Tales and Tropes: Gender Roles from Word Embeddings in a Century of Children's Books*

Anjali Adukia[1], Patricia Chiril[1], Callista Christ[1], Anjali Das[2]
Alex Eble[2], Emileigh Harrison[1], Hakizumwami Birali Runesha[1]
(1) University of Chicago
{adukia, pchiril, callistac, harrisone, runesha}@uchicago.edu
(2) Columbia University
ad3888@columbia.edu, eble@tc.columbia.edu

Abstract

The manner in which gender is portrayed in materials used to teach children conveys messages about people's roles in society. In this paper, we measure the gendered depiction of central domains of social life in 100 years of highly influential children's books. We make two main contributions: (1) we find that the portrayal of gender in these books reproduces traditional gender norms in society, and (2) we publish StoryWords 1.0, the first word embeddings trained on such a large body of children's literature. We find that, relative to males, females are more likely to be represented in relation to their appearance than in relation to their competence; second, they are more likely to be represented in relation to their role in the family than their role in business. Finally, we find that non-binary or gender-fluid individuals are rarely mentioned. Our analysis advances understanding of the different messages contained in content commonly used to teach children, with immediate applications for practice, policy, and research.

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 depictions, such messages can shape how children view the roles that they themselves, as well as others, can occupy in society. In this paper, we apply Natural Language Processing (NLP) tools to analyze the gendered association of different attributes (e.g., *traits*, *occupations*, *physical characteristics*) to measure how females and males are portrayed in children's books.[1]

To measure portrayal, we use word embeddings, a prediction-based method for analyzing the semantic meaning of words in context using high-dimensional vectors. We supplement our analysis by training a model to detect individual sentences containing stereotypes in order to gain a 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.

Messages about gender-specific abilities and roles may influence children's beliefs and career paths (Leslie et al., 2015; Riley, 2017; Bian et al., 2017, 2018). Gender representation in children's content has traditionally been measured by manual content analysis, in which one or multiple annotators slowly read through the text of written content to classify the messages within 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 disadvantage is that it is highly labor-intensive, making it prohibitively costly to comprehensively characterize representation in large bodies of content, and requires a high degree of fidelity in the management and training of the coders (Krippendorff, 2018). It is difficult to avoid human biases in this type of traditional content analysis, though these biases are of course also baked into any content analysis, including computerized approaches (Buolamwini and Gebru, 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;

*All the authors contributed equally to this work.

[1]As gender is not a binary construct, we wished to include characters who identified as non-binary or gender fluid, in addition to those who identified as female or male. In an analysis of a subset of books which center LGBTQIA+ experiences, we only found two characters with non-binary identities (0.37% of total characters). Because there would not be a sufficient sample size to estimate embeddings for this group, we limit our main analysis to females and males.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3086–3097
October 12–17, 2022.

Gentzkow et al., 2019). Simple token counts, however, primarily capture 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 address this gap by using word vectors to measure *how* different genders are depicted, *vis-a-vis* societal roles, in English-language, award-winning children's books commonly found in schools and homes over the past century, complementing existing measurement of *whether* they appear (Adukia et al., 2022). 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. Our study makes two primary contributions:

(1) We apply established NLP tools to a policy-relevant body of text with clear implications for child development and education in order to understand how roles are differentially portrayed by gender (cf. Section 5). Specifically, we examine the gender associations of societal domains such as *appearance* vs. *competence*, *family* vs. *business*, and *female* vs. *male professions*.

(2) We release a word embeddings dataset trained on our sample of award-winning children's books (named the `StoryWords 1.0` dataset) so that other researchers can use these data (cf. Section 3).[2]

How different identities are portrayed in these books has the potential to shape children's beliefs about themselves as well as their beliefs about 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 messages contained in bodies of text being considered for use in curricular settings. This has clear and immediate applications for both the practice of education and for research on the linkages between the content of books and on the educational outcomes of children exposed to them.

2 Related Work

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 how frequently females are present in the text relative to males over time regardless of the measure, for example, in gendered pronouns and in the gender of named characters (Adukia et al., 2022). 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.

