Gender Roles from Word Embeddings in a Century of Children's Books

Anonymous ACL submission

Abstract

When presenting content to children, educators and parents not only want to know *whether* characters of different backgrounds are represented; they also want to understand *how* these characters are depicted. In this paper, we measure the gender portrayal of central domains of social life as depicted in highly influential children's books using word co-occurrence and word embeddings. We find that females are more likely than males to be associated with words related both to family and appearance, while males are more associated with business-related words. The gender associations with appearance and business have endured over time, whereas family word associations have become more gender-neutral. We make two main contributions: one, we create a word embeddings data set, StoryWords 1.0, of 100 years of award-winning children's literature, and two, we show inequality in the portrayal of gender in this literature, which in turn may convey messages to children about differential roles in society. We include our code and models as supplemental data associated with this manuscript.

1 Introduction

Educators and parents use books to teach children messages about society, conduct, and the world. These messages may be encoded in how different identities are, and are not, represented. If there are systematically different associations between specific identities and particular depictions, such messages can shape how children view the roles of themselves and others in society. In this paper, we apply natural language processing (NLP) tools to analyze the gendered association of different domains of social life (e.g., family, business, appearance) to measure how females and males are portrayed in children's books.[1] We use two methods: word co-occurrence, a frequency-based approach, and word embeddings, a prediction-based approach.[2] These tools can enable deeper understanding of the implicit and explicit messages conveyed to children by the books they read. This awareness can, in turn, also help inform content-selection decisions of educators and caregivers.

Early exposure to messages about gender-specific roles and abilities may influence children's beliefs, academic performance, and career paths (Bian et al., 2017; Rodríguez-Planas and Nollenberger, 2018). Gender representation in children's content has traditionally been measured by manual content analysis, in which one or multiple human beings slowly read through the text of a paper to capture the messages on one or multiple dimensions (Neuendorf, 2016). The key advantage of this approach is that it is able to measure deep meaning in books; the main disadvantages are that it is highly labor-intensive, costly to comprehensively characterize a large body of content, and requires a high degree of fidelity in the management and training of the coders (Krippendorff, 2018).

Advances in computer-driven content analysis began to address these concerns through automation. Early efforts focused on a numerical accounting of words which represented different genders – such as counts of pronouns and the genders of named entities – and these counts were then compared across bodies of text (Krippendorff, 2018; Gentzkow et al., 2019). Simple token counts, however, capture only superficial representation. If a female or male is frequently present but portrayed in a stereotypical or narrow manner, then the mere existence of representation will not only be insufficient but also possibly counterproductive.

In this paper, we use word vectors to measure *how* females and males are depicted, vis-a-vis soci-

[1] We refer to "domains of social life" as "domains" for the remainder of the paper.

[2] Static word embeddings are also referred to as word vectors in the literature. For the remainder of this paper, we refer to these as "word embeddings."

etal roles, in award-winning children's books commonly found in schools and homes over the past century, which complement existing measurement of *whether* they appear. This involves converting high-dimensional measures of the semantic meaning of words in text into one-dimensional measures of gender representation in children's books.

We first create co-occurrence matrices to observe how often gendered and domain words appear in the same sentence. We aggregate pairwise counts within categories and scale these values to create co-occurrence frequencies for each gender and each domain. This allows us to compare how frequently females and males are represented in relation to specific domains, both overall and over time.

We then estimate the word embeddings for associations between females and males and specific societal domains. We use the word embeddings analogue of the Implicit Association Test (IAT) to generate group-to-domain cosine similarity measures between the word embeddings for each group with each domain, allowing us to compare the representation of females and males (Garg et al., 2018). We compare the group-to-domain similarity measures overall and over time.

Co-occurrence shows how frequently group words appear in the same sentence as domain words but not how related group and domain words are. Word embeddings estimate deeper semantic relationships between groups and domains but not necessarily whether the words commonly appear together in the same sentence for example. Put together, they provide a more holistic picture of how gender roles are depicted in children's books.

