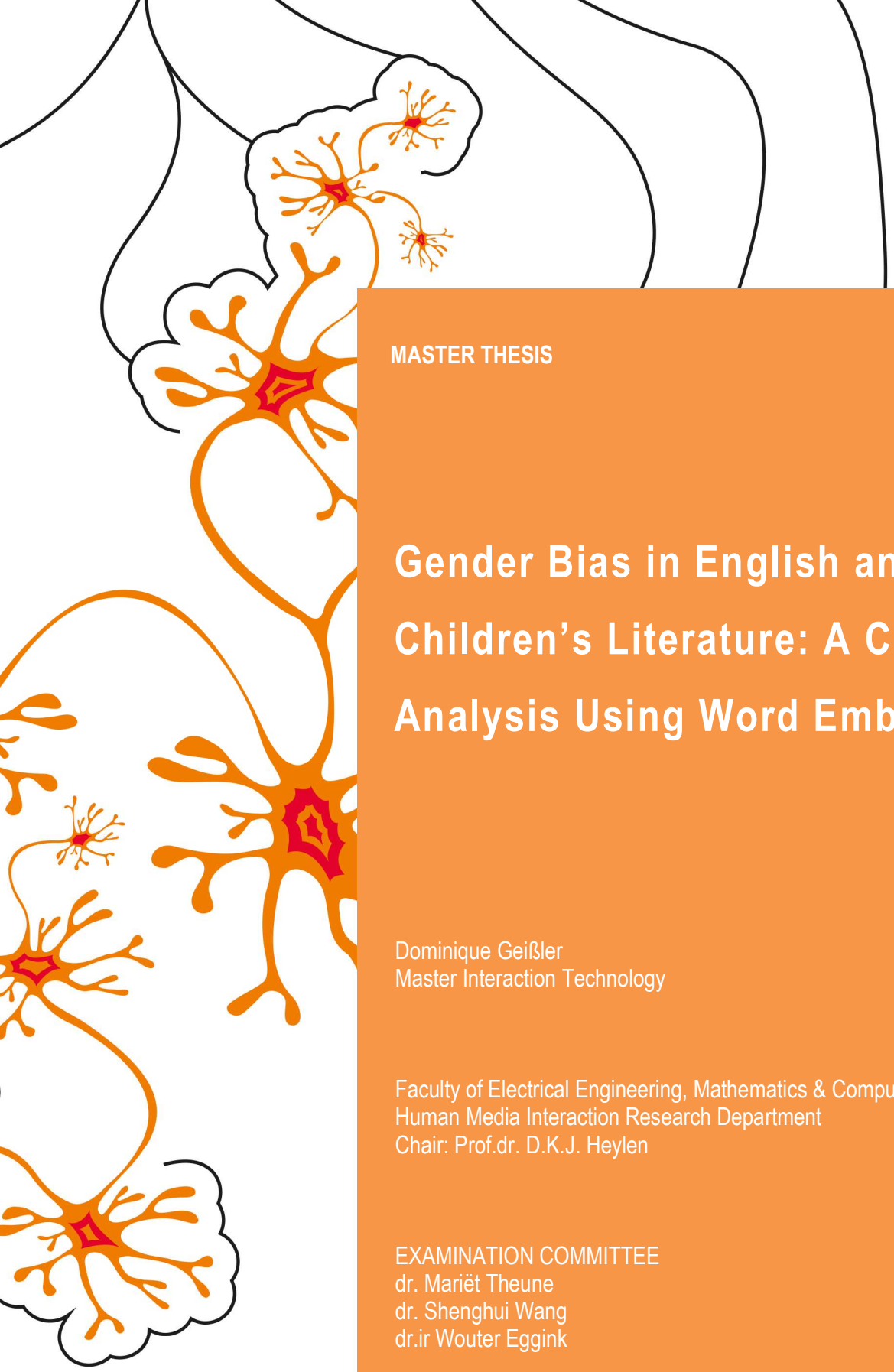


MASTER THESIS

Gender Bias in English and German Children's Literature: A Computational Analysis Using Word Embeddings

Dominique Geißler
Master Interaction Technology

Faculty of Electrical Engineering, Mathematics & Computer Science
Human Media Interaction Research Department
Chair: Prof.dr. D.K.J. Heylen

EXAMINATION COMMITTEE
dr. Mariët Theune
dr. Shenghui Wang
dr.ir Wouter Eggink

24th of September 2021

UNIVERSITY OF TWENTE.

ACKNOWLEDGEMENTS

This thesis marks the end of my Master's degree in Interaction Technology and my time in Enschede. I am grateful for having been given the opportunity to switch academic fields, learning and growing in this new discipline. I have learned skills I never thought I would. I am happy and proud of my professional as well as personal development. Throughout my Master's thesis, I have received a great deal of support and assistance.

