Disagreeable, *Slovenly*, *Honest* and *Un-named* Women? Investigating Gender Bias in English Educational Resources by Extending Existing Gender Bias Taxonomies

◆Haotian Zhu and ◆Kexin Gao and ◆Fei Xia and ♡Mari Ostendorf
University of Washington
◆Department of Linguistics
♡Department of Electrical & Computer Engineering
{haz060, kexing66, fxia, ostendor}@uw.edu

Abstract

Gender bias has been extensively studied in both the educational field and the Natural Language Processing (NLP) field, the former using human coding to identify patterns associated with and causes of gender bias in text and the latter to detect, measure and mitigate gender bias in NLP output and models. This work aims to use NLP to facilitate automatic, quantitative analysis of educational text within the framework of a gender bias taxonomy. Analyses of both educational texts and a lexical resource (WordNet) reveal patterns of bias that can inform and aid educators in updating textbooks and lexical resources and in designing assessment items.

1 Introduction

Educational materials for children such as reading comprehension articles or test assessments often protagonize real or fictional characters with gender information, rendering the materials more engaging (Brugeilles et al., 2009). They, however, could carry implicit gender bias and thus potentially reinforce gender stereotypes via children's learning process (Waxman, 2013; Doughman et al., 2021).

One example of such gender bias in educational materials lies in the asymmetrical distribution of males and females in human-generated text such as textbooks, where male and female characters tend to take on different social roles (Brugeilles et al., 2009). Additionally, such gender bias surfaces in the lexical entries and definitions in dictionaries. An open letter (Flood, 2023) calls on Oxford University Press to change its "sexist" definitions of the word "*woman*."

Most research on gender bias in the educational field relies on qualitative methodologies suitable for small-scale analyses (e.g., Namatende-Sakwa (2018); Phan and Pham (2021)). In contrast, gender bias studies in the field of NLP mostly attempt to identify, quantify and mitigate gender bias in NLP applications (Savoldi et al., 2021; Zhao et al., 2019; Bordia and Bowman, 2019), with few looking at educational texts (Li et al., 2020).

Towards the aim to identify and analyze gender bias in educational data using NLP methods, in this paper, we first review recently developed gender bias taxonomies (§3) with an extension to incorporate new types of bias in text. Using NLP techniques, we extract gendered mentions[1] from educational materials (e.g. textbooks, reading materials, etc.) and a lexical resource (WordNet[2] (Miller, 1992)). We quantify different types of gender bias therein to reveal the linguistic patterns most closely associated with such bias. Our contributions include: (1) adopted and extended existing gender bias taxonomies and developed a pipeline for the extraction of person mentions and linguistic features (§4); (2) designed an analysis method for identifying various types of gender bias in text in different dimensions (§5); and (3) applied the analysis method to educational datasets to demonstrate the presence of different types of gender bias.

2 Bias Statement

In this work, we attempt to examine gender bias in human-generated text and specialize it to educational resources such as textbooks, test assessment items and lexicons. We adopt the definition of gender bias as given in Doughman et al. (2021): "an exclusionary, implicitly prejudicial, or generalized representation of a specific gender as a function of various societal stereotypes." Here we employ and extend existing gender bias taxonomies (Hitti et al., 2019) and examine different types of gender bias in educational resources.

People implicitly associate certain behaviors or

[1]We recognize and acknowledge that gender is a spectrum rather than binary; however, in this work, we focus solely on investigating gender bias concerning male and female genders, as explicit non-binary entries in available data are scarce.

[2]https://wordnet.princeton.edu/

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 219–236
August 16, 2024

traits to a specific gender, creating gender stereotypes. Such bias in educational resources can be learned by children through the early process of learning (Waxman, 2013; Doughman et al., 2021) and further perpetuates gender stereotypes. For example, it has been shown that women are generally less represented in textbooks and often associated with family-related roles and traits, whereas men are over-represented and often associated with work-related roles. Such differentiated representation of male and female genders in textbooks, which often serve an instructional purpose, creates a false imagery for children with respect to what roles men and women are expected to undertake, producing unnecessary and harmful gender stereotypes. Furthermore, lexical resources such as WordNet are often used to train NLP systems or as external knowledge bases. The implicit bias within these resources can be passed on to produce biased system outputs that can potentially cause representational harms (Blodgett et al., 2020).

Here, we investigate gender bias in educational resources only for male and female genders for the following reasons: (1) the datasets used for analysis are not recent and up-to-date (all educational datasets are published before 2018). Therefore, the number of people mentioned in those datasets whose gender is non-binary gender is limited; (2) the NLP systems such as coreference resolution in the current pipeline to extract person mentions cannot reliably detect and extract people of non-binary gender. In future work, once trustworthy NLP systems that can reliably detect and extract people of non-binary gender become accessible, the analyses can be extended to incorporate the comparison between binary and non-binary genders by using the same overall pipeline and analysis methods (e.g. odds ratio analysis).

3 Related Work

In this study, we focus on gender bias in educational data. We first discuss a taxonomy of gender bias in human-generated text and then review previous research on gender bias in the educational field and in NLP research.

3.1 Taxonomy of Gender Bias

To meaningfully categorize various kinds of gender bias, Hitti et al. (2019) propose two types of gender bias in text: **structural** and **contextual** bias. **Structural** bias "occurs when bias can be traced down from a specific grammatical construction," including gender generalization (e.g., generic *he*) and explicit marking of sex (e.g., "*chairman*" vs. "*chairwoman*"). **Contextual** bias "requires the learning of the association between gender marked keywords and contextual knowledge," which includes societal bias, where traditional gender roles reflect social norms, and behavioral bias, which is a generalization of attributes and traits onto a gendered person. Examples are given in Table 1 (**B3** (1) and (2)).

Based on Hitti et al. (2019), Doughman et al. (2021) and Doughman and Khreich (2022) provide a more fine-grained taxonomy with five types of gender bias, linking each type to possible real-world implications. Our work builds on and expands the taxonomies, as further described in §4.2.

3.2 Gender Bias Studies in Educational Field

There exists substantial research on gender bias in educational settings for various languages and regions, including: English textbooks in Uganda (Namatende-Sakwa, 2018) and Vietnam (Phan and Pham, 2021), in Vietnamese story textbooks (Vu, 2008) and Arabic textbooks (Izzuddin et al., 2021).

Research on gender bias in educational corpora mostly resorts to traditional approaches such as content analysis (Stemler, 2001) and critical discourse analysis (CDA) (Locke, 2004). Despite their obvious strengths in providing in-depth understanding of gender bias, manual coding is required, which is impractical for widespread use.

In this work, we study gender bias in an educational setting by building on linguistic constructs associated with qualitative categories of bias, but enable scalable quantitative analysis by applying NLP methods.

3.3 Measuring Gender Bias in Text

Cryan et al. (2020) explore automating bias analysis in text by developing lexicon-based and machine learning algorithms for gender stereotype detection from a corpus manually coded for gender stereotypes. This approach is limited to the particular gender stereotypes used in annotation.