We find that these books are more likely to emphasize females' role in the family and their appearance as compared to depicting males in relation to their role in business, or at work. This trend attenuates over time for the association with family roles, but is consistent for the associations with appearance and business. Patterns remain similar when using word embeddings or co-occurrence.

We make two primary contributions: One, we apply established NLP tools to a policy-relevant body of text with clear implications for child development and education; specifically, the award-winning children's books we examine (and thus the representations they contain) are among those most commonly found in schools and homes. How different identities are portrayed in these books has the potential to shape children's beliefs about themselves and others, which affects their effort in school, future educational decisions, and later life outcomes. Our work also demonstrates how NLP tools can be used to measure the deep meanings contained in bodies of text being considered for use in curricular settings. This has immediate applications for both the practice of education and for research on the linkages between the content of books and the educational outcomes of children exposed to them. Two, we release a word embeddings data set from children's literature (named the StoryWords 1.0 data set) so that other researchers can access these data.

2 Background

External stimuli may have important influences in shaping beliefs, actions, and outcomes (Bian et al., 2017; Bordalo et al., 2017; Rodríguez-Planas and Nollenberger, 2018). For example, historical analysis of changes in textbooks using a quasi-experimental framework has shown that such changes shape both people's preferences and their view of history (Fuchs-Schündeln and Masella, 2016; Cantoni et al., 2017). Less is known about the representation of identities in the content in these books and how these identities are depicted.

Recent work has attempted to address this question by estimating the frequency of female and male presence in stories. Research enumerating gender counts in children's books shows inequality in the frequency of presence of females relative to males over time regardless of the measure, for example, gendered pronouns as compared with gender of characters (Adukia et al., 2021). While these findings are illustrative, they show only superficial representations and neglect to demonstrate whether the trend towards numeric equality is inclusive or rather one of an increased incidence of imbalanced representations. If the frequency of inclusion of underrepresented identities increases without a change in the underlying equity in the manner of representation, simple frequency-based measures might overstate the equity of representation in books that children are given.

In this paper, we address this gap by measuring how females and males are associated with different domains in the text of children's books. We show how NLP tools can help isolate messages in content, converting high-dimensional concepts into one-dimensional parameters of the messages related to gender.

3 Data

3.1 Primary Data: Children's Books

School libraries and classrooms serve as major purveyors of sanctioned visual content for children. The books they offer are accompanied by an implicit state-sanctioned stamp-of-approval. These books are chosen because their content is perceived to be appropriate for children. They are often intended to transmit clear narratives about appropriate conduct, an account of important historical moments, or other, often identity-specific messages.

We draw from a set of children's books that are likely to be found in school libraries – namely, those that received awards administered or featured by the Association for Library Service to Children, a division of the American Library Association. Each book in our sample of 1,130 books is associated with one of 19 different awards.

In order to understand whether representation differs depending on the focus of efforts to highlight different kinds of books, we divide these award-winning corpora into two "collections": the "Mainstream" collection and the "Diversity" collection. Figure 1 shows the sample size of each collection by decade and overall.

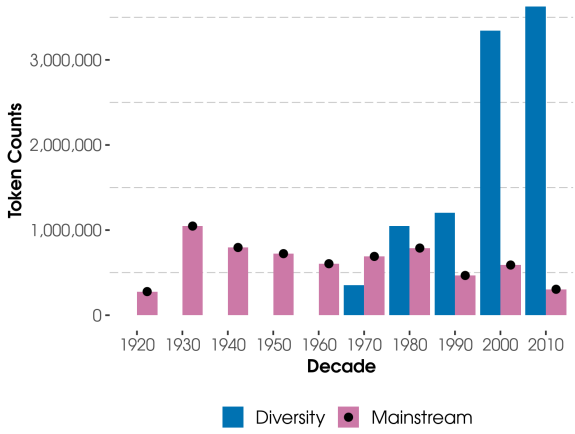


Figure 1: Sample size of the Mainstream and Diversity collections over time. The aggregate number of words in the Mainstream collection is 6,289,116 words and for Diversity is 9,599,638 words.