I would first like to thank my supervisors Dr. Shenghui Wang and Dr. Mariët Theune for giving me your time, interest and support in this process. Shenghui, I would like to thank you for pointing me towards the topic of gender bias and for your continuous help whenever I struggled with the fuzziness of research papers and word embeddings. And thank you Mariët, for hopping on board so enthusiastically and contributing to this research with your reflected insights and academic experience. The insightful feedback you both provided pushed me to sharpen my thinking and brought my work to a higher level. You have enriched my work with your professional knowledge while your kind words and patience have guided me on a personal level. Thank you, for giving me the freedom to explore and find myself in this research, steering me in the moments I lost sight of the goal.

In addition, I would like to thank my family and my friends for their continuous encouragement and support during my Master's thesis. Thank you, for discussing endless research ideas and approaches until I found the right one. Thank you, for continuously motivating and pushing me. And thank you, for having my back when Corona made life difficult. This accomplishment would not have been possible without your support, guidance and friendship.

ABSTRACT

Gender inequality is a general problem across all societies. Women and girls face discrimination and are underrepresented at all levels of political and economic participation. However, the structural disadvantaging of women and girls is unconnected to physical variation and is instead culturally channelled. Children learn gender roles through sex-role socialisation, the practice of teaching the appropriate behaviours for the sexes. Books play an important role in this as they teach children what the world outside of their environment looks like. Researchers in the area of social science have conducted a thorough analysis of gender bias in children's literature, however, computational efforts to study gender bias in children's books, in particular, are missing. Hence, the research question is: **To what extent can language models be used to examine gender bias in children's books across time and culture?** To fill this research gap, the research leverages the characteristics of word embeddings to examine gender bias computationally. Four language models are trained: two English and two German models trained on full-text books and book descriptions. The research question is answered in four steps. First, the potential of current methods to find gender bias in children's literature is established. Second, methods are tailored to children's books to account for the difference in language use. Third, the methods are used to examine change in gender bias over time. Fourth, a cross-cultural approach is being taken where the results of English language models are compared to German ones.

The research finds that English children's books have a limited view on females, confining them to being nurturing and caring as well as reducing them to their appearance. Males, in return, are characterised by diversity, allowing boys more freedom in their personal development. Over time, gender bias has changed mostly in the reduced association of males with the military. Across culture, many differences between English and German children's books are found, with the latter containing much less stereotypical ideas. The research contributes academically by computationally confirming social science findings and increasing generalisability. In the practical sphere, the research can sensitise parents and educators to gender bias and help them to be more selective in the readings they provide to their children.

TABLE OF CONTENTS

List of Figures

List of Tables

LIST OF ABBREVIATIONS

BLS	U.S. Bureau of Labor Statistics
DNB	Deutsche Nationalbibliothek / German National Library
IAT	Implicit Association Test
NLP	Natural Language Processing
OCLC	Online Computer Library Centre
PCA	Principal Component Analysis
POS	Part-of-Speech
UBE	Unsupervised Bias Enumeration Algorithm
VIAF	Virtual International Authority File
WE	Word Embeddings
WEAT	Word-Embedding Association Test
WEFAT	Word-Embedding Factual Association Test

1 INTRODUCTION

Gender inequality is a general problem across all societies. Women are underrepresented at all levels of political leadership and occupy only 28% of managerial positions across the world [2]. General workforce participation has stagnated at 31% worldwide, with women being at higher risk of poverty [2]. Women face discriminatory laws in all countries and social norms hinder gender parity [2]. While one may refer to societal gender differences being a direct result of physical variation, Vianello and Hawkesworth [3] state that the structural disadvantaging of women and girls across all cultures is unconnected to nature. “Biological diversity is [...] culturally channelled, thereby creating a plethora of differences, which are manifested in social roles, divisions of labor, status hierarchies, structures of consciousness, stylizations of bodies, as well as individual desires and aspirations.” [3, p. 117]. Culture has shaped individual consciousness and shared social expectations of the genders. But also unconsciously, a collective image of the genders has evolved. Females are connected to the arts while males are to follow mathematics and science [4, 5]. Men are rational, skilled and heroic, posing extra challenges for women who want to enter domains where these qualities are required [3]. Females are caring, moral and emotional, hence these traits are not considered manly. From a young age on, boys are not allowed to display behaviours considered female or weak, leading to anxiety surrounding sex-role behaviours and over-excessive displays of masculinity [6].

Gender roles[1] being defined by culture means they originate outside of the individual. They exercise external forces onto people to fit themselves into these predefined roles, disregarding native endowments [6]. This happens through sex role socialisation, the practice of teaching the appropriate behaviours for the sexes. This process starts with the birth of the individual and continues throughout life. Already from a young age on, children can make sex-role distinctions and express sex-role preferences they have [7]. Sex-roles refer to “all the personal qualities, behavioral characteristics, interests, attitudes, abilities, and skills which one is expected to have because one occupies a certain status or position” [6, p. 457]. Since gender bias is already formed at a very young age, this research will focus on the sex roles children are confronted with.