An alternative approach is to compute some statistic associated with gendered mentions in different linguistic contexts, leveraging NLP analysis tools to automatically annotate linguistic contexts. For example, Zhao et al. (2017) investigate and define gender bias based on the ratio of the joint probability of an activity (e.g., a verb) and a gender group (e.g., female). Bordia and Bowman (2019)

Type	ID	Subtype	Example	Dataset
Structural Bias	**B1**	Explicit Marking of Sex	*policeman*: a member of a police force	WordNet
	B2	Generic *he*	*researcher*: a scientist$_i$ who devotes himself$_i$ to doing research.	Both
Contextual Bias	**B3**	Contextual Bias	(1) *slovenly* woman vs. *rich* man (2) Women are *incompetent* at work.	Both
Additional Bias	**B4**	Distributional Bias	for `textbook` dataset, 32,884 male mentions and 14,308 female mentions are extracted.	Both
	B5	Namedness	for `textbook` dataset, 73.46% male mentions are named, while 32.02% females are named	Corpora
	B6	Definitional Bias	*horseman*: a man skilled in equitation *horsewoman*: a woman horseman	WordNet

Table 1: Taxonomy with types and subtypes of gender bias examined in this study, along with the dataset on which specific subtype is investigated and examples. Additional bias types are newly added to this taxonomy. In the examples, red indicates male gender; blue female; green neutral. Mentions that refer to the same person are indicated by i. Examples in B1, B2, B3 (1) and B6 are the definitions of entries from WordNet. Example (2) in B3 is from Doughman et al. (2021).

use a point-wise mutual information (PMI) based statistic. The odds ratio (OR) is often adopted statistic for measuring gender bias in text (Valentini et al., 2023), and will be adopted in our work. An advantage of this approach of using statistics on a range of linguistic contexts is that it can reveal biases not anticipated in manual coding.

Studies that have taken this approach with texts for children include Li et al. (2020), which explores gender and cultural bias in U.S. history textbooks used in Texas and Toro Isaza et al. (2023), which investigates gender bias in fairy tales for children. Our work is informed by these studies, but it is grounded in a bias taxonomy, and we also investigate a lexical resource.

3.4 Gender Bias Studies in NLP research

For NLP models, researchers look at the existence of gender bias in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; May et al., 2019), large language models (LLMs) (Bordia and Bowman, 2019; Fatemi et al., 2023), and in tasks such as coreference resolution (Zhao et al., 2018), machine translation (Savoldi et al., 2021), among others. Another important aspect of gender bias studies in NLP concerns bias mitigation in NLP applications (Savoldi et al., 2021; Bolukbasi et al., 2016; Park et al., 2018). These efforts are ultimately concerned with downstream application impact. In our work, the use of NLP is as a linguistic annotation tool, and bias detection is aimed to support human authors of educational texts.

4 Methodology

In this work, we adopt and expand the existing taxonomies for gender bias in human-generated text and attempt to identify different types of gender bias in our datasets. We look at two types of data[3]: educational corpora (denoted corpora henceforth) and lexical resource (WordNet).

4.1 Datasets

There are two major types in the educational corpora: **Content** and **Exam** (listed in Table 2). **Content** datasets mainly include open source textbooks (Michigan, 2014; Siyavula, 2014; CK12, 2007) and reading articles for K-12 education (e.g., `CCS_doc`[4], `wee_bit` (Vajjala and Meurers, 2012), and `OneStop` (Vajjala and Lučić, 2018)); **Exam** datasets contain test items administered either in the U.S. or internationally, including `pisa` (Pisa, 2015), `naep_science` and `naep_math`.[5] These educational corpora cover a wide range of subjects such as math, science, history etc., and diverse linguistic phenomena, offering a rich source for the investigation of gender bias.

For lexical resources, we opt for WordNet [6] for a few reasons. It is widely used in the NLP field and may thereby perpetuating potential biases in downstream tasks. Also, it serves as a rich lexical resource with definitions and semantic relationships among words, which benefits our analysis. Lastly, it offers users convenient and free access to word entries and related information.

[3]Both types of educational materials examined in this paper are in **English**.

[4]https://corestandards.org/assets/Appendix_B.pdf

[5]https://nces.ed.gov/nationsreportcard/

[6]The latest version 3.1 contains only database files but no code is available, therefore we use Version 3.0. https://wordnet.princeton.edu/

Dataset	Content				Exam		
	textbook	CCS_doc	wee_bit	OneStop	pisa	naep_science	naep_math
# of Documents	32,626	168	10,486	567	48	123	446
Avg. # of Sent	4.78	28.55	1.82	35.06	13.10	5.93	2.46
Avg. Sent Length	15.09	19.47	14.02	21.95	18.35	12.08	14.83
Year of Release	2007, 2014	-	2012	2018	2015	-	-

Table 2: Description of educational corpora. The definition of **Instance** differs by datasets: for **Content**, an instance means an article or a paragraph; for **Exam**, an instance is a test item. - indicates the publication year is unavailable.

4.2 Different Types of Gender Bias to Identify

As noted earlier, important related work on detecting gender bias in text (e.g., Li et al. (2020); Toro Isaza et al. (2023)) does not incorporate recent taxonomies of gender bias. To systematically understand what kinds of gender bias exist in educational materials, we adopt and extend the gender bias taxonomy from Hitti et al. (2019) and Doughman et al. (2021). In our study, we first consider **structural bias** and **contextual bias** (as defined in §3.1). We also add three new types of bias: **distributional bias**, **definitional bias** and **namedness**. Table 1 lists all bias types and the datasets used to conduct the analyses, along with examples.

4.2.1 Structural Bias

Explicit Marking of Sex (B1): At the morphological level, explicit marking of sex[7] manifests when gender-neutral entities are denoted using gender marker such as "*-man*" and "*-woman*." Here, the term "gender marker" refers not to markers of grammatical gender but to free morphemes (e.g., "*-woman*" in "*needlewoman*") or head nouns in compound phrases (e.g., "*woman*" in "*slovenly woman*"). **B1** in Table 1 presents an example where "*policeman*" contains the marker "*-man*" but the definition denotes a gender-neutral meaning.

Generic *he* (B2): We also examine the generic usage of gendered pronoun "*he*" where the pronoun is co-indexed with a neutral common noun. As shown in the example from **B2** of Table 1, the word *scientist* is gender neutral but is co-indexed with a male reflexive pronoun "*himself*".

4.2.2 Contextual Bias

In Hitti et al. (2019), contextual bias has two subtypes: societal bias, where a gender is stereotypically assigned a social role, and behavioral bias, where certain attributes or traits associated with a gender can lead to generalized gender stereotypes.

In our work, we use the same word **contextual bias** (**B3**) to refer to societal and behavioral bias due to the nuanced distinction between societal and behavioral bias. For example, the sentence from Doughman et al. (2021) illustrates societal bias: "The event was kid-friendly for all the mothers working in the company," where "*mothers*" are stereotypically assigned the role of caretakers, representing societal bias. However, "*mothers*" are also stereotypically associated with the trait of "caring for kids", which falls under behavioral bias. In our study, stereotypical bias emerges when a specific gender is stereotypically ascribed a social norm or attributed certain traits.

4.2.3 Additional Bias

We add three gender bias types to the taxonomy:

Distributional Bias (B4): This type of bias refers to the uneven distribution of different genders. For example, in our textbook dataset, male mentions appear more frequently than female ones.

Namedness (B5): People in text can be mentions with a real or fictional name or referred to with a common noun such as "*scientist*." Through preliminary examination of the educational corpora, we found that female characters show up as anonymous more frequently than their male counterparts (e.g. "*mother*" vs. "*John*"). Thus, we choose to explore this bias type where males are often given names while females are not. For example, in a corpus, the percentage of male proper nouns is higher than that of females (see statistics **B5** in Table 1). This issue is denoted as namedness bias in our taxonomy.

Definitional Bias (B6): The nuanced definitions given to male and female words implicate the differentiated representation of men and women in lexical resources, which we denote definitional bias. As shown in **B6** in Table 1, the definition given to "*horseman*" only refers to men and is detailed, whereas "*horsewoman*" is defined solely based on the male version: "*horseman*".

[7]The word "**sex**" in this terminology is used by the original author. We keep this terminology in this work for the sake of consistency but do not use sex and gender interchangeably.