Mainstream Collection. The Mainstream collection comprises books that have received recognition through the Newbery or Caldecott Medals, the two oldest children's book awards in the United States starting in the 1920s to present day. These books are selected for their perceived literary value and not popularity. Receipt of the award facilitates the book's entry into the canon of children's literature (Smith, 2013; Koss et al., 2018).

Diversity Collection. To examine how purposeful efforts to highlight typically excluded or marginalized identities perform, we draw from books likely to be placed on "diversity lists" such as during Black History Month or Women's History Month. Specifically, we examine books that have received recognition from the following awards: American Indian Youth Literature, Américas, Arab American, Asian/Pacific American Award for Literature, Carter G. Woodson, Coretta Scott King, Dolly Gray, Ezra Jack Keats, Middle East, Notable Books for a Global Society, Pura Belpré, Rise Feminist, Schneider Family, Skipping Stones Honor, South Asia, Stonewall, and Tomás Rivera Mexican American Book Awards. Awards in this collection were first distributed in 1970, with a gradual rollout of different awards over the following decades.

We might expect that books recognized to center one underrepresented identity may also center other underrepresented identities. We can compare the estimates for the Mainstream and Diversity collections to examine whether intentional efforts to highlight underrepresented identities more equitably portray females and males compared to unintentional, "general" efforts.

We provide word embeddings for these collections as supplemental data (StoryWords 1.0).

3.2 Data Pre-Processing

We use Google Vision Optical Character Recognition to extract text from scanned pages of each children's book.[3] Once the text is extracted, we pre-process the data to reduce variability and noise. We first divide each award corpus into sentences using the pre-trained Punkt tokenizer from Python's NLTK library (Bird et al., 2009). For each sentence, we lowercase the text and remove digits, line breaks, punctuation, and special characters. We refrain from removing "stop words" – words that appear frequently and do not contribute to the content of the story – because the learning process does not benefit from their removal (Qiao et al., 2019).[4]

Our goal is to characterize how females and males are represented in each collection of books,

[3]This process is restricted to the conversion of scanned text into ASCII characters.

[4]We check the sensitivity of our results to the inclusion or exclusion of stop words prior to the learning process and find that our results remain similar.

both overall and by decade. We therefore combine the data at two levels: (1) at the collection level, in order to measure overall representations between each of the collections, and (2) at the collection-by-decade level, to measure changes over time.

3.3 Supplemental Data

HistWords. In addition to the children's books, we incorporate data from the HistWords data set, a collection of books gathered from over 40 university libraries containing more than 361 billion English words (Michel et al., 2011). These books span from 1800 to 2000 and are composed of a variety of genres.[5] We include these data as a numeraire, capturing the representations of females and males in books intended for adult consumption, rather than children's consumption. Because the only publicly available data for HistWords is in the form of word2vec embeddings, we directly incorporate the embeddings they provide in our final visualizations rather than running the lexicon through our pipeline, as outlined in Section 4.[6]

Group and Domain Words. We develop a vocabulary of words that comprise two gender groups (females, males) and three domains (appearance, family, business).[7,] The words associated with females, males, appearance, and family were generated by drawing upon commonly used words for each category, in addition to incorporating words from sources such as those lists given by Caliskan et al. (2017) and Senel et al. (2018). We fine-tune the categories to the linguistic particularities of the domain of children's literature by incorporating vocabulary that is commonly used in these books. For example, words such as "princess" and "king" are included our gender group word lists, but are not in prior group word lists, such as those in Caliskan et al. (2017) and Garg et al. (2018). [8] The final sizes of our lists are as follows. female: 77 words, male: 75 words, appearance: 154 words, family: 29 words, and business: 221 words and each word within a given category is exclusive to that category only. Specifically, the family category is notably smaller than other lists because many "family" words are gendered and therefore were included in the male or female lists instead of the "family" list. These word lists are available in the supplemental data associated with this manuscript.

4 Methods

We use two NLP tools to characterize gender representation: co-occurrence and word embeddings. We release all code and models as supplemental data associated with this manuscript.