Recent work has addressed how characters of different genders are portrayed. Xu et al. (2019) analyzes female characters' emotional dependency on male characters in a collection of books, movie synopses and movie scripts. That study defines narratives in which a man serves as a woman's path to a happy, fulfilling life as characterized by the '*Cinderella complex*'. Using pretrained word2vec models, they constructed a vector representing the dimension of *happy* vs. *unhappy* that was used for calculating the *'happiness scores'* of words surrounding specific female and male characters. They first selected the movie synopsis of Cinderella; calculating happiness scores for it, the study shows that the happiness of Cinderella depends on the prince, but not vice versa. Further testing on different movie genres showed that the happiness score of the female characters portrayed in the same context as male characters was higher than when the females were portrayed alone. They also find that male characters are more likely to be described using verbs, while female characters are more likely to be described using adjectives.

[2]The data and associated code are available at: `https://github.com/miielab/GenderEmbeddingsPaper`

2.1 Gender in Language Models

Word embeddings have become one of the most used types of features in many NLP models and are widely used for a variety of downstream tasks. However, these word representations have been proven to reflect social biases (such as race and gender) inherited from data used to train them (Caliskan et al., 2017). To automatically quantify these biases, several fairness metrics (i.e., functions that measure the association degree between target and attribute words in a word embedding model) have been proposed in the past few years (Caliskan et al., 2017; Garg et al., 2018; Sweeney and Najafian, 2019; Ethayarajh et al., 2019; Dev and Phillips, 2019). More recently, researchers have started quantifying, analyzing and mitigating the gender bias exhibited by contextualized embeddings (Zhao et al., 2019; Kurita et al., 2019; Tan and Celis, 2019; Guo and Caliskan, 2021). Their results show that contextualized word models inherit human-like biases, which are then propagated to downstream tasks.

2.2 Gender Stereotypes in Social Sciences

Gender stereotypes are defined by the Office of the High Commissioner for Human Rights (OHCHR) as *'a generalised view or preconception about attributes, or characteristics that are or ought to be possessed by women and men or the roles that are or should be performed by men and women'*.[3]

One significant consequence of gender stereotypes is the reinforcement of gender inequality; within this framework, *agency* (i.e., traits such as competence and independence) and *communion* (i.e., concerns about the welfare of others and relationship with them) are the core dimensions used to characterize gender stereotypes. Although biological attributes may impact a person's behaviour and choice of occupational roles, research indicates that gender differences in beliefs about gender stereotypes develop over time, and that they are influenced by family, friends and education (Dhar et al., 2018; Eble and Hu, 2020, 2022). For example, one set of gender stereotypes posits that women are communal, kind and family oriented, whereas men are more agentic, skilled and work oriented (Ellemers, 2018).

In light of changes in the positions occupied by women in society, as well as the broadening of opportunities presented to women, Haines et al. (2016) characterize the extent to which gender stereotypes have changed between 1983 and 2014. In that study, participants assessed the likeliness of a set of gendered characteristics (e.g., *traits*, *behaviours*, *occupations*, *physical characteristics*) to belong to a typical man or woman, similar to the methods used by Deaux and Lewis (1984). The study assessed whether people's beliefs changed over time in parallel with changes in society. They also measured whether these beliefs vary by age, measuring this for people from 19 to 73 years of age, as opposed to the college students studied in Deaux and Lewis (1984). Surprisingly, the authors find no indication of a substantial change in basic stereotypes over time in spite of many relevant societal changes.

Although widely studied in psychology, communication studies and social science (Allport et al., 1954; Crawford et al., 2002; Beike and Sherman, 2014; Biscarrat et al., 2016), in NLP, gender stereotypes have been studied mainly to detect or remove gender bias in word embeddings or word association graphs (Bolukbasi et al., 2016; Park et al., 2018; Madaan et al., 2018; Dev and Phillips, 2019; Du et al., 2019) as well as to identify disparity across gender in various applications like co-reference resolution (Zhao et al., 2018) and sentiment analysis (Felmlee et al., 2019). A notable exception is the work by Chiril et al. (2021) who use gender stereotypes detection as an auxiliary task to improve sexism classification.