One way to pass down societal values to children is through books. Books teach children what the world outside of their environment looks like and persuade them to accept cultural values. They are read to children by people they hold in high regard, like parents or other family members, leaving lasting impressions on the children [8]. Frawley [9] concluded that children learn gender schema from books to the extent that future information intake is influenced. When presented with stories inconsistent with their existing gender schema, children tended to distort the information so that it responded to their gender biases. Each sex comes with acceptable behaviours that the children are to adhere to. Also, the relative worth of the sexes in society and the personality characteristics are passed down. In Western societies, boys are valued more than girls and are expected to be active and ambitious whereas women are asked to be passive and dependent [10]. Children also adopt occupational sex-role bias and aspire to occupations their respective gender tends to occupy [11]. At the same time, children’s occupational gender

[1] While we note the difference between gender as a social construct and sex as biological attributes, a distinction will not be made in this research and the two words will be used interchangeably.

bias can be altered by presenting them with selective readings, showing the influence of books on children's views and biases [12, 13].

Challenging rigid gender roles and increasing gender equality are important tasks for societies and will not only lead to less psychological pressure on individuals but also greater societal progress. Parity of the genders has been connected to economic growth, a decrease in crime and violence as well as an increase in overall living standards and quality of life. Education of girls and inclusion of women in the workforce leads to prosperity, stability and even sustainability [2, 14]. To defy stereotypical sex roles, they need to be understood in great depth. Given the great influence of books on the development of gender roles in children, this research aims to study the portrayal of sex roles in children's literature.

Researchers in the area of social science have conducted a thorough analysis of gender bias in children's literature by studying the number of males and females present in titles and text, the roles they take in stories, the occupations they go after and how they are described. The most varied study includes 5,618 books from 101 years of American literature. A computationally driven study could widen the scope of books that may be studied, including more books, more years, different cultures and languages. It would also open the opportunity to study gender bias with new methodologies that can give novel insights into how males and females are presented differently in children's books. Hence, the present research aims to expand social science efforts with computational methods, in particular language models. They could be used to look at gender bias computationally and open up possibilities for across time and culture analysis. Hence, the research question is: **To what extent can language models be used to examine gender bias in children's books across time and culture?**

The following chapter will discuss related works within the field of gender bias in children books from a social science and computational perspective (Chapter 2). Subsequently, the research question will be discussed in more depth and sub-research questions are drawn (Chapter 3). Chapter 4 presents and analyses the datasets collected for this research and Chapter 5 proposes a methodology on how to conduct the research. Afterwards, the results of each sub-research question are presented and discussed. Chapter 6 looks at current methods to find gender bias computationally and their applicability to children's literature. Next, the methods are adapted to fit the language models better, which will be presented and discussed in Chapter 7. Following up, change of gender bias across time is researched, presented and discussed in Chapter 8. The research concludes with a comparison of English and German gender bias which is presented and discussed in Chapter 9. The report closes with Chapter 10 with conclusions and recommendations for future work. This Chapter, as well as the Chapters on Related Work, Research Question, Data and Methodology are largely taken from the accompanying proposal of this thesis [15].

2 RELATED WORK

In this Chapter, related work around gender bias in children's literature is presented. First, a definition of bias in general and tailored to language models is given. Second, social science findings on gender representation in children's books are discussed. Third, computational methods for studying gender bias in word embeddings are presented, followed up by a summary of other computational ways to find gender bias. Last, research is discussed, where children's books form the foundation for machine learning and language models.

2.1 Definition of Bias

First, a definition of bias, in particular gender bias, in the scope of the research is given. In general, bias can lead to two types of harms: allocational and representational harms. The former addresses issues of unfair allocation of resources or opportunities to different social groups while the latter is about unfavourable representation or diminishing of social groups or denial of their existence altogether [16]. In Natural Language Processing (NLP), in specific, allocational harm is present when systems perform better on data of the majority group and representational harm occurs when model parameters capture associations between groups and concepts in the following categories: denigration, stereotyping, recognition, under-representation [17]. "Briefly, denigration refers to the use of culturally or historically derogatory terms; stereotyping reinforces existing societal stereotypes; recognition bias involves a given algorithm's inaccuracy in recognition tasks; and under-representation bias is the disproportionately low representation of a specific group" [17, p. 1631]. This research will focus on representational harm as it is about the association of groups, in this case, the genders, with certain concepts and stereotypes. Considering the categories of representational harms, this research positions itself to study representational harm in the form of stereotyping, denigration and under-representation.

Computer systems and algorithms can also contain bias. They may suffer from three types of biases: 1) preexisting, 2) technical, and 3) emergent. The first refers to bias that is present in the input data, the second to technical constraints and differences that lead to bias, e.g. through overfitting, and the third occurs during evaluation of results and contexts of applications [18]. This research will focus on preexisting bias present in the input text, in specific gender bias in children books, which can be captured and even amplified by language models. By studying preexisting bias present in children's books in the form of representational harms, inferences can be made on the gender stereotypes children are confronted with from an early age on [12, 13].