4.1 Word Co-occurrence

We first measure how females and males are represented in the text of these stories by estimating word co-occurrence. Co-occurrence is a simple and easily interpretable tool that looks at how frequently pairs of words co-occur in a set context window which, in our study, is a sentence.[9,10]

Because we are interested in the relationship between genders and specific domains, we focus on co-occurrence of pairwise words within these categories. We created a matrix X with gender words as columns and domain words as rows. Each value (denoted as $X_{(i,j)}$, where i = domain words, j = gender words) in X contains the number of times each word pair appeared within the same sentence in that collection, both overall and by decade. To understand what this matrix shows about co-occurrence of each gender and domains, we reduce the dimensions and aggregate counts by group. This results in a matrix Y with two columns (female, male) and three rows (appearance, family, business) containing summed pairwise counts. For example, the $Y_{(bus,fem)}$ count is calculated as $\sum_{bus} \sum_{fem} X_{(bus,fem)}$ where bus represents each business word and fem represents each female word. The issue with comparing how females and males are represented in relation to domains using raw counts is that there are more instances of male words than there are female words, and therefore the gendered counts cannot be easily compared. To account for this, we divide each value in the female column by the number of sentences that

[5] We limit analysis of HistWords starting in the 1920s; the first book in the children's collections was published in the 1920s, and the last book in HistWords is from 2000.

[6] The aggregate model for the HistWords collection is not publicly available. We discuss how we estimate HistWords collection-level measures for word embeddings in Section 4.2.

[7] Note that this analysis is limited to the binarization of gender; analysis of other gender identities represents an important area for future work.

[8] Our choice of gendered vocabulary is over 3 times as large as the gendered word lists used in Garg et al. (2018), who use 20 male words and 20 female words, and approximately 9 times larger than the gendered word lists in Caliskan et al. (2017), who use 8 male words and 8 female words.

[9] Co-occurrence is categorized as a frequency-based embedding because it examines raw counts.

[10] We also test co-occurrence using context windows between three and six words and found similar results.

have a female word, and each value in the male column by the number of sentences that have a male word. By transforming counts to frequencies, we can accurately compare how often females and males co-occur with specific domains. We define "gender-skew" as female frequency minus male frequency for a given domain and collection, calculating gender-skews for each domain in each collection, both aggregate and by decade.

We then examine whether the difference between female and male frequencies for a domain is significant. To do this, we test the hypothesis $H_0 : \mu_d = 0$ versus $H_A : \mu_d \neq 0$ using a paired two-sample t-test where μ_d is the population mean of the differences in weighted frequencies for female versus male over all words in a domain for a given collection. We test this hypothesis in the appearance, family, and business domains to gain an understanding of whether the co-occurrence gender frequencies in these domains are significantly different.

4.2 Word Embeddings

Another way to capture how gender is represented in text is through word embeddings. Word embeddings operate under the assumption that words which appear in similar contexts have similar meanings (Firth, 1951). In practice, word embeddings are neural networks that map each word to a high-dimensional vector representation of that word. Each word vector encapsulates semantic and syntactic information by incorporating information from the nearest neighbors (context) of that word. Word embeddings permit analysis between sets of vectors, including calculating similarity measurements between words using cosine distance.[11]

We use word2vec from Python's Gensim library to estimate word embeddings (Řehůřek and Sojka).[12] Our word2vec implementation uses the Skip-gram model architecture for training, which uses a given word to predict context words (words that appear within a certain window of the current word) in a sentence.[13] During the training process, the model learns the word vector representation of each word in the set of vocabulary contained in a given text.[14] After training, the algorithm outputs 300-dimensional vectors of every word in the lexicon of each book.[15] We train separate word2vec models on the aggregate collection data as well as on the collection-by-decade data discussed in Section 3.2.[16] We name the resulting data set of word embeddings StoryWords 1.0.

We then apply a textual analysis variant of the Implicit Association Test (IAT), a method used to detect bias in speech (Caliskan et al., 2017). The word embeddings analogue of the IAT involves taking the estimated vectors for words belonging to a given group (e.g., "she" and "queen" belong to the group "female") and examines their relationship with a given domain (e.g., "hair" and "shirt" belong to the domain "appearance").