3 Data

3.1 Primary Data: Children's Books

School libraries and classrooms serve as major purveyors of sanctioned literary 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 written in English that are likely to be found in U.S. 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. Out of the 3,447 books that either won an award or received an honourable

[3]Source: https://www.ohchr.org/en/women/gender-stereotyping, accessed September 14, 2022.

mention, we were able to collect and digitize a sample of 1,130 books using both library and online resources.

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.

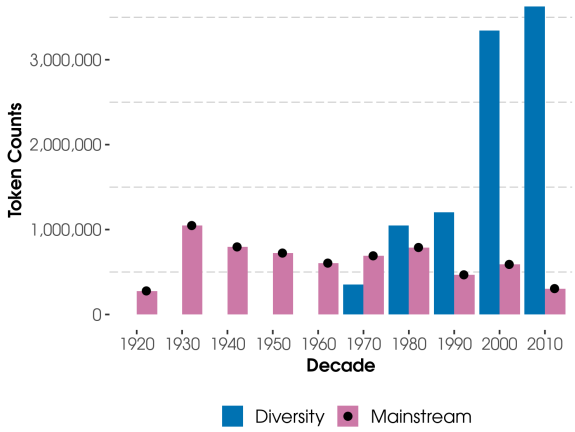

Figure 1: The sample size of the `Mainstream` and `Diversity` collections over time. The aggregate number of words in the `Mainstream` collection is 6,289,116 words and in the `Diversity` collection is 9,599,638 words.

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 contribution to children's literature and not popularity. Receipt of the award facilitates the book's entry into the canon of U.S. children's literature (Smith, 2013; Koss et al., 2018). These books are all in English, but are likely to be translated into other languages.

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.[4] Awards in this collection were first distributed in 1970, with a gradual rollout of different awards over the following decades.

We 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 hypothesize that books which are recognized for highlighting one underrepresented identity may also highlight other underrepresented identities.

3.2 Data Collection and Pre-processing

We use Google Vision Optical Character Recognition (OCR) to extract text from scanned pages of each children's book.[5] Note that this process is restricted to the conversion of scanned text into ASCII characters. A manual error analysis on a random sample of 10 children's books shows that the average Word Error Rate (WER) of the text extracted using OCR was 2.62%. Since our sample of children's books contains many illustrations, most of the error can be attributed to random characters added to the extracted text when the OCR software mistook a shape in a illustration as an ASCII character.

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 then lowercase the text and remove digits, line breaks, punctuation, and special characters.[6]

Our goal is to characterize how females and males have been overall represented in each collection of books, as well as how this representation has changed over time. 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.

[4]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.

[5]Source: https://cloud.google.com/vision/docs/ocr

[6]We refrain from removing stopwords because a preliminary inspection of the sensitivity of our results to the inclusion or exclusion of stopwords prior to the learning process showed that our results remain similar.

3.3 Supplemental Data: HistWords

In addition to the children's books, we incorporate data from the `HistWords` dataset, 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 contain text from of a variety of genres.[7] We include these data as a numeraire, capturing the representations of females and males across the last two centuries 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.[8]

3.4 Gender as a Non-binary or Fluid Construct

Our main goal is to measure the relationship between different gender groups and societal domains. Gender is not a binary construct and can comprise females, males, non-binary, and gender-fluid individuals. However, as far as we are aware, there is no systematic way of measuring non-binary or gender-fluid identities in off-the-shelf NLP packages such as those we use. Instead, to evaluate the presence of non-binary identities in our data, we manually search for non-binary characters in a set of books that received awards for centering LGBTQIA+ experiences (i.e., Stonewall Book Awards), which we expect would have a greater representation of individuals who identify as non-binary or gender-fluid. This exercise entails manually coding the gender for each human named entity (e.g., *Mary*) mentioned more than once in each book as measured by spaCy's Named Entity Recognition software (Honnibal and Montani, 2017). The manually coded gender labels are: *male*, *female*, *non-binary*, or *unknown*. We use context clues within the books to determine gender. These context clues include the character's own identification or, if the character's input is absent, pronouns and gendered descriptions (e.g., *Character X was a woman.*). In the absence of sufficient context clues, we label the gender as *unknown*.