4.3 Analysis Methods

We detect different bias types in our datasets by employing a generic pipeline comprising four steps: (1) preprocessing, (2) person mention extraction, (3) gender labeling, (4) bias analysis.

4.3.1 Preprocessing

Corpora: In preprocessing, we use the Stanford `CoreNLP` package[8] (Manning et al., 2014) with steps of sentence segmentation, tokenization, truecasing, POS tagging, named entity recognition, dependency parsing and coreference resolution.

WordNet: In WordNet, an entry can either be a single word (e.g., "*horsewoman*") or a compound phrase (e.g., "*honest woman*"). If a word or phrase has multiple senses, each sense is treated as a distinct entry. Each entry includes a definition and additional details such as syntactic category (e.g., "`NOUN`") and lexicographer (e.g., "`noun.person`"). We extract entries and their definitions from WordNet using the `NLTK` package[9] (Bird et al., 2009) and analyze the dependency structure of the definitions using `CoreNLP`.

4.3.2 Person Mention Extraction

Corpora: We first extract all proper nouns, common nouns and pronouns as mention candidates. We use named entity information and the WordNet sense (i.e., "`noun.person`") information to determine if each candidate is a person. Lastly, in coreference chains, if at least one mention in a chain is considered a person from the previous step, then the rest of the chain is also considered a person. Implementation detail is given in Appendix A.

WordNet: For WordNet, we extract all entries in the "`noun.person`" lexicographer file. We consider these entries as the ones denoting people.

4.3.3 Gender Labeling

Gender labeling procedure outputs three labels: *M* for male, *F* for female and *N* for neutral [10].

Corpora: We label the gender of mentions in corpora based on a two-step heuristic. First, we determine the gender of individual mentions using a list of seed words for pronouns (e.g., "*she*", "*he*") and common nouns (e.g., "*woman*", "*man*") and the `Gender Guesser` API[11] for the first names of proper nouns. Then, using coreference chains, we resolve the gender for mentions whose gender is not determined from the previous step. For example, for common nouns such as "*scientist*," the gender cannot be determined in the first step because it is a profession that can be undertaken by any gender. Through coreference chain where it is co-referred by a gendered pronoun, its gender then can be resolved. Implementation detail is given in Appendix B.

WordNet: The extracted entries are grouped into the three gender categories based on gender indications in their definitions. We create three seed word lists containing terms with obvious gender information (e.g., colored words in the first three examples in Table 3). If the root of the dependency structure of the entry definition or the modifier of the root matches predefined terms, we assign the corresponding gender label to the entry.

Then, unlabeled entries are categorized using those labeled entries. If the root of a definition matches a labeled entry, the unlabeled entry is assigned the corresponding gender label. As the last example in Table 3 shows, the gender of "*roughrider*" is assigned based on the gender of "*horseman*." This iterative process repeats until no further male or female labeling occurs, leaving the remaining unlabeled entries as neutral.

Entry	Definition	Label
horseman	a man skilled in equitation	*M*
actress	a female actor	*F*
needlewoman	someone who makes or mends dresses	*N*
roughrider	a horseman skilled at breaking wild horses to the saddle	*M*

Table 3: Example of entries and definitions from WordNet, along with gender labels assigned through pipeline.

4.3.4 Pipeline Validation

To validate the accuracy of the person mention extraction and gender labeling components in our NLP pipeline, we manually labeled 100 examples from the `pisa`, `naep_math` and `naep_science` datasets. All gendered person mentions (pronouns, proper nouns and common nouns) are annotated with respect to gender. The annotated validation set contains 365 mentions in total (176 male mentions and 189 female mentions). The system identified 368 mentions and the number of correctly extracted mentions is 345.

[8] Version 4.5.3, release date: 3/15/2023, https://stanfordnlp.github.io/CoreNLP/index.html

[9] Version 3.8.1, https://www.nltk.org/index.html

[10] The label *N* for neutral gender can refer to person mentions of either gender (e.g., "*someone*") and groups of people of mixed genders (e.g., "*they*").

[11] https://pypi.org/project/gender-guesser/

Precision	Recall	F-1
93.7%	94.5%	94.1%

Table 4: Evaluation results for person extraction on the hand-labeled evaluation set.

The pipeline can achieve high precision, recall and F-1 scores in extracting the person mentions (see Table 4). The extraction module can produce some false positive extractions such as animal names (e.g., "*Dolly*" (the famous clone sheep)) and planet names (e.g., "*Venus*"). The named entity recognition package can miss some human names (e.g., "*Stacie*", "*Sue*").

For the gender labeling component, the labeling accuracy is 100% for the 100 validation instances where the gold standard mentions match the extracted mentions, because all person mentions in the validation set are in coreference chains and they are co-referred with a gendered pronoun. For larger datasets, the accuracy is not perfect because of several limitations. First, the `Gender Guesser` API is based on a list of proper first names. If a name is not in the list, then the gender cannot be correctly resolved. Second, for non-English names such as Chinese first names, most of the time the gender cannot be determined without further coreference information.

4.3.5 Bias Analysis

Corpora: For distributional bias (**B4**), we count the frequencies of males and females. Linguistic features are extracted to assess their association with gender to examine generic *he* (**B2**), contextual bias (**B3**) and namedness (**B5**).

First, we correlate the POS tags of gendered mentions with gender to investigate generic *he* (**B2**) and namedness (**B5**). By categorizing the verbs that serve as the root of gendered mentions using the agency connotation framework (Sap et al., 2017), we examine what types of verbs are more likely to be associated with a specific gender (**B3**). Agency is attributes of the agent of the verbs, denoting whether the action implies power and decisiveness. For example, "*he obeys*" implies the person "*he*" has low agency, while "*he chooses*" implies "*he*" has high agency. We also extract gendered possessive pronouns and the possessed common nouns. Via a list of kinship terms (e.g., "*mother*", "*father*") (full list in Appendix D), the association between gender of possessive pronouns and kinship terms is measured (**B3**).

WordNet: Initially, we extract proper nouns (usually names of famous persons or fictional figures) from person entries using heuristics, and look into distributional bias (**B4**) based on the frequency of their gender labels. Next, we investigate the use of gender pronouns such as "*he*" (**B2**) in defining gender-neutral entries. Additionally, we employ rule-based techniques to extract person entries ending with gender markers of "*-man*," "*-woman*," and "*-person*"[12] and assess the tendency for gender-specific markers to encompass gender-neutral connotations, indicative of explicit marking of sex (**B1**). Lastly, we scrutinize potential stereotypical bias (**B3**) in entries associated with gender-specific markers and definitional bias (**B6**) by examining how roles marked by "*-man*" and "*-woman*" are depicted.

4.3.6 Gender Bias Statistic

In the analysis of feature bias, we conduct significance testing on the association between gender and a binary feature of interest using Fisher's exact test[13] to obtain p-values[14] at $\alpha = 0.05$ level. In addition, we use odds ratio (OR) to determine the direction and magnitude of association. The odds ratio of a binary related feature $x \in \mathbf{X}$ that measures gender bias in favor of males is given by:

$$OR_x = \frac{M_x / M_{not\ x}}{F_x / F_{not\ x}} \tag{1}$$

where M_x is the count of male mentions with feature x and $M_{not\ x}$ without x. F_x and $F_{not\ x}$ are defined similarly. If the p-value ≤ 0.05, the association is deemed significant. If $OR > 1$, then we observe gender bias toward men, and toward women for $OR < 1$. We choose odds ratio as the statistic to measure association between a specific gender and a feature because it is interpretable and commonly used to measure association between binary categorical variables and it is independent of the marginal distributions, which is desirable for our case since the distributions of male and female mentions are highly asymmetrical.

5 Experiments and Results

In this section, we present our experimental design and results for the corpora and WordNet.