Blodgett et al. [19] used the classification of bias into allocational and representational harms to study motivations and techniques of 146 papers about bias in NLP systems. They criticise the papers to have vague or inconsistent motivations and to lack normative reasoning for why the defined behaviour is biased and to whom. The techniques used to analyse bias in the studied papers are unconnected to the mentioned motivation and efforts are not grounded in relevant literature outside of NLP. To combat their findings, Blodgett et al. [19] suggest three recommendations to guide research in the area of bias in NLP systems. First, researchers are to ground their work in literature outside of NLP that studies the relationship of language and

social hierarchies. Second, they are to describe why the bias is harmful, in what ways and to who. Third, researchers should engage with the experience of social groups that are affected by the bias. These recommendations will be followed up on in the subsequent sections.

2.2 Gender Representation in Children's Books

One way gender bias shows itself is through implicit human bias. Since the bias is implicit, individuals are often not aware of it, making it difficult to measure it. The Implicit Association Test (IAT) has been frequently used to find implicit bias in individuals. In the test, participants are asked to pair concepts and the time it takes them to match the pairs is measured. Findings show that concepts are paired much faster if the subjects consider them similar than when they consider them different [20]. This has shown intrinsic bias of subjects when pairing female names with family faster than with career words, compared to male names [4]. While the test has been frequently used to measure gender bias in various texts, children's books have not been analysed with it yet. Tailored to children books, gender bias is often measured through the number of males and females present in titles and texts, their roles within the story, assigned occupations and characteristics.

Weitzman et al.'s [10] classic study focuses on award-winning picture books for preschool children as they are "read to children when they are most impressionable before other socialization influences (such as school, teachers, and peers) become more important at later stages in the child's development" [10, p. 1126-1127]. The study includes Caldecott award-winning books from 1967 - 1972, Newbery Award winners, Little Golden Books (a popular American book series) and etiquette books (books that explicitly teach gender roles to children). The findings suggest that females are greatly underrepresented in titles, central characters and illustrations, for example, in pictures the male-female ratio is 11:1. The most unequal category is animal characters with a highly unequal male-female ratio of 95:1. At the same time, the study shows that both genders are depicted reinforcing traditional stereotypes: male characters are involved in instrumental, adventurous activities and are engaged in various occupations, while females are shown in dependent and passive activities and are generally presented as wives and mothers. Weitzman et al.'s [10] study also specifically discusses the effects of these gendered portraits on the self-image and aspirations of children. Children feel trapped by these rigid gender roles and it leads to unhappiness, even hampering their full intellectual and social development.

Another important baseline study is Czaplinski's [21] study of Caldecott-award winning picture books of 1940 throughout 1971. Counting the number of males and females in texts and pictures, they found males outnumbering females with 65% to 35% in texts and 63% to 37% in pictures. Surprisingly, these inequalities didn't decrease over time but rather increased with a proportion of 51% females in 1950 to only 23% females in the 1960s. A follow-up study by Davis and McDaniel [22] found the proportion of females to be 39% in texts and 40% in pictures in books published between 1972 and 1997. Overall, the proportion of females in texts didn't exceed 1950s levels according to the study.

Kortenhaus and Demarest [23] studied the distribution and roles of gender in 125 non-award picture books and 25 Caldecott winners or runners-up published between 1940 and 1980. In their study, they looked at the number of males and females in titles, central roles and pictures as well as the number of male and female animals. Their results show that for each female there were three to five males in the same story with titles being the most unequal category. Contradicting the findings of [21, 22], Kortenhaus and Demarest [23] found more equal male/female ratios in the 1970s when comparing the books across decades. Similar to [10], a content analysis was done to study the activities that the main characters of the book undertook. "The 18 that were most prevalent were categorized as either instrumental independent (i.e., actions that involved a lot of self-initiated movement, decision making, and/or creativity) or passive depen-

dent (i.e., actions that required little movement and/or more help from others).” [23, p. 223]. In line with the findings of [10], the distribution of activities on the genders was found to be highly unequal. Males were dominantly presented in instrumental and independent activities while females were involved in passive or dependent actions. When females were active there was usually a male character who was even more active, e.g. a boy and a girl are riding on a horse together but the boy would sit in the front and steer the horse while the girl would sit in the back, holding onto the boy for balance.

Taking a cross-cultural approach outlook, Shahnaz, Fatima and Qadir [24] studied the number of male and female characters in stories published in a popular Pakistani children’s magazine in the years 2006, 2007, 2011 and 2012. Their analysis includes the number of male and female adults, children and elderly depicted in the stories as well as their professions and whether they hold major or supporting roles. Supporting characters are the ones helping or assisting the major characters. When it comes to picture books, they counted the number of male and female front page and in-story pictorials. Their results show that males tend to get assigned professions while females don’t, even if females dominate the front page pictorial. A similar trend can be observed with occupying major or supporting roles. Males tend to be main and supporting characters more often than females, independent of whether there is a male or a female dominating the pictorial on the cover. Also when it comes to assigned attributes, there is a disbalance between the genders as females are not involved in worthwhile activities nor described in detail while “male characters are observed with traits of aggression, creativity, strategic planning, physically assertiveness and the one guiding, dodging or conveying information.” [24, p. 475]. Shahnaz, Fatima and Qadir relate their findings to the wider picture of Pakistani society and conclude that women are withdrawn from the public domain in large parts and if they do appear they are supposed to be passive, silent and in the roles of caregivers and housewives.