We calculate the pairwise cosine similarities between the vocabulary words of each group and the vocabulary words of each domain. For example, to calculate the association between female-to-appearance, we calculate the cosine similarities between each female word and each appearance word (excluding words that do not appear in the text of the collection).

We then average each set of the pairwise group-to-domain cosine similarities to obtain a single association value. This association describes the extent to which groups (females or males) are associated with a given domain (appearance, family, or business) (Caliskan et al., 2017).

We then estimate a domain-specific parameter of "gender-centeredness" by subtracting the association between a given domain and male from the association between the domain and female. If the value of a domain's gender-centeredness is positive, we classify the domain as more female-centered; if negative, we classify it as male-centered.

We estimate logistic regression models for each domain by collection – in aggregate and by decade – to test whether the female and male cosine similarities are significantly different. Using pairwise cosine similarities for (domain, female) words and (domain, male) words as the predictor variable and

[11] Word embeddings are categorized as prediction-based embeddings because they use machine learning to predict context words.

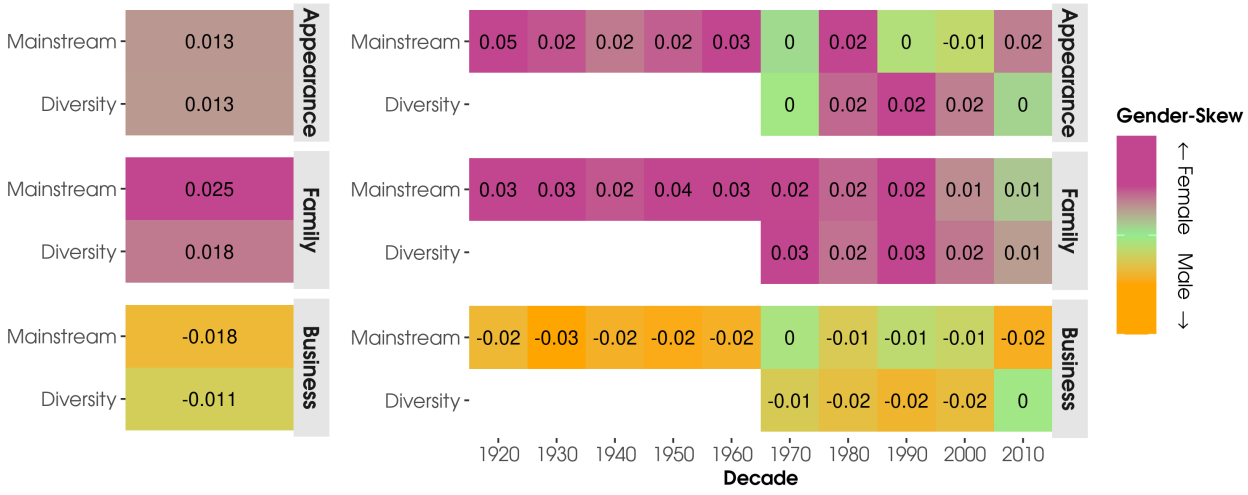
[12] While we show results from the implementation of word2vec, our results are similar when we use GloVe, another commonly used algorithm.

[13] We chose the Skip-gram architecture as it outperforms other architectures on semantic relationship tests and is more accurate on larger data sets in general (Mikolov et al., 2013).

[14] Words that occur fewer than ten times are excluded from the analysis, for these words appear too infrequently to obtain reliable vectors.

[15] We use the word2vec defaults for the remaining parameters and hyperparameters of these models.

[16] Because aggregate measures are not available at the collection level for HistWords, we average the measures for HistWords for each decade starting from the 1920s through the 1990s to estimate an overall measure for this collection and are not able to calculate statistics to generate an overall measure.



(a) Overall

(b) Over Time

Figure 2: Gender-skew calculated from co-occurrence matrices for Mainstream and Diversity collections (a) overall and (b) over time. Magenta indicates a more female-skewed domain (denoted by positive values), green indicates a more neutral domain, and orange indicates a more male-skewed domain (denoted by negative values).