[12] We plan to analyze more gender markers such as "*-or*" in "*actor*" and "*-ess*" in "*actress*" in future works.

[13] We opt for Fisher's exact test instead of Chi-square test because the number of co-occurrences of gender and certain features is too small.

[14] Adjusted via False Discovery Rate for multiplicity.

5.1 Educational Corpora

By extracting gendered mentions with their linguistic features, we investigate four types of gender bias in corpora.

5.1.1 Distributional Bias (B4)

Distributional bias in corpora is examined through comparing the number of extracted male and female mentions. We have observed the evidence for distributional bias in favor of male mentions for all content corpora (Table 5), which adheres to our hypothesis that male mentions are over-represented in text while females are under-represented with respect to mention frequency.

Dataset	Gender		
	M	*F*	*Total*
textbook	32,884*	14,308	47,192
naep_math	159	156	315
naep_science	28	47	75
pisa	97	88	185
wee_bit	2,389*	1,408	3797
CCS_doc	2,127*	810	2937
OneStop	8,178*	2,999	11,177

Table 5: Number of male and female extracted mentions. We only include *M* and *F* counts here since our analysis only considers these two genders. * indicates significance of a one-sided binomial test on the number of male mentions against female mentions at α = 0.05.

5.1.2 Generic *He* in Corpora (B2)

To inspect the usage of generic *he* in corpora, we look at extracted mentions that are only common nouns with no gender information per se in comparison to those that are inherently gendered common nouns. Generic common nouns such as "*researcher*" denote nouns that can address any person in general, while gendered common nouns such as "*mother*" refer to a specific gender in particular. Our finding (Table 6) shows that for all datasets examined, male common noun mentions are typically generic rather than gendered, while female mentions are more likely to be gendered.

Dataset	*Gendered*		*Generic*		OR
	M	*F*	*M*	*F*	
textbook	4,532	6,976	1,652	252	0.10*
wee_bit	234	288	109	16	0.12*
CCS_doc	262	180	210	1	0.01*
OneStop	478	624	422	56	0.10*

Table 6: Gendered vs. generic common nouns in the corpora. We ignore naep_math, naep_science and pisa in this analysis because the counts are too small. **OR** denotes odds ratio. Fisher's exact test performed at α = 0.05. * indicates significance of association. Same notation is used for Table 7 and 8.

5.1.3 Possessive Pronoun and Kinship (B3)

To approach contextual bias where a specific gender is associated with certain societal roles, we create a list of kinship terms such as "*mother*" and "*father*" to categorize the common nouns possessed by a gendered possessive pronoun. Possessive pronouns (e.g., "*his*", "*her*") that occur frequently in the datasets carry important gender information. We examine which gender is more likely to be associated with kinship terms, indicating a stereotypical association of a specific gender with family-related roles. Significant association with kinship terms is observed for the OneStop and CCS_doc datasets with OR < 1: female possessive pronouns (e.g., "*her*") are more likely to co-occur with kinship nouns, while male ones do not.

5.1.4 Agency of Gendered Mentions (B3)

In addition to the previous finding on contextual bias, to examine what kinds of behavior are stereotypically associated with a specific gender, we categorize the verbal roots that head the person mentions in the nominal subject position in the sentences according to the connotation framework in Sap et al. (2017). Significant association (Table 7) between female mentions and low agency verbs in the textbook dataset is detected with an OR < 1, indicating females mentions in textbook are more often associated with low-agency verbs than males do, consistent with the findings in Sap et al. (2017). For the other datasets except naep_math, while insignificant, the OR < 1, displaying a similar trend to textbook.

Dataset	*NEG*		*POS*		OR
	M	*F*	*M*	*F*	
textbook	1,740	884	6,792	2,964	0.86*
naep_math	25	17	56	64	1.68
naep_science	1	10	8	20	0.25
pisa	7	10	45	20	0.31
wee_bit	162	93	555	268	0.84
CCS_doc	177	57	542	173	0.99
OneStop	505	172	3,300	978	0.87

Table 7: Gendered mentions against agency of root verbs. *NEG* refers to verbs for which the subject has lower agency than the object; *POS* means the opposite.

5.1.5 Namedness of Gendered Mentions (B5)

We investigate namedness using the POS tags of gendered mentions. There are three types of male and female person mentions: pronoun (PRP), common noun (NN) and proper noun (NNP). By comparing the distribution of NN and NNP, we discover that males are more likely to be tagged as proper nouns, while females tend to be common nouns. Proper nouns have explicit name information, whereas common nouns can refer to any person in general. The significant correlation (Table 8) between males and whether or not they are proper nouns implies that males tend to receive names, but females typically remain more generic and anonymous. This observation represents previously unreported structural bias where females appear less identifiable through proper names.

Dataset	**POS Tag**				**OR**
	NN		*NNP*		
	M	*F*	*M*	*F*	
`textbook`	6,184	7,228	17,120	3,564	0.18*
`naep_math`	3	11	95	80	0.23*
`naep_science`	10	4	6	24	10.00*
`pisa`	11	26	42	38	0.38*
`wee_bit`	343	304	1,075	544	0.57*
`CCS_doc`	472	181	392	102	0.68*
`OneStop`	900	680	3,052	824	0.36*

Table 8: Male and female mentions against NN and NNP in the corpora.

5.2 WordNet

We conduct experiments on the person entries and definitions extracted from WordNet to elucidate instances of five bias types.

5.2.1 Distributional Bias (B4)

Table 9 shows the number of entries we extract from WordNet. Among all entries in WordNet, 21,463 are person entries.

Among person entries, we define 8,652 proper nouns (e.g., names of famous persons or fictional figures). Labeling the gender of proper names by their definitions is challenging (e.g., the definition of "*Sand*" is "French writer known for ...," exhibiting no gender cue). Therefore, we randomly pick 100 proper nouns and determine their gender based on the information on their Wikipedia pages: 85 of them are males, 14 are females, and 1 entry ("*salian*") refers to a group of people. Among the 99 entries that are individuals, 91 are real persons, 8 are fictional. This adheres to the distributional bias that males are represented more in this lexical resource, possibly due to historical reasons.

The rest of person entries are grouped into *M*, *F*, and *N* based on their definitions (see Section 4.3.3).

All Entries	**Person Entries**				
	Total	*NNP*	*M*	*F*	*N*
227,733	21,463	8,652	592	726	11,493

Table 9: Number of all entries and person entries under the proper noun (*NNP*) group and each gender category in WordNet.

5.2.2 Generic *He* (B2)

Among the neutral person entries (column *N* in Table 9), we find there are 100 entries wherein the roots in the dependency structures of the definitions are either co-referred or co-indexed with gendered pronouns such as "*himself*" (see example in **B2** of Table 1). We count the frequency of gendered pronouns and gender-inclusive pronouns (e.g., "*he or she*" or "*they*"). We find that usage of generic *he* widely occurs in WordNet definitions. Among the 100 definitions, the male generic pronoun is employed in 67 definitions to denote gender-neutral roots, whereas only 33 instances feature gender-inclusive language.

5.2.3 Explicit Marking of Sex (B1)

For person entries that are not proper nouns, we collect those ending with the gender markers ("*-man*," "*-woman*," and "*-person*"). Table 10 displays the breakdown of their gender labels determined by the definitions.

Marker	**Gender**			**Total**
	M	*F*	*N*	
-man	79	0	303	382
-woman	0	61	16	77
-person	0	0	113	113
Total	79	61	432	572

Table 10: Number of unique person entries in WordNet that end with "*-man*," "*-woman*," or "*-person*."