The largest study conducted about gender bias in children’s books has been conducted by McCabe et al. [25]. They widened the typical scope of children’s book research to a total of 5,618 books covering the 20th century of U.S. children’s literature. The books include full-text Caldecott award-winning books from 1938-2000, Little Golden Books full texts from 1942-1993 and the Children’s Catalog. The last is a collection of book titles and summaries from 1900-2000 designed for librarians and educators to assign appropriate readings to children. McCabe et al. [25] studied the ratio of males to females in titles and main characters in the books. They also looked at the difference over time using straight time series analysis. Results showed males were in up to 100% of the book titles each year with an average of 36.5% whereas female representation didn’t exceed 75% of the book titles each year with an average of 17.5%. When it comes to main characters, the greatest parity was found in children central characters (male-female ratio of 1.3:1) and greatest disparity in animal characters (male-female ratio of 2.6:1). Concerning the different types of books in this study, Golden Books had the highest disparity, followed by Caldecotts and Catalog, but overall disparities favouring male characters were present. In contrast to [21, 22], analysis over time revealed that the books published between the 1930s and 1960s had the greatest disparity in titles and characters. In return, books published before or after the mid-century had more equal representation with the 1910s and 1990s featuring lightly more girl than boys as central characters. While titles and human central characters are trending towards parity in 1970-2000, animal characters do not.

Having grounded the work in literature outside of NLP (see Table 2.1 for summary), the first recommendation by [19] is adhered to. The following section describes why the bias is harmful, in what ways and to who, by doing so fulfilling the second recommendation of [19]. Research consistently found males to be present in more titles, texts, and pictorials as well as major and supporting roles. Females were underrepresented in all categories across all decades. Since children draw upon books for sex-role guidance [9], under-representation of females leads to young girls thinking that women are not to part-take in society [24] or hamper their intellectual

and social development [10]. Moreover, females were rarely portrayed as occupying a profession in the children books while males were engaged in a wide variety of jobs. Since children aspire to occupations their respective gender tends to inhabit [11], this leads to girls believing they can do fewer jobs than boys and to aspire to become a homemaker or a job with a caregiver function [13]. This development may be reversed if males and females were represented more equally in the books as selective reading can alter a child's vision on occupation [12, 13]. In addition, males were portrayed to take part in instrumental independent activities while females took passive dependent actions [10, 23]. This forces rigid characteristics and activities on children that they are to follow [6] and can make them unhappy and hamper development [10].

Author	**Variables**	**Covered Years**	**Scope**
Weitzman et al. [10]	Males and females in titles Males and females in illustrations Characters' roles Characters' activities	1967-1972	18 Caldecott award winners and runners-up
Czaplinski [21]	Males and females in texts Males and females in illustrations Change over time	1940-1971	31 Caldecott award winners
Davis and McDaniel [22]	Males and females in texts Males and females in illustrations Change over time	1972-1997	25 Caldecott award winners
Kortenhaus and Demarest [23]	Males and females in titles Males and females in illustrations Male and female animals Characters' roles Characters' activities Change over time	1940-1980	125 nonaward books; 25 Caldecott winners or runners-up
Shahnaz, Fatima and Qadir [24]	Males and females in text Males and females in illustrations Characters' roles Characters' occupations Characters' activities Characters' attributes	2006-2007, 2011-2012	36 children's magazines
McCabe et al. [25]	Males and females in titles Characters' roles Change over time	1900-2000	263 Caldecott winners; 1,023 Little Golden Books; 4,485 Children's Catalog (abstracts)

Table 2.1: Summary Studies on Gender Bias in Children's Books

2.3 Computational Examination of Gender Bias in Word Embeddings

The previous studies are all situated within the domain of social sciences and focus mainly on rather limited data scopes. The biggest study included 5,618 books from 101 years. With computational methods, much bigger datasets can be analysed including more books and a wider time range. Moreover, Machine Learning methods could be applied to different languages to compare gender bias across cultures. A computationally driven study can enhance knowledge about gender bias in literature as it allows looking at the issue from a new perspective.