0 (male) and 1 (female) as the predicted variable, we aim to predict gender from cosine similarities. If there is no association between cosine similarities and gender, then the log-odds ratio, β_1, will be 0. We consider the observed difference between female and male cosine similarities to be different from zero if the the test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is significant at the $\alpha = 0.05$ level. This word embedding analysis provides insights in how females and males may be differently represented with respect to different domains.

5 Results: Word Co-occurrence

We first report estimates of gender-skews using co-occurrence matrices for the appearance, family, and business domains. We show how these estimates differ across collections overall and over time.

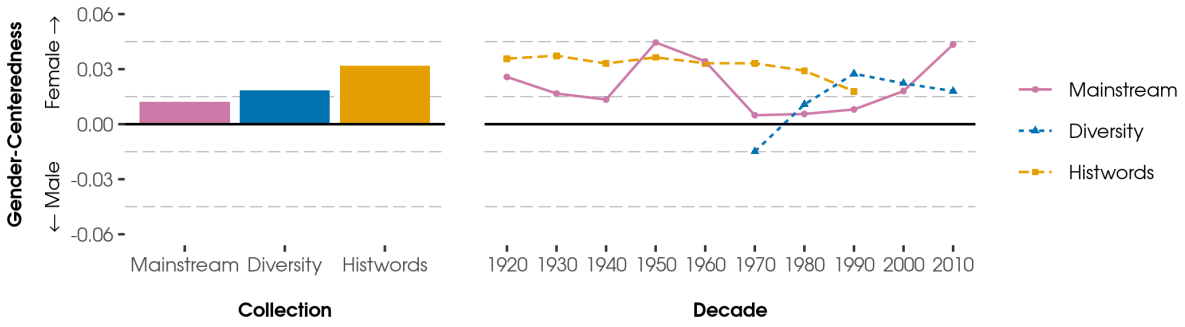
Figure 2 illustrates our findings for the gender-skews of the three domains. Overall, females are generally more likely to appear in the context of their appearance (body parts, clothing-related words) than males, with no apparent difference between the Mainstream and Diversity collections. This female-skew for appearance persists over time. Moreover, females are generally more likely to appear in the context of family than males overall, with a noticeably stronger female association with family in the Mainstream collection than the Diversity collection. In contrast to our estimates of gender-skew in appearance, our estimates of gender-skew in family attenuate over time.

Finally, we see that males are generally more likely to be referenced in the context of business than females, with a noticeably stronger male-skew in the Mainstream collection than the Diversity collection. This male-skewed association with business persists over time. The paired t-tests comparing differences of frequencies for co-occurrence are not statistically significant.

6 Results: Word Embeddings

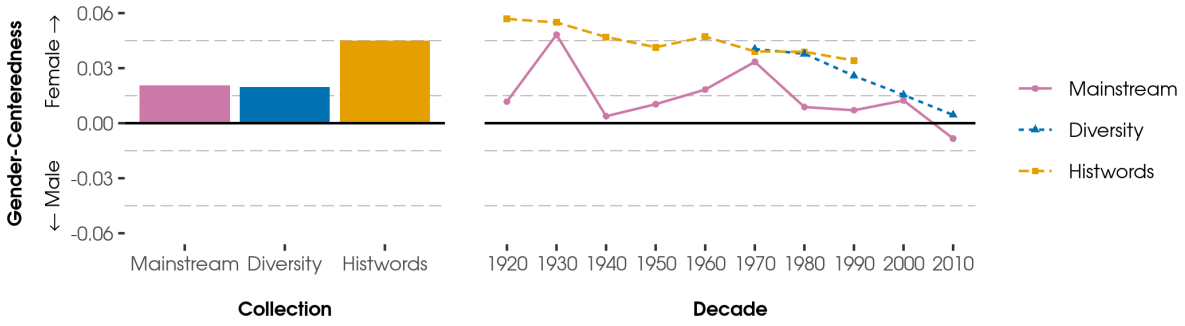
In this section, we report word embedding estimates of the gender-centeredness of each domain across collections overall and over time.