There are notably 303 entries ending with "*-man*" featuring gender-neutral definitions. Also, while the neutral label of the 16 entries with "*-woman*" may seem perplexing, they are deemed neutral due to the absence of gender-specific words in their definitions (see example of "*needlewoman*" in Table 3). We consider gender markers ("*-man*" vs. "*-woman*") and the gender labels of the definitions (*M* and *F* vs. *N*) and observe that the marker "*-man*" is inclined towards denoting gender-neutral entries,[15] providing evidence for explicit marking of sex.

[15] Fisher's exact test: OR = 14.623, $p \ll 0.05$.

5.2.4 Contextual Bias (B3)

In Table 10, some entries have variants representing the same role. For instance, "*chairman*," "*chairwoman*," and "*chairperson*" share the same root morpheme but differ in markers. We classify person entries containing gender markers based on the number of associated variants in Table 11 (Full word lists in Appendix F and example definitions in Appendix G).

Entries w/	Marker	Gender *M*	*F*	*N*	Total
(1) one variant	(1a)-man	50	0	260	310
	(1b)-woman	0	11	1	12
	(1c)-person	0	0	85	85
(2) two variants	(2a) -man	19	0	28	47
	-woman	0	34	13	
	(2b) -woman	0	3	0	3
	-person	0	0	3	
	(2c) -man	2	0	8	10
	-person	0	0	10	
(3) three variants	-man	8	0	7	15
	(3a)-woman	0	13	2	
	-person	0	0	15	

Table 11: Number of entries ending with different gender markers, grouped by number of variants. Numbers investigated in the experiments are marked into red.

In Table 11, row **(1a)** shows that out of the 310 entries marked only with "-*man*", 50 are defined as male, lacking corresponding "-*person*" or "-*woman*" variants. These entries typically pertain to occupational roles (e.g., "*seaman*", "*mailman*"). Row **(1b)** identifies 11 entries solely marked with "-*woman*", some of which carry sexist connotations like "*loose woman*", "*kept woman*", and "*honest woman*", where asymmetric social expectations are imposed on women in contrast to men.

Row **(2)** shows entries with only two markers. Specifically, Row **(2b)** features 3 entries without the "-*man*" variant, all of which ("*disagreeable woman*", "*slovenly woman*", and "*unpleasant woman*") convey negative connotations. Row **(2c)** highlights 10 entries lacking the "-*woman*" version. Notably, the two male entries with "-*man*" ("*rich man*" and "*wealthy man*") lack female counterparts.

In this table, 52 male entries lack "-*woman*" variants[16] and 14 female entries lack "-*man*" variants.[17] We perform Sentiment Analysis on the definitions of these two entry groups using the `vaderSentiment` (Hutto and Gilbert, 2014) API. Results reveal a significant difference,[18] with female entries having a lower average sentiment score (-0.141) compared to male ones (0.056).[19]

The presence of entries like "*disagreeable woman*" and "*rich man*" raises initial concerns, since the modifiers directly convey their meaning, rendering their inclusion in lexical resources less necessary. Moreover, these entries may reinforce gender stereotypes. These observations indicate societal bias, reflecting not only the allocation of certain social roles exclusively to males but also the differentiated sentiment associated with gender.

5.2.5 Definitional Bias (B6)

Furthermore, we examine the definitions of the 62 entries that have both "-*man*" and "-*woman*" variants.[20] We find 10 entries whose definitions for "-*man*" variant are detailed, whereas the corresponding "-*woman*" entries receive simpler definitions derived from their "-*man*" or "-*person*" counterparts (see example of "*horseman*" and "*horsewoman*" in row B6 in Table 1). This approach renders the understanding of "*horsewoman*" reliant on the definition of "*horseman*." For the purpose of ensuring semantic comprehensiveness, meticulous definitions for all variants should be provided, incorporating senses conveyed by all morphemes within the entries to facilitate reader comprehension and mitigate potential bias.

6 Discussion

Our investigation has revealed the pervasive existence of various types of gender bias within both educational corpora and WordNet. Specifically, we have noted the prevalence of distributional bias evidenced by the uneven distributions of males and females across both datasets, alongside explicit marking of sex and the generic use of male pronouns within WordNet. Additionally, a diverse array of syntactic patterns within the corpora has been identified as displaying gender bias.

In this work, we only explore gender bias in English educational materials. The extraction pipeline and gender labeling procedure proposed contain language-dependent components that are unique to English (e.g. using a coreference resolution system to determine gender of a common noun based on gendered pronouns). For languages such as Man-

[16] 52 is the sum of 50 from **(1a)** and 2 from **(2c)** in Table 11

[17] 14 is the sum of 11 from **(1b)** and 3 from **(2b)**

[18] Unpaired two-sample t-test: $t = -2.15$, $p = 0.035$.

[19] The sentiment score ranges from -1 to 1, where [-1, 0) indicates negative sentiment, and (0, 1] indicates positive.

[20] 62 is the sum of **(2a)** and **(3a)** totals in Table 11

darin Chinese where the gender of the pronouns is indistinguishable without orthographic information, the pipeline may integrate language-specific NLP systems to resolve the gender of person mentions. Moreover, the way that gender bias manifests in text can differ from language to language (and culture to culture). Thus, the bias patterns used to detect gender bias in this work will be different.

The presence of gender bias in educational resources carries significant implications. Exposure to those materials can potentially shape children's perceptions through implicit gender bias, fostering the development of gender stereotypes. This perpetuation of biased narratives has far-reaching consequences for societal attitudes and inequality. Moreover, NLP models reliant on lexical resources such as WordNet, wherein gender bias is discernible in multiple forms, may inadvertently perpetuate said biases in downstream tasks.

However, our work offers actionable insights for educational resource developers, offering guidance on elements to consider during the creation process to mitigate bias. Moreover, our study on WordNet pinpoints the bias issues that warrant monitoring and maintenance by developers.

7 Conclusion

In this study, based on the existing taxonomy of gender bias in text, we have examined 7 types of gender bias in educational corpora and WordNet. The analysis has shown that many types of gender bias exist in both types of data, emphasizing the necessity for meticulous examination of such biases in associated resources. Our future work aims to identify additional linguistic features correlated with gender. Furthermore, deeper exploration is warranted into corpora from other domains and lexical resources beyond WordNet.

8 Limitations

There are several limitations to our study: (1) we only consider binary gender in this paper; (2) the small data size of some of the assessment items limits the use of statistical analyses; (3) WordNet as a proxy for a dictionary does not suffice due to its lack of comprehensive entries and definitions and it is not regularly maintained; (4) in this study, we employ odds ratio as the statistic for gender bias, which only considers correlation instead of causation; (5) in this work, we only work with the English language, while gender bias can appear in educational materials in other languages as well.

9 Ethical Considerations

We identify several ethical considerations that are related to our work. (1) First, the educational assessment items typically are not made publicly available, which presents a challenge for multiple researchers to compare methods on the same data and to reproduce our analysis results. However, this type of educational data assumes vital importance to look at, so mechanisms are needed to enable these types of studies. (2) This work is not subjected to privacy concerns since the datasets do not contain identifiable information about individuals. However, famous people (dead or alive) appear in our datasets, and they are potentially used for analysis. (3) Our gender labeling procedure only labels male, female and neutral gender, without consideration of non-binary genders. Such limited consideration and inclusion of binary gender constrains the scope of our study within the binary gender framework, particularly in neglect of stereotypes and bias directed towards non-binary gender community.

References

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Carole Brugeilles, Sylvie Cromer, and Centre développement. 2009. *Analysing gender representations in school textbooks*.

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

CK12. 2007. *CK-12 Free Online Textbooks, Flashcards, Adaptive Practice, Real World Examples, Simulations*.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. 2020. Detecting gender stereotypes: Lexicon vs. supervised learning methods. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Jad Doughman and Wael Khreich. 2022. Gender bias in text: Labeled datasets and lexicons. *CoRR*, abs/2201.08675.

Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. Gender bias in text: Origin, taxonomy, and implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44, Online. Association for Computational Linguistics.

Zahra Fatemi, Chen Xing, Wenhao Liu, and Caiming Xiong. 2023. Improving gender fairness of pre-trained language models without catastrophic forgetting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1249–1262, Toronto, Canada. Association for Computational Linguistics.

Alison Flood. 2023. Fresh call for oxford dictionaries to change 'sexist' definitions. *The Guardian*.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

C. J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Izzuddin, Reza Pahlevi Dalimunthe, and Sulistiyono Susilo. 2021. The portrayal of women in arabic textbooks for non-arabic speakers. *SAGE Open*, 11.

Lucy Li, Demszky Dorottya, Bromley Patricia, and Jurafsky Dan. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas u.s. history textbooks. *AERA Open*, 6.

Terry Locke. 2004. *Critical discourse analysis*. Continuum research methods series. Continuum, London.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *ArXiv*, abs/1903.10561.

Michigan. 2014. *Michigan Open Book Project*.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Lydia Namatende-Sakwa. 2018. The construction of gender in ugandan english textbooks: a focus on gendered discourses. *Pedagogy, culture society*, 26(4):609–629.

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, Brussels, Belgium. Association for Computational Linguistics.

Anh Ngoc Quynh Phan and Tho Xuan Pham. 2021. Gender stereotypes as hidden curriculum: A case of vietnamese english textbooks. *International Journal of Education*.

OECD Pisa. 2015. *Draft science framework*.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Siyavula. 2014. *Open Textbooks | Siyavula*.

Steve Stemler. 2001. An overview of content analysis. *Practical assessment, research & evaluation*, 7(17):1–6.

Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023. Are fairy tales fair? analyzing gender bias in temporal narrative event chains of children's fairy tales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6509–6531, Toronto, Canada. Association for Computational Linguistics.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Francisco Valentini, Germán Rosati, Damián Blasi, Diego Fernandez Slezak, and Edgar Altszyler. 2023. On the interpretability and significance of bias metrics in texts: a PMI-based approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 509–520, Toronto, Canada. Association for Computational Linguistics.

Phuong Anh Vu. 2008. Gender stereotypes in story textbooks for primary school students in vietnam.

Sandra Waxman. 2013. Building a Better Bridge. In *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us*, pages 292–296. Oxford University Press.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

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.

A The Pipeline for Extracting Person Mentions from Educational Corpora

This appendix describes in detail the implementation of the person mention extraction procedure for educational corpora. The corpora first are preprocessed by using the Stanford `CoreNLP` package. After preprocessing the educational corpora, we extract individual person mentions. Person mentions include three kinds: pronouns, proper nouns and common nouns. We first recognize the three types of mentions from text as individual mention candidates using their POS tag information. Using named entity recognition (NER) information and the supersense obtained from WordNet, we determine if each candidate mention is a person if and only if the NER assigns a "*PERSON* tag or its supersense is "`noun.person`". By leveraging coreference resolution, we then form coreference chains. In each coreference chain, if at least one mention in the chain is determined as a person in the previous step, the rest of the chain is deemed as person mentions. The last step is to ensure that common nouns that are missed from the second step are correctly extracted.

B Gender Labeling for Corpora

In this appendix, we describe the gender labeling procedure for the educational corpora.

After extracting person mentions from the corpora, we resolve the gender of the mentions based on a two-step heuristic:

The first step in gender labeling is to check whether or not a mention is in fixed lists of pronouns and common nouns that have salient gender information: for example, "he", "she", "woman", "man" (full lists in Appendix C). If a mention is in the list, then the gender labeling function will output a label from the set $\{M, F, N\}$, where N stands for neutral gender. If a mention is not in the list, we then send the first token of the mention (assuming that the remaining mention is a proper noun) to the `Gender Guesser` API[21]. This API has a list of first names from various countries that have corresponding gender information. If the mention is in the name list, then it will output one label from {*male, female, mostly_male,mostly_female,andy, unknown*}, where *andy* stands for androgynous, meaning a name that is equally probable for male and female. If a mention is not in the name list, then the API will return *unknown*. We group *male* and *mostly_male* to be *M* and *female* and *mostly_female* to be *F*.

Note that there are some issues with this `Gender Guesser` API: it does not predict gender of mentions with only last names. Within the datasets used in this project, there are many last names of famous people of whom the gender is clearly retrievable. Also, the word lists for pronouns and common nouns in Appendix C are not comprehen-

[21]`https://pypi.org/project/gender-guesser/`

sive. To resolve these two concerns, we choose to leverage the coreference cluster information, where we obtain the gender of a mention by the genders of its cluster, if any. The next issue with this API is that it is largely US-centric (although it has an option for country) and does not consider variations across different cultures. We do not attempt to solve this issue in this work.

The gender labeling function using cluster information works as follows:

1. Remove all $unknown$ genders from the cluster if there are other genders in the cluster, e.g. $\{M, F, unknown\}$ becomes $\{M, F\}$
2. If there is a three-way tie between M, F and $andy$, return $andy$.
3. If there is a two-way tie between M and F, return $andy$.
4. If there is a two-way tie between either M or F and $andy$, return M or F. For example, for $\{M, M, andy, andy\}$, return M.
5. If there is no tie, return the most frequent gender.

C Word Lists for Person Pronouns and Person Common Nouns

This appendix contains the word lists for male, female and neutral gendered and neutral person pronouns (excluding "it") and for male, female and neutral gendered person common nouns. The list for common nouns are not exhaustive.

Neutral Pronouns: I, me, we, our, us, myself, ourself, ourselves, let's my, mine, they, them, their, you, your, themself, themselves, yourself, yourselves.

Male Pronouns: he, him, his, himself.

Female Pronouns: she, her, hers, herself.

Female common nouns: girl, woman, mrs, ms, mother, mom, aunt, niece, sister, wife, daughter, grandmother, grandma, grandmom, granddaughter, bride, girlfriend, gal, madam, lady, female, waitress, actress, governess, spinster, empress, heroine, hostess, landlady, stewardess, princess.

Male common nouns: boy, man, mr, father, dad, uncle, nephew, brother, husband, son, grandfather, grandpa, granddad, grandson, groom, boyfriend, guy, gentleman, bachelor, male, actor, emperor, prince.

Neutral Person Common Nouns: people, adult, adults, person, people, child, children.

D Kinship Terms for Detecting Societal Bias (B3)

This appendix provides the list for kinship terms for the analysis of stereotypical bias (**B3**) for educational corpora.

family, son, daughter, brother, child, sister, father, mother, dad, daddy, mum, mom, mummy, niece, nephew, parent, sibling, stepdaughter, wife, husband, spouse, stepfather, stepdad, stepmother, stepmom, grandchild, grandfather, grandmother, grandma, grandmom, grandpa, granddad, grandson, granddaughter, baby[22].

E Example of Instances from the Educational Corpora

This appendix provides instance examples for all educational corpora used in this study.

E.1 `CCS_doc`

A medieval fisherman is said to have hauled up a three-foot-long cod, which was common enough at the time. And the fact that the cod could talk was not especially surprising. But what was astonishing was that it spoke an unknown language. It spoke Basque. This Basque folktale shows not only the Basque attachment to their orphan language, indecipherable to the rest of the world, but also their tie to the Atlantic cod, Gadus morhua, a fish that has never been found in Basque or even Spanish waters. The Basques are enigmatic. They have lived in what is now the northwest corner of Spain and a nick of the French southwest for longer than history records, and not only is the origin of their language unknown, but also the origin of the people themselves remains a mystery also. According to one theory, these rosy-cheeked, dark-haired, long-nosed people where the original Iberians, driven by invaders to this mountainous corner between the Pyrenees, the Cantabrian Sierra, and the Bay of Biscay. Or they may be indigenous to this area. They graze sheep on impossibly steep, green slopes of mountains that are thrilling in their rare, rugged beauty. They sing their own songs and write their own literature in their own language, Euskera. Possibly Europe's oldest living language, Euskera is one of only four European languages–along with

[22]The term "baby" is tricky because it can be used for intimate, non-family members, but when its possessive pronouns are gendered such as "his", "her", it is more likely that "baby" refers to a child.