A novel method to study gender bias is the use of word embeddings. In type-level word embeddings, each word is represented by a vector in a 300-dimensional space based on the co-occurrences of words [26]. They are a popular way of representing text in Natural Language Processing applications but contain and might even amplify biases present in text corpora. Gender is represented by a direction in the vector space and gender-neutral and gender-definition words can be linearly separated in the word embeddings. Two types of biases can be present in word embeddings: direct and indirect bias. The former is about the closer association of gender-neutral words with the genders such as `football` being closer to males and `receptionist` being closer to females. The latter is more hidden and results from nuanced correlations of supposedly gender-neutral words. For example, the word `bookkeeper` is closer to `softball` than `football` as they are both more closely associated with the female gender [27].

Bolukbasi et al. [27] used analogies to test for gender bias in embeddings by automatically finding pairs that have a similar relationship as `Man` to `Woman`. By querying the word embeddings for pairs with a similar relationship, analogies can be formed that contain gender bias, for example, `Man:Doctor - Woman:Nurse`. However, analogies have been criticised as a way to measure bias in text corpora. Nissim, van Noord and van der Groot [28] point out that in many studies, all four terms in an analogy are to be distinct, making it impossible to relate both genders to the same word. Querying for `Man:Doctor - Woman:?` could then not return `Doctor` as a fourth term, hence forcing the analogy to find the next best match: `Nurse`. Moreover, a distinction needs to be made between analogies that have a logical fourth term and analogies where this is not the case. In the latter, it needs to be addressed what the desired term would be. For example, what would the logical answer for `Man:Computer_programmer - Woman:?` be?

Another way to study gender bias in word embeddings is the Word-Embedding Association Test (WEAT). Building on the Implicit Association Test, WEAT is a statistical test to see if Machine Learning algorithms can capture bias present in text corpora [26]. WEAT compares the distance of vectors of so-called target words to the gender dimension (attribute words). Distance is measured using cosine similarity. Pre-trained word embeddings by GloVe [29] were used for the test. Findings of the WEAT match those of the IAT, showing that word embeddings contain bias. For example, [26] compared target words representing career and family to the genders and found that male pronouns are significantly closer to career words such as `business` or `office` and female words are closer to family target words, e.g. `home` or `children`. They also looked at the difference of math vs. arts and science vs. arts and found a male association with maths and science and female association with arts.

While WEAT tests for gender bias between two sets of target words, the Word-Embedding Factual Association Test (WEFAT) can be used to relate gender bias present in word embeddings with real-world data. Caliskan et al. [26] compared gender association of occupation words and workforce participation and found the word embeddings to be correlated to the percentage of women in the working field. Occupations with a greater female association such as `kindergarden teacher` also had higher participation of females in the real world. This is in line with the findings of Nosek et al. [5], who suggest that gender bias concerning occupations is linked to gender gaps in the workforce. However, the two might be mutually reinforcing with language bias contributing to the gender gap in science engagement.

Garg et al. [30] used word embeddings as a quantitative lens to study gender and ethnic biases in the 20th and 21st centuries in the U.S. Their focus lies on expanding current qualitative methods to studying bias and comparing bias over time, goals that this research also pursues. First, Garg et al. [30] create word lists for each gender and ethnicity as well as a list of neutral words such as adjectives and occupations. Second, for the gender and ethnicity lists, representative group vectors are calculated. Third, the similarity of the representative group vectors and each word vector in the neutral list is calculated using the average l_2 norm. Fourth, bias is then seen as the differences of the average l_2 norms. Garg et al. [30] suggest that the choice of similarity metrics is not important and that cosine similarity could be used instead of the l_2 norm because the metrics highly correlate with each other.

Demographic occupation data was drawn upon as a comparison to validate the biases that were found. Real-world data is seen as an objective metric of social change and the embeddings are to reflect gender participation in the workforce. Findings suggest that the biases present in the word embeddings indeed correspond to the real-world occupation frequencies: occupations with nearly 50-50 gender participation have no measurable gender bias whereas male-dominated occupations have male bias and the other way around. Having established that word embeddings reflect real-world biases, Garg et al. [30] move on to analysing word embeddings per decade and comparing them to occupation data. Overall, the gender bias in the embeddings is negative, meaning closer association with men than with women. However, the bias moves closer to 0 from 1950 to 1990, representing the increasing participation of females in the workforce. Nevertheless, an analysis of word embeddings and gender stereotypes towards occupation suggests that the embeddings seem to be closer related to human stereotypes than to real-world occupation data.

The embeddings were also used to quantify historical changes in gender stereotypes in literature and culture. Garg et al. [30] used the correlation of word embeddings and human-annotated adjectives to find changes across time. Findings suggest that word embeddings can accurately represent gender stereotypes in literature and culture. Using this knowledge and expanding on the work of [26], two additional gender bias categories were researched by [30] using WEAT: personal descriptions in the intelligence vs. appearance sphere and physical as well as emotional weakness vs. strength. It was found that portrayals of men and women have changed over time and that real-life events, such as the women's movement in the 1960s and 1970s, have influenced these changes systematically. For example, the association of women with intelligence has increased over time, especially after 1960.