Appearance. We first analyze the gender-centeredness of words in the appearance domain to understand who is more likely to be associated with words related to one's looks. In Figure 3, we show similar trends as found from co-occurrence matrices: females are much more likely than males to be associated with appearance words both overall and over time. All three collections are female-centered in the appearance domain, with the largest difference between females and males exhibited in the HistWords collection. Regression outputs show that the overall female association with appearance is significantly stronger than the male association in both the Mainstream collection ($T = 9.58, p = 9.99E-22$) and Diversity collection ($T = 13.22, p = 6.45E-40$). Only one decade in the Mainstream collection does not show a statistically significant difference between fe-



(a) Overall (b) Over Time

Figure 3: Appearance gender-centeredness (a) overall and (b) by decade for each collection



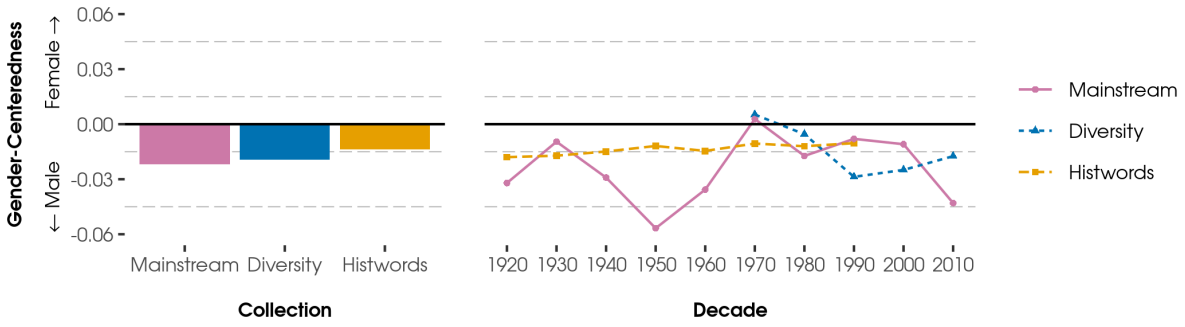
(a) Overall (b) Over Time

Figure 4: Family gender-centeredness (a) overall and (b) by decade for each collection

males and males. All decades in the HistWords and Diversity collections are significant at the $\alpha = 0.05$ level, although the first decade in the Diversity collection indicates a male-centered association.

Family. We next examine gender-centeredness in the family domain. In Figure 4, we show that females are more likely to appear in the context of family than males overall for each collection. The Mainstream and Diversity collections share a similar level of gender-centeredness in their representation of the family domain, though this masks a slightly more gender-equal representation of this domain among the Mainstream collection books in the years in which there are also Diversity collection books (see the measurements from 1970 onwards in Figure 4). While the overall association with family words is about twice as skewed towards females in the Histwords collection than in the children's book collections, the gender-centeredness in this domain appears to be aligned with the Diversity collection for the years which they overlap. Like the family gender-skew results, the family gender-centeredness values attenuate over time, becoming more gender-neutral for all three collections. Logistic regressions show that the female-centeredness estimates observed overall in both the Mainstream ($T = 4.37, p = 1.27E-05$) and Diversity collection ($T = 2.73, p = 0.00641$) are significant. Further, all of the 8 HistWords decades, 3 of 10 Mainstream decades, and 3 of 5 Diversity decades show a significant difference between female and male cosine similarities at the $\alpha = 0.05$ level.

Business. We examine the gender-centeredness of words in the business domain to understand who is more associated with business. We observe similar trends as seen with co-occurrence gender-skews: Figure 5 shows that males are more likely to be associated with business words than females in each collection overall. The male-centeredness for business persists over time in each collection. Models testing the strength of this association have the most significant results across all domains, likely partially due to the large sample size of business words. The associations in Mainstream ($T = -18.04, p = 9.39E-73$) are similar to those in Diversity ($T = -20.47, p = 3.88E-93$), and both have significant differences between female and male cosine similarities. Statistically signifi-



(a) Overall

(b) Over Time

Figure 5: Business gender-centeredness (a) overall and (b) by decade for each collection

cant differences appear in every HistWords decade, 9 of 10 decades in Mainstream, and 4 of 5 decades in Diversity.