Estonian, Finnish, and Hungarian–not in the Indo-European family. They also have their own sports, most notably jai alai, and even their own hat, the Basque beret, which is bigger than any other beret.

E.2 `naep_math`

A bag contains two red candies and one yellow candy. Kim takes out one candy and eats it, and then Jeff takes out one candy. For each sentence below, fill in the oval to indicate whether it is possible or not possible.

E.3 `naep_science`

Bacteria and laboratory animals are sometimes used by scientists as model organisms when researching cures for human diseases such as cancer. Describe one possible advantage and one possible disadvantage of using bacteria as models to help find cures for human diseases. Advantage: Disadvantage: Describe one possible advantage and one possible disadvantage of using laboratory animals such as mice, guinea pigs, and monkeys as models to help find cures for human diseases.

E.4 `OneStop`

The Duke and Duchess of Cambridge have won the first part of their fight for privacy. A French magazine was told to stop selling or reusing photos of the royal couple. The pictures show the duchess sunbathing topless while on holiday in the south of France. It is possible that the magazine editor and the photographer or photographers will also have to go to a criminal court. The French magazine Closer was told to give digital files of the pictures to the couple within 24 hours. Closers publisher, Mondadori Magazines France, was also told to pay 2,000 in legal costs. The magazine will have to pay 10,000 for every day it does not give the couple the files. The court decided that every time Mondadori the publishing company owned by the ex Italian Prime Minister Silvio Berlusconi publishes a photograph in the future in France, they will get 10,000 fine. The couple welcome the judges decision. They always believed the law was broken and that they had a right to their privacy. The royal couple are pleased with the decision, but they want to have a much more public criminal trial against the magazine and photographer or photographers. Under French law, if you do not respect someones privacy, you may have to spend a maximum of one year in prison and pay a fine of 45,000. This punishment would send a message to the world and, the couple hope, stop paparazzi taking photos like this in the future. On Saturday the Irish Daily Star also published the photos. And the Italian celebrity magazine Chi published a special edition of 26 pages with the photos of the future queen.

E.5 `pisa`

Mimi and Dean wondered which sunscreen product provides the best protection for their skin. Sunscreen products have a Sun Protection Factor (SPF) that shows how well each product absorbs the ultraviolet radiation component of sunlight. A high SPF sunscreen protects skin for longer than a low SPF sunscreen. Mimi thought of a way to compare some different sunscreen products. She and Dean collected the following: ... Mimi and Dean included mineral oil because it lets most of the sunlight through, and zinc oxide because it almost completely blocks sunlight. Dean placed a drop of each substance inside a circle marked on one sheet of plastic, and then put the second plastic sheet over the top. He placed a large book on top of both sheets and pressed down. Mimi then put the plastic sheets on top of the sheet of light-sensitive paper. Light-sensitive paper changes from dark gray to white (or very light gray), depending on how long it is exposed to sunlight. Finally, Dean placed the sheets in a sunny place.

E.6 `textbook`

Conclusions The scientist must next form a conclusion. The scientist must study all of the data. What statement best explains the data? Did the experiment prove the hypothesis? Sometimes an experiment shows that a hypothesis is correct. Other times the data disproves the hypothesis. Sometimes it's not possible to tell. If there is no conclusion, the scientist may test the hypothesis again. This time he will use some different experiments. No matter what the experiment shows the scientist has learned something. Even a disproved hypothesis can lead to new questions. The farmer grows crops on the two fields for a season. She finds that 2 times as much soil was lost on the plowed field as compared to the unplowed field. She concludes that her hypothesis was correct. The farmer also notices some other differences in the two plots. The plants in the no-till plots are taller. The soil moisture seems higher. She decides to repeat the experiment. This time she will measure soil moisture, plant growth, and the total amount of water the plants consume. From now on she will use no-till methods of farming. She will also research other factors that may reduce soil erosion.

E.7 wee_bit

Nicole Thompson and her third-grade social studies students at Greenbriar Academy in North Carolina wanted to learn about world geography. So late last year, they sent an e-mail message to 100 people. Readers were asked to send the e-mail message to people in other places. Readers were also asked to write something about themselves as well. About six weeks later, Thompson and her students received more than 60,000 e-mail replies! Messages came from every state in the United States and from 120 countries. According to Thompson, the students' favorite response was written by a carpenter at McMurdo Station in Antarctica. "It was a huge deal. We didn't think we would hear from Antarctica!" Thompson said.

F Full Word List for Table 11

This appendix provides the comprehensive word list corresponding to each row of Table 11.

F.1 Row 1a (310 entries that only have -man marker)

freshman, ablebodied seaman, able seaman, abominable snowman, adman, aircraftman, aircraftsman, aircrewman, alderman, apeman, artilleryman, assistant foreman, backup man, backwoodsman, baggageman, bagman, bandsman, bargeman, barman, barrowman, batman, batsman, beadsman, bedesman, beef man, bellman, best man, big businessman, boatman, bookman, border patrolman, bowman, brahman, brakeman, broth of a man, bushman, busman, cabman, cameraman, career man, cattleman, cavalryman, cave man, caveman, chapman, chargeman, chinaman, churchman, city man, clergyman, coachman, coalman, coastguardsman, college man, company man, con man, confidence man, conjure man, corner man, cousingerman, cow man, cowman, cracksman, craftsman, cragsman, crewman, "customers man", dairyman, dalesman, deliveryman, deskman, dirty old man, divorced man, doorman, dragoman, draughtsman, dustman, earthman, elder statesman, elevator man, end man, ent man, everyman, exserviceman, exciseman, family man, feral man, ferryman, fieldsman, fingerprint man, fireman, first baseman, fisherman, foeman, footman, fourminute man, frogman, front man, fugleman, gman, gagman, garbage man, garbageman, gasman, "gentlemans gentleman", government man, groomsman, groundsman, guardsman, gunman, handyman, hangman, hardwareman, hatchet man, heman, head linesman, headman, headsman, heidelberg man, helmsman, henchman, herdsman, highwayman, hired man, hit man, hitman, hodman, holdup man, hotelman, houseman, huntsman, husbandman, iceman, infantryman, ingerman, iron man, ironman, jazzman, journeyman, klansman, "ladies man", landman, landsman, lawman, leading man, ledgeman, lensman, letterman, liegeman, liftman, lighterman, lineman, linesman, linkman, linksman, liveryman, lobsterman, lockman, longbowman, longshoreman, lookout man, lowerclassman, lumberman, machoman, mailman, maintenance man, maltman, marksman, matman, meatman, medical man, medicine man, medieval schoolman, merman, middleaged man, middleman, midshipman, military man, military policeman, militiaman, milkman, minuteman, miracle man, moneyman, motorcycle policeman, motorman, mountain man, muffin man, muscleman, navy man, night watchman, nurseryman, oddjob man, oilman, ombudsman, organization man, outdoor man, packman, pantryman, party man, patrolman, penman, pigman, piltdown man, pitchman, pitman, pivot man, placeman, plainclothesman, plainsman, plantsman, ploughman, plowman, pointsman, posseman, postman, potman, poultryman, pr man, preacher man, pressman, privateersman, property man, propman, publicity man, quarryman, raftman, raftsman, railroad man, railway man, railwayman, red man, remittance man, renaissance man, repairman, rewrite man, rhodesian man, rifleman, righthand man, roadman, roundsman, sandwichman, schoolman, seaman, second baseman, section man, seedman, seedsman, service man, serviceman, sheepman, showman, sidesman, signalman, skilled workman, soundman, spaceman, sporting man, squaw man, stableman, steelman, steersman, stickup man, stockman, straw man, strawman, strongman, superman, swagman, switchman, swordsman, tman, tallyman, taximan, taxman, third baseman, timberman, tollman, townsman, tradesman, trainbandsman, trainman, traveling salesman, travelling salesman, trencherman, tribesman, triggerman, tv newsman, underclassman, utility man, vice chairman, vigilance man, visiting fireman, warehouseman, watchman, waterman, weatherman, widowman, wild man, wingman, wireman, wise man, wolfman, woodman, woodsman, workingman, workman, yardman, yeoman, yesman