Out of 539 named human entities, only two characters identified as non-binary (0.37%). As a result, the sample size in our data would be insufficient to accurately and precisely estimate embeddings for non-binary and gender-fluid identities. As a result, performing the computational methods used in this paper on non-binary groups, separate from females and males, would not yield reliable results. In light of this, we focus our analysis on only females and males, though we note that this extremely small proportion of explicitly non-binary characters in the collection of books most likely to represent them is an important finding in itself.

4 Methodology

4.1 StoryWords 1.0

We use word embeddings to capture the ways in which gender is represented in these texts. Word embeddings operate under the assumption that words which appear in similar contexts have similar meanings (Firth, 1951; Harris, 1954). In practice, word embeddings are generated by neural networks which 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.[9]

We use the word2vec package from Python's Gensim library to estimate word embeddings (Rehurek and Sojka, 2010).[10] Our word2vec implementation uses the Skip-Gram with Negative Sampling (SGNS) model architecture introduced by Mikolov et al. (2013).[11] When setting the hyperparameters of our models, we followed the recommendations of Levy et al. (2015). After training, the algorithm outputs 300-dimensional vectors of every word in the lexicon of each book. We train separate word2vec models on the *aggregate collection* data as well as on the *collection-*

[7] We limit analysis of `HistWords` starting in the 1920s as the first book in our corpus was published in the 1920s.

[8] The aggregate model across all decades for the `HistWords` collection is not publicly available. We discuss how we estimate `HistWords` collection-level measures for word embeddings in Section 4.1.

[9] Word embeddings are categorized as prediction-based embeddings because they use Machine Learning to predict context words.

[10] While we show results from the implementation of word2vec, our results are similar when we use GloVe (Pennington et al., 2014).

[11] We chose the SGNS architecture as it outperforms other architectures on various linguistic tasks. It is fast to train and inexpensive in terms of memory consumption and disk space (Levy et al., 2015).

by-decade data (cf. Section 3.2). 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.

Each time a word2vec model is trained with the exact same hyperparameters, word neighborhoods may change, which can generate different embedding estimates for each round of training (Hellrich and Hahn, 2016; Antoniak and Mimno, 2018; Burdick et al., 2018). To minimize the influence of idiosyncratic variation in shaping our results, we train 50 separate word2vec models with identical hyperparameters on the *collection-by-decade* and *aggregate collection* data (cf. Section 3.2).[12] We name the resulting embeddings dataset `StoryWords 1.0`. We make this available on our GitHub.

4.2 Word Embedding Association Tests

Group (Gender) Words. We develop a vocabulary of words that comprise two gender groups (*females*, *males*). The words associated with females and males were generated by drawing upon commonly used words for each category, in addition to incorporating words from sources such as those lists provided 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 in our gender group word lists, but are not in prior group lexicons, such as those in Caliskan et al. (2017) and Garg et al. (2018).[13] Our lexicon includes 71 pairs of (*females*, *males*) words. Each word within a given category is exclusive to that category only. The final list of gendered words can be found on our GitHub.

Domain Words. We seek to understand how females and males are depicted within these children's books in relation to different attributes (e.g., *traits*, *behaviours*, *occupations*, *physical characteristics*). The choice of these attributes is based on their importance for children's beliefs and perceptions of themselves and others. Each of these is commonly portrayed in children's literature (Nodelman, 2008; Rudd, 2012; Beauvais, 2015).

Our empirical analysis follows Caliskan et al. (2017), supplemented with analysis of whether females are more associated with descriptions of appearance and related terms than males. Our decision to add this analysis follows prior research indicating that men are often described by words that pertain to behaviour, whereas women are typically described by adjectives that refer to their physical appearance and sexuality (Caldas-Coulthard and Moon, 2010).

