific”	1841
Falkner, J. M.	Male	Moonfleet	1898
Ingelow, J.	Female	Mopsa the Fairy	1869
Ewing, J. H.	Female	Mrs. Overtheway’s Remembrances	1869
Nesbit, E.	Female	Nine Unlikely Tales	1901
Barrie, J. M.	Male	Peter and Wendy (Peter Pan)	1911
Lang, A.	Male	Prince Prigio. From “His Own Fairy Book”	1889
Nesbit, E.	Female	The Book of Dragons	1899
Anstey, F.	Male	The Brass Bottle	1900
Mrs. Molesworth	Female	The Carved Lions	1895
Marryat, F.	Male	The Children of the New Forest	1847
Ballantyne, R. M.	Male	The Coral Island. A Tale of the Pacific Ocean	1858
Martineau, H.	Female	The Crofton Boys	1841
Mrs. Molesworth	Female	The Cuckoo Clock	1877
Yonge, C. M.	Female	The Daisy Chain, or Aspirations	1856
Yonge, C. M.	Female	The Dove in the Eagle’s Nest	1866
Reed, T. B.	Male	The Fifth Form at Saint Dominic’s: A School Story	1887
Grahame, K.	Male	The Golden Age	1895
Wilde, O.	Male	The Happy Prince, and Other Tales	1888
Yonge, C. M.	Female	The Heir of Redclyffe	1853
Kipling, R.	Male	The Jungle Book	1894
Ruskin, J.	Male	The King of the Golden River; or the Black Brothers: A Legend of Stiria	1841
Yonge, C. M.	Female	The Little Duke: Richard the Fearless	1854

Martineau, H.	Female	The Peasant and the Prince	1841
MacDonald, G.	Male	The Princess and the Goblin	1872
Nesbit, E.	Female	The Railway Children	1905
Strickland, A.	Female	The Rival Crusoes; Or, The Ship Wreck	1826
Thackeray, W. M.	Male	The Rose and the Ring	1854
Burnett, F. H.	Female	The Secret Garden	1911
Martineau, H.	Female	The Settlers at Home	1841
Marryat, F.	Male	The Settlers in Canada	1844
Nesbit, E.	Female	The Story of the Amulet	1906
Nesbit, E.	Female	The Story of the Treasure Seekers	1899
Crockett, S. R.	Male	The Surprising Adventures of Sir Toady Lion With Those of General Napoleon Smith	1897
Potter, B.	Female	The Tale Of Benjamin Bunny	1904
Potter, B.	Female	The Tale of Jemima Puddle-Duck	1908
Potter, B.	Female	The Tale of Peter Rabbit	1902
Potter, B.	Female	The Tale of Squirrel Nutkin	1903
Potter, B.	Female	The Tale of the Flopsy Bunnies	1909
Potter, B.	Female	The Tale of Two Bad Mice	1904
Mrs. Molesworth	Female	The Tapestry Room: A Child's Romance	1879
De La Mare, W.	Male	The Three Mulla-mulgars	1910
Kingsley, C.	Male	The Water-Babies	1863
Grahame, K.	Male	The Wind in the Willows	1908
Carroll, L.	Male	Through the Looking-Glass	1871
Hughes, T.	Male	Tom Brown's Schooldays (By An Old Boy)	1857

Stevenson, R. L.	Male	Treasure Island	1883
Anstey, F.	Male	Vice Versa or A Lesson to Fathers	1882
Henty, G. A.	Male	Winning His Spurs. A Tale of the Crusades	1882
Henty, G. A.	Male	With Clive in India. Or, The Beginnings of an Empire	1884
Jefferies, R.	Male	Wood Magic. A Fable	1881

II. Group Words

a. Men words

he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews

b. Women words

she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, femen, sisters, aunt, aunts, niece, nieces