Combining the categories used by Caliskan et al. [26] (Career vs. Family, Maths vs. Arts and Science vs. Arts) and Garg et al. [30] (Intelligence vs. Appearance and Strength vs. Weakness), Chaloner and Maldonado [31] studied gender bias in four different embedding algorithms across four domains, including the gender-balanced GAP corpus [32]. Findings suggest that the presence of gender bias differs per domain and algorithm with Google News embeddings being most and biomedical PubMed embeddings being least biased. Findings on the GAP corpus were ambiguous as the corpus was too small to contain all test words. As the five bias categories by [26] and [30] may not be exhaustive, Chaloner and Maldonado [31] developed a method to automatically detect gender bias categories by clustering the words. To measure statistical association with gender, the words in each cluster are then related to known female and male attribute words, e.g. man, male, he, woman, female, and she. According to their findings, the clustered word groups have coherent topics and contain significant bias.

Building on the work of Caliskan et al. [26] as well as Chaloner and Maldonado [31], Kurpicz-Briki [33] applied the WEAT to German and French embeddings to see whether they contain similar biases as English embeddings. They test on gender and origin bias using the following tests from literature: pleasant vs. unpleasant terms and their association with names of different origin; career vs. family terms and their association with gendered names; and math vs. arts as well as science vs. arts terms and their association with gendered terms. While the first two

WEATs were significant, the last two were not. Moreover, Kurpicz-Briki [33] defined two WEATs that are specific to German culture. On the one hand, they test the association of different study choices with gendered terms along the hypothesis that males prefer natural sciences and females the humanities. On the other hand, they test for the association of typically male and female characteristics extracted from 1800 dictionary entries with gendered terms. Both WEATs were both significant. This indicates that the WEAT is able to find gender bias in German and French embeddings to some extent, and can, hence, also be applied to gendered languages.

Inspired by WEAT and its power to unveil bias in word embeddings, Swinger et al. [34] developed the Unsupervised Bias Enumeration Algorithm (UBE). The UBE leverages the principles of parallelism and clustering to automatically extract target word lists for WEAT tests. It takes as input the word embeddings and a list of attribute tokens, e.g. words that represent gender, and finds terms that are statistically associated with the group. They use clustering to group the target and attribute words into meaningful categories. By doing so, they can automatically find bias without the need for hand-crafted target word lists.

Friedman et al. [35] related their findings on gender bias to real-world data. They trained word embeddings on tweets from 100 countries posted in 2018. Similar to [30], they used word lists to group together word vectors of gender and neutral words. Gender bias is then measured as the average axis projection of the neutral set onto the male-female axis. The trained embeddings are correlated with the Global Gender Gap Index. Findings show that word embeddings are an appropriate way of characterising and predicting gender gaps across cultures. For example, their analysis revealed that women have more political influence and power in countries where the political language word list has more female bias. Besides this, GloVe, Word2Vec, CBOW Word2Vec, and FastText embedding algorithms and axis projection, relative l_2 norm difference and relative l_2 norm ratio as metrics were compared. "The Word2Vec approach with axis projection yields the highest coefficient of determination— for both direct and indirect correlation— across all three gender gap and word set pairs" [35, p. 22].

2.4 Other Computational Methods to Study Gender Bias

Besides type-level word embeddings, gender bias can also be studied using other types of Machine Learning (see Table 2.2 for an overview of the methods to study gender bias computationally). Gender bias in contextual word embeddings such as BERT [36] has been studied by Kurita et al. [37] using simple template sentences. The sentences include a target word (gendered-word) and an attribute word (e.g career-related word). First, the target word is being masked to measure the association between target and attribute: "[MASK] is a programmer". Subsequently, both are masked: "[MASK] is a [MASK]". The probability for the attribute to be male or female then constitutes the prior probability of the gender. The bias is calculated as the log of the association and the prior probability. Using this method, Kurita et al. [37] managed to find gender bias in the BERT embeddings. Examining the effects of bias on downstream applications, they found models using BERT to face more difficulties performing coreference resolution when the gender pronoun is female and the topic contains male bias. Moreover, they tried implementing WEAT on BERT by forming sample sentences containing the WEAT word lists. However, statistically significant bias could not be found. This indicates that WEAT cannot be used to measure bias in contextual word embeddings, instead, new methods, like the one form Kurita et al. [37], are needed.

Another approach to studying gender bias computationally is through connotation frames. Connotation frames are about implied sentiment and stereotypes of entities that are transferred through the use of certain predicates. For example, *x violated y* implies: "(1) writer's perspective: projecting x as an "antagonist" and y as a "victim", (2) entities' perspective: y probably dislikes x, (3) effect: something bad happened to y, (4) value: y is something valuable, and (5) mental state: y is distressed by the event" [38, p. 311]. Connotation frames consist of the

relationship of the predicate and other entities modelling the relationship of perspective, effect, value and mental state [38].