We constructed our final word lists as follows. We first began with the domain lists provided by Caliskan et al. (2017) and Senel et al. (2018). We then manually augmented them by drawing upon a set of commonly used words for each domain category: *appearance* (93 words; e.g., *alluring*, *elegant*), *competence* (93 words such as *persuasive*, *reasonable*), *family* (39 words), and *professions* (340 words; e.g., *dancer*, *educator* for women; *architect*, *professor* for men).[14] Our augmentation of these lists was performed through ConceptNet (Speer et al., 2017), a multilingual knowledge graph for natural language words or phrases in their undisambiguated forms. The final list of domain words can be found on our GitHub.

Each word within a given category is exclusive to that category only. For example, the *family* category is notably smaller than other lists because many 'family' words are gendered and therefore were included in the *male/female* lists instead of the *family* list.

4.3 Gender Stereotype Detection

While word embedding association tests help us understand gender stereotypes in collections of text, it is also important to be able to identify specific stereotypes found in individual books so parents can make informed decisions about which books are appropriate for their children. A recently published report shows that nearly two-thirds of preteenagers in America read for pleasure at least once a week (Rideout et al., 2022). As a parent, monitoring what content their children do or do not consume is difficult without information to guide

[12] Because there is only one embedding model published for each decade of the `HistWords` dataset, we cannot perform this exercise on the `HistWords` data.

[13] 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.

[14] We used the occupation census data provided by Garg et al. (2018).

their decisions. Information on gender stereotypes in specific books can be obtained, for example, by using online platforms such as Common Sense Media. In a preliminary exploration of this resource, we found that only 25% of the books included in our corpus have a review.

To automatically identify potential stereotypical topics, we employ SentenceBERT, a modification of BERT that derives semantic sentence embeddings that can be compared using cosine similarity (Reimers and Gurevych, 2019). We leverage three manually annotated corpora with gender stereotype information from previous studies: the Automatic Misogyny Identification (AMI) dataset collection from both `IberEval` (Fersini et al., 2018b) and `Evalita` (Fersini et al., 2018a)), and the dataset released by Chiril et al. (2021). We selected these datasets as they are freely available to the research community.[15] This method classifies a sentence as containing a stereotype if the cosine similarity between the sentence and another sentence which has already been labeled as containing a stereotype is higher than a threshold (T). For our analysis, we apply SentenceBERT to a subset of six books from our corpora of children's books that had commentary in the *What Parents Need to Know* section of Common Sense Media reviews. This section highlights topics that may be of particular concern to parents, flagging content to which they might not want their kids being exposed. We experimentally set the threshold to $T = 0.45$.

5 Results and Discussion

Word Embeddings. For conducting the experiments, we relied on the Word Embedding Fairness Evaluation (WEFE) framework (Badilla et al., 2020), an open source software that encapsulates, evaluates and compares different fairness metrics proposed in the literature. Here, we present the results obtained by using the WEAT metric from Caliskan et al. (2017), the most commonly used association test for word embeddings.[16] WEAT assesses the extent to which a model associates two sets of target words (i.e., *females* and *males*) with sets of attribute words (i.e., *appearance* and *competence*, *family* and *business*, *female* and *male professions*). With values that can range between -2.0 and 2.0, a positive score means that `females` are more associated with words related to `appearance`, `family` or `female professions`, and a negative score means that `males` are more associated with the aforementioned attributes.

The representation and visibility of women has increased substantially over the last century, including in occupations that have been traditionally dominated by men (Goldin, 2014). Despite this considerable progress, differential treatment of women in many dimensions of the economy persists into the 21st century (Blau and Kahn, 2017). Our analysis reflects these patterns, highlighting the incidence and persistence of professional role stereotyping in these corpora (cf. Figure 2 (a)). In addition, our results also highlight that females are more likely to be associated with words related to family than words related to business, relative to these likelihoods for males (cf. Figure 2 (b)). While the association between females and family appears to decrease slightly in the `HistWords` data over time, we see no evidence of a similar decline in our children's book collections (c.f. Figure 2 (e)).

Finally, we quantify the degree to which the language used to describe females and males is different. Figure 2 (c) shows that females are much more likely to be associated with words related to their looks (as opposed to males, who tend to be associated with words related to their competence). This association between females and appearance is decreasing over time in the `HistWords` and `Mainstream` collections, but increases in the most recent decades within the children's book collections.