F.2 Row 1b (12 entries that only have -woman marker)

charwoman, cleaning woman, comfort woman, foolish woman, honest woman, kept woman, lollipop woman, loose woman, needlewoman, washwoman, widow woman, wonder woman

F.3 Row 1c (85 entries that only have -person marker)

abandoned person, aliterate person, bad person, bereaved person, bisexual person, blind person, british people, clumsy person, color-blind person, colored person, crabby person, creative person, dead person, deaf-and-dumb person, deaf person, deceased person, diseased person, displaced person, disreputable person, dutch people, eccentric person, emotional person, english people, english person, epicene person, famous person, fat person, forgetful person, french people, french person, good person, handicapped person, heterosexual person, homeless person, hunted person, illiterate person, important person, incompetent person, inexperienced person, influential person, insured person, irish people, irish person, juvenile person, large person, learned person, literate person, nonperson, nonreligious person, nude person, oriental person, poor person, primitive person, professional person, psychotic person, religious person, retired person, scholarly person, self-employed person, selfish person, shy person, sick person, silent person, slavic people, sleepless person, small person, spanish people, stateless person, street person, stupid person, swiss people, thin person, uneducated person, unemotional person, unemployed person, unfortunate person, ungrateful person, unkind person, unperson, unskilled person, unsuccessful person, unusual person, unwelcome person, very important person, visually impaired person

F.4 Row 2a (47 entries that have -man and -woman markers)

-man

airman, assemblyman, beggarman, bionic man, bondsman, bondsman, bondsman, bondman, bondman, clansman, committeeman, congressman, cornishman, councilman, countryman, countryman, englishman, fancy man, fancy man, freedman, freeman, frenchman, frontiersman, gay man, gentleman, horseman, irishman, juryman, laundryman, madman, newspaperman, nobleman, oarsman, outdoorsman, point man, policeman, scotchman, scotsman, selectman, sportsman, statesman, stunt man, unmarried man, vestryman, washerman, yachtsman, yellow man

-woman

airwoman, assemblywoman, beggarwoman, bionic woman, bondswoman, bondswoman, bondswoman, bondwoman, bondwoman, clanswoman, committeewoman, congresswoman, cornishwoman, councilwoman, countrywoman, countrywoman, englishwoman, fancy woman, fancy woman, freedwoman, freewoman, frenchwoman, frontierswoman, gay woman, gentlewoman, horsewoman, irishwoman, jurywoman, laundrywoman, madwoman, newspaperwoman, noblewoman, oarswoman, outdoorswoman, point woman, policewoman, scotchwoman, scotswoman, selectwoman, sportswoman, stateswoman, stunt woman, unmarried woman, vestrywoman, washerwoman, yachtswoman, yellow woman

F.5 Row 2b (3 entries that have -woman and -person markers)

-woman

disagreeable woman, slovenly woman, unpleasant woman

-person

disagreeable person, slovenly person, unpleasant person

F.6 Row 2c (10 entries that have -man and -person markers)

-man

anchorman, common man, draftsman, holy man, layman, public relations man, rich man, straight man, wealthy man, working man

-person

anchorperson, common person, draftsperson, holy person, layperson, public relations person, rich person, straight person, wealthy person, working person

F.7 Row 3a (15 entries that have -man, -woman and -person markers)

-man
black man, businessman, chairman, counterman, enlisted man, foreman, foreman, kinsman, married man, newsman, old man, salesman, spokesman, white man, young man
-woman
black woman, businesswoman, chairwoman, counterwoman, enlisted woman, forewoman, forewoman, kinswoman, married woman, newswoman, old woman, saleswoman, spokeswoman, white woman, young woman
-person
black person, businessperson, chairperson, counterperson, enlisted person, foreperson, foreperson, kinsperson, married person, newsperson, old person, salesperson, spokesperson, white person, young person

G Example Definitions of Entries in Table 11

This appendix provides the example definitions of entries from Table 11.

G.1 Examples from the 50 entries in row (1a)

able-bodied seaman: a seaman in the merchant marine; trained in special skills
able seaman: a seaman in the merchant marine; trained in special skills
backwoodsman: a man who lives on the frontier
bagman: a salesman who travels to call on customers
beef man: a man who raises (or tends) cattle
best man: the principal groomsman at a wedding
career man: a man who is a careerist
cattleman: a man who raises (or tends) cattle
coachman: a man who drives a coach (or carriage)
cow man: a man who raises (or tends) cattle
dirty old man: a middle-aged man with lecherous inclinations
divorced man: a man who is divorced from (or separated from) his wife
elevator man: a man employed to operate an elevator
family man: a man whose family is of major importance in his life
ferryman: a man who operates a ferry

G.2 Examples from the 11 entries in row (1b)

charwoman: a human female employed to do housework
cleaning woman: a human female employed to do housework
comfort woman: a woman forced into prostitution for Japanese servicemen during World War II
foolish woman: a female fool
honest woman: a wife who has married a man with whom she has been living for some time (especially if she is pregnant at the time)
kept woman: an adulterous woman; a woman who has an ongoing extramarital sexual relationship with a man
lollipop woman: a woman hired to help children cross a road safely near a school
loose woman: a woman adulterer
washwoman: a working woman who takes in washing
widow woman: a woman whose husband is dead especially one who has not remarried
wonder woman: a woman who can be a successful wife and have a professional career at the same time

G.3 Examples from the 47 entries in row (2a)

airman: someone who operates an aircraft
airwoman: a woman aviator

assemblyman: someone who is a member of a legislative assembly
assemblywoman: a woman assemblyman

oarsman: someone who rows a boat
oarswoman: a woman oarsman

policeman: a member of a police force
policewoman: a woman policeman

statesman: a man who is a respected leader in national or international affairs
stateswoman: a woman statesman

G.4 Examples from the 3 entries in row (2b)

disagreeable woman: a woman who is an unpleasant person
disagreeable person: a person who is not pleasant or agreeable

slovenly woman: a dirty untidy woman

slovenly person: a coarse obnoxious person

unpleasant woman: a woman who is an unpleasant person
unpleasant person: a person who is not pleasant or agreeable

G.5 Examples from the 2 entries in row (2c)

rich man: a man who is wealthy
rich person: a person who possesses great material wealth

wealthy man: a man who is wealthy
wealthy person: a person who possesses great material wealthy

G.6 Examples from the 15 entries in row (3a)

businessman: a person engaged in commercial or industrial business (especially an owner or executive)
businesswoman: a female businessperson
businessperson: a capitalist who engages in industrial commercial enterprise

newsman: a person who investigates and reports or edits news stories
newswoman: a female newsperson
newsperson: a person who investigates and reports or edits news stories