7 Conclusion

We make two primary contributions. One, we create a word embeddings data set from children's books, StoryWords 1.0. Two, we analyze how gender roles are portrayed in children's literature using NLP methods that convert high-dimensional data into single-parameter estimates that succinctly describe the relationship between females and males with respect to words that represent appearance, family, and business. We use two tools in particular: word co-occurrence and word embeddings. Both methods find that females are more likely than males to be represented in relation to their appearance and their role in the family, while males are more likely than females to be represented in relation to their roles in business. Only in the family domain do we see an attenuation of the female-centeredness of the representation over time. We find no evidence that the Diversity collection, meant to highlight typically excluded or marginalized identities, portrays females more equitably than the "general" efforts of the Mainstream collection.

It is important to note that the results from co-occurrence and word embeddings are not directly comparable. While co-occurrence analysis shows whether a group and domain word pair appear in the same sentence, word embeddings does not focus on observed counts. Word embeddings infer associations using context windows, and can predict associations between word pairs that may never appear in the same sentence. Moreover, group-to-domain associations may not capture direct relationships; words may appear in the same context but not actually refer to one another. The patterns are consistent regardless of the approach used, which increases our confidence in these results.

Our study has two key limitations. The first is accurately characterizing gender for each instance of a gendered word. We chose not to include proper nouns in our vocabulary lists because of their relative infrequency (and therefore their estimated embeddings would be relatively unstable). Exclusion of these words then means that we cannot estimate the portrayal of names with an identifiable gender. Another important limitation is that our measure of gender representation binarizes gender constraining it as female or male and does not account for non-binary or gender-fluid identities.

Future work includes using more precise tools, such as coreference resolution, to better understand and disentangle the indirect and direct messages contained in these texts. In addition, researchers or practitioners using these tools could expand their analysis to other categories, such as different groups to understand how other identities may be differentially represented, additional domains, or adjectives that convey different societal meanings. Additionally, we can expand definitions of gender to account for non-binary and gender-fluid identities. This work could also account for polysemous words by using contextualized word vectors.

Our paper demonstrates how NLP tools can be used to reveal systematically different associations between females and males and their societal roles, as transmitted through children's stories. These findings underscore the importance of tracking not only whether different identities are included in stories, but also how characters of different backgrounds are portrayed.

References

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2021. What we teach about race and gender: Representation in images and text of children's books. *NBER Working Paper 29123*.

Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. 2017. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323):389–391.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2017. Memory, attention, and choice. *The Quarterly Journal of Economics*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. Publisher: American Association for the Advancement of Science.

Davide Cantoni, Yuyu Chen, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang. 2017. Curriculum and ideology. *Journal of Political Economy*, 125(2):338–392.

John Rupert Firth. 1951. *Papers in Linguistics (1934–1951)*. Oxford University Press, Oxford, UK.

Nicola Fuchs-Schündeln and Paolo Masella. 2016. Long-lasting effects of socialist education. *Review of Economics and Statistics*, 98(3):428–441.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Matthew Gentzkow, Jesse Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Melanie D Koss, Nancy J Johnson, and Miriam Martinez. 2018. Mapping the diversity in caldecott books from 1938 to 2017: The changing topography. *Journal of Children's Literature*, 44(1):4–20.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, , Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv e-prints*.

Kimberly A. Neuendorf. 2016. *The content analysis guidebook*. Sage.

Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the behaviors of bert in ranking. *arXiv preprint:1904.07531*.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora.

Núria Rodríguez-Planas and Natalia Nollenberger. 2018. Let the girls learn! It is not only about math... it's about gender social norms. *Economics of Education Review*, 62:230–253.

Lutfi Kerem Senel, Ihsan Utlu, Veysel Yucesoy, Aykut Koc, and Tolga Cukur. 2018. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Vicky Smith. 2013. The "Caldecott effect". *Children Libraries: The Journal of the Association for Library Service to Children*, 1(1):9–13.