Sap et al. [39] use the concept of connotation frames to study gender bias in Hollywood movie scripts. They built a connotation lexicon that can reveal gender stereotypes of the fictional characters and hence contribute to studying bias beyond superficial analyses of screen time, the number of females and the Bechdel test. The analysis of [39] focuses on the theory of power and agency. A character has power if they can influence the actions of others and has agency if they can influence their destiny. Characters can have high or low power as well as high or low agency. Characters are ascribed certain amounts of power and agency through verbs and this influences the audience's perception of the characters. Assumptions about characters can have negative consequences if they play into existing gender stereotypes, e.g. men are assertive and in power and women are helpless and in need of rescuing. In their analysis, each character was given connotation scores for the four metrics (high power, low power, high agency, low agency) based on the actions they did or were subject to. Logistic regression was used to measure the association of gender and the metrics. The findings show that men are assigned more agency than women and tend to use more powerful verbs in a narrative, giving them more authority. In return, women are assigned low-agency verbs, leaving them to contribute to the aesthetics of the movie rather than the plot. When it comes to speech acts, again, men have more power and agency than women. They use inhibitory language more frequently, putting them in a position of blocking or allowing actions. While men show assertiveness through the use of imperative sentences, women speak in hedges, characterising them as uncertain, ambiguous and indecisive.

2.5 Children's Books as the Foundation for Machine Learning

While this research focuses on finding preexisting bias present in children's books to shed light on the type of stereotypes children are confronted with, it can also be relevant when children's literature is at the basis of NLP systems. Using biased children's literature as the input data to NLP systems can lead to the biases being represented or even amplified by the algorithms. Hill et al. [40] examined how statistical models can use language context to make predictions on language content. They used The Children's Book Test dataset consisting of 108 full-text children books retrieved from Project Gutenberg [41]. In their test, they query the model to predict different types of missing words based on smaller contexts (nearby words) or several sentences. Humans can predict all types of words equally well. For high-frequency verbs and prepositions, small contexts are enough but a wider context is needed for named entities and nouns. In return, Recurrent Neural Networks with Long-Short Term Memory tend to perform extremely well on predicting prepositions and verbs but do not achieve human performance when it comes to named entities or nouns. Memory Networks tend to outperform current state-of-the-art NLP models when it comes to predicting nouns and named entities since they rely on local information and the explicit representation of the wider context. For good performance, the latter is critical and small windows are enough if they are centred around important words in the context. Given that Hill et al.'s [40] research is about predicting words, their models might learn underlying associations between certain verbs or nouns and gender. While they might do so accurately given the biased input data, this can replicate or even amplify bias when using the models in new contexts.

Method	Author	Specification	Data
Analogies	Bolukbasi et al. [27]	Ranked Lists; Query for pairs with similar relation as `Man-Woman`	word2vec on Google News
WEAT	Caliskan et al. [26]	Career vs. Family; Maths vs. Arts; Science vs. Arts	GLoVe
	Garg et al. [30]	Intelligence vs. Appearance; Strength vs. Weakness	word2vec on Google News
	Chaloner and Maldonado [31]	all of the above; Clustering to automatically detect WEAT categories	word2vec on Google News; Twitter word2vec; PubMed word2vec; FastText GAP corpus
	Kurpicz-Briki [33]	male vs. female study choice; male vs. female 1800 characteristics;	FastText on German and French CommonCrawl and Wikipedia
	Swinger et al. [34]	UBE to automatically detect WEAT categories	word2vec on Google News; FastText Web corpus; GloVe Web corpus
WEFAT	Caliskan et al. [26]	Relating occupation data to word embeddings	GLoVe
Group vectors	Garg et al. [30]	Difference in l_2 norm	Google Books / COHA; GloVe on New York Times
	Friedman et al. [35]	Axis projection	Twitter word2vec
Template sentences	Kurita et al. [37]	Masking of attribute and target words	BERT
Connotation Frames	Sap et al. [39]	Power vs. Agency of characters	Movie Scripts

Table 2.2: Summary Computational Methods to Study Gender Bias

3 RESEARCH QUESTION

This research aims to address gender stereotypes found in children's literature. The related works show that great efforts have been undertaken to analyse the role of the sexes through a social science lens. However, computational efforts to study gender bias in children's books, in particular, are missing. The research question is: **To what extent can language models be used to examine gender bias in children's books across time and culture?** To answer the research question, several sub-questions are needed. The sub-questions are:

1. How effective are the methods listed in Table 2.2 to study gender bias in children's books?
2. To what extent can the WEAT gender bias categories be adapted to children's literature?
3. How does gender bias in children's books change over time?
4. How does gender bias differ between English and German children's books?

First, the potential of current methods to find gender bias in domain-specific texts will be established. Hence, the first sub-research question is: *How effective are the methods listed in Table 2.2 to study gender bias in children's books?* Answering this sub-research question will establish whether current methods can be used to find gender bias that is inherent in children's books. In specific, will the use of analogies reveal gender-biased pairs? If so, are those similar to the results achieved by [27]? Are the categories and lists of words of the WEAT appropriate to study gender bias in the literature for very young