These results show that all three collections (i.e., `Diversity`, `Mainstream`, and `HistWords`) contain biased representations. We found 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.

Stereotype Detection. We next apply the method for detecting particularly salient incidences of gender stereotypes in text, as described in Section 4.3. First, we access reviews from the Common Sense Media platform to identify books that are likely to possess highly gender stereotypical language. We then apply stereotype detection to iden-

[15] These datasets contain tweets that are annotated at different granularity levels. While the AMI corpora only indicates the presence of a stereotype, the dataset released by Chiril et al. (2021) offers a finer characterization.

[16] Looking across results from the 50 models, we observe a small amount of variation in our results, both for the aggregate and over time measures. In light of this, our results report averages over the 50 models.

(a) Female vs. Male Professions (b) Family vs. Business (c) Appearance vs. Competence

(d) Female vs. Male Professions (e) Family vs. Business (f) Appearance vs. Competence

Figure 2: Which domains are *females* more associated with relative to *males*?

Note: We present the WEAT effect sizes, which show whether `females`, as compared to `males`, are more associated with one set of attribute words relative to another. We present these both overall in panels (a)-(c), and over time in panels (d)-(f). Panels (a) and (d) show `female professions` relative to `male professions`; panels (b) and (e) show `family` relative to `business`; panels (c) and (f) show `appearance` relative to `competence`.

tify specific sentences associated with the themes that were highlighted in the review and that present egregious cases of this type of messaging.

Here we present two sentences from a book in our corpora that this method identifies as containing such stereotypes. This book lends a commentary on rape culture.:

(1) *And I know you're on your period, but there's no need to get cranky with me. ($T = 0.552$)*

(2) *If I had boobs like that, I'd wear a burka or something. ($T = 0.466$)*

Sentence (1) appears to reinforce the stereotype of the hysterical menstrual woman, while sentence (2) discriminates against women based on the widely held belief that the exposure of the chest is a sexual act. It is important to acknowledge that this method does not take into account the context surrounding these sentences when identifying them as containing stereotypes. These detected sentences, together with the results from our word embeddings analysis, provide us with evidence of how highly gender stereotypical messages can appear in the text of children's books, even those recognized in highly prominent national book awards.

Future work should more deeply interrogate the stereotypes being transmitted in these children's books, as highlighted by these few examples. Such work should also include an exploration of the context of such phrases.

6 Conclusion

In this paper we demonstrate how NLP tools can be used to investigate the incidence of 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 they are portrayed. We make two primary contributions. First, we analyzed how gender roles are portrayed

in children's literature. Second, we created the first word embeddings trained on a century of award-winning children's literature, `StoryWords 1.0`. Consistent with previous research, our results show that females are more likely than males to be represented in relation to their appearance and their roles in the family.

While we cannot speak to what 'optimal' representation would look like, our tools make it possible for practitioners, policymakers, and parents with a given goal to measure representation in a given set of books in order to help them make their choices.

Important directions for future work include 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 targets: different groups (to understand how other identities may be differentially represented), as well as additional attributes that convey different societal meanings. Furthermore, researchers must expand definitions of gender to account for non-binary and gender-fluid identities. In the future, we also plan to account for polysemous words by using contextualized word vectors.

Ethical Approval. The research reported in this article involved no human participants and so no human subjects review was sought. Our use of the text data in the children's books in our study is transformative, analyzing the books' content and transforming it, via this analysis, into separate and distinct data, which we conduct under the fair use principle.

An important limitation is that our measure of gender representation binarizes gender, constraining it as *female* and *male*, and does not account for non-binary or gender-fluid identities. This comes as a direct result of the low number of characters identified as non-binary (cf. Section 3.4). With respect to the stereotypical language identified through the sentence similarity approach, we make no claims about the intentions of the authors of these books. The context of these phrases remains to be explored in future work.

Future work should innovate to address these challenges and spur new developments in this under-explored area.

Acknowledgements

We thank Celia Anderson for her contributions to the early stages of this project. For excellent research assistance, we thank Simon Mahns, Noah McLean, Qurat ul ain, Khemraj Shukla, and Charlie Wang. For helpful feedback, we thank Kevin Gimpel, Allyson Ettinger, Chenhao Tan, and others. For financial support, we thank Becker-Friedman Institute at UChicago, Center for Data and Applied Computing at UChicago, Data Science Institute at UChicago, National Academy of Education, and Spencer Foundation. The research reported here was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A200478 to the University of Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We acknowledge the University of Chicago Research Computing Center for support of this work.

References

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2022. What we teach about race and gender: Representation in images and text of children's books. *NBER Working Paper 29123*.

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. Wefe: The word embeddings fairness evaluation framework. In *IJCAI*, pages 430–436.

Clémentine Beauvais. 2015. *The Mighty Child: Time and Power in Children's Literature*, volume 4. John Benjamins Publishing Company.

Denise R Beike and Steven J Sherman. 2014. Social inference: Inductions, deductions, and analogies. *Handbook of social cognition*, pages 209–285.

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.

Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. 2018. Evidence of bias against girls and women in contexts that emphasize intellectual ability. *American Psychologist*, 73(9):1139.

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

Laurence Biscarrat, Marlène Coulomb-Gully, and Cécile Méadel. 2016. One is not born a female CEO and...won't become one! *Gender equality and the media - a challenge for Europe. Routledge, ECREA Book Series*.

Francine D Blau and Lawrence M Kahn. 2017. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.

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

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, pages 77–91.

Laura Burdick, Jonathan K Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102.

Carmen Rosa Caldas-Coulthard and Rosamund Moon. 2010. 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, 21(2):99–133.

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.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "Be nice to your wife! The restaurants are closed": Can Gender Stereotype Detection Improve Sexism Classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844.

Matthew T Crawford, Steven J Sherman, and David L Hamilton. 2002. Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of Personality and Social Psychology*, 83(5):1076.

Kay Deaux and Laurie L Lewis. 1984. Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of Personality and Social Psychology*, 46(5):991.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, pages 879–887.

Diva Dhar, Tarun Jain, and Seema Jayachandran. 2018. Intergenerational transmission of gender attitudes: Evidence from India. *Journal of Development Studies*, pages 1–21.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6132–6142. Association for Computational Linguistics.

Alex Eble and Feng Hu. 2020. Child beliefs, societal beliefs, and teacher-student identity match. *Economics of Education Review*, 77:101994.

Alex Eble and Feng Hu. 2022. Gendered beliefs about mathematics ability transmit across generations through children's peers. *Nature Human Behaviour*, pages 1–12.

Naomi Ellemers. 2018. Gender stereotypes. *Annual Review of Psychology*, 69:275–298.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

Diane Felmlee, Paulina Inara Rodis, and Amy Zhang. 2019. Sexist slurs: Reinforcing feminine stereotypes online. *Sex Roles*, pages 1–13.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 task on Automatic Misogyny Identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on Automatic Misogyny Identification. In *Proceedings of the Third*

Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) colocated with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

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.

Claudia Goldin. 2014. A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Elizabeth L Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing… or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Johannes Hellrich and Udo Hahn. 2016. Bad company—neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical papers*, pages 2785–2796.

M Honnibal and I Montani. 2017. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application*.

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.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from Bollywood movies. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 92–105.

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, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Kimberly A. Neuendorf. 2016. *The Content Analysis Guidebook*. Sage.

Perry Nodelman. 2008. *The Hidden Adult: Defining Children's Literature*. JHU Press.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. ELRA.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Victoria Rideout, Alanna Peebles, Supreet Mann, and Michael Robb. 2022. Common sense census: Media use by tweens and teens 2021. *Common Sense Media*.

Emma Riley. 2017. Increasing students' aspirations: The impact of Queen of Katwe on students' educational attainment. In *CSAE Working Paper WPS/2017-13*.

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.

David Rudd. 2012. *The Routledge Companion to Children's Literature*. Routledge.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220.

Huimin Xu, Zhang Zhang, Lingfei Wu, and Cheng-Jun Wang. 2019. The Cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS One*, 14(11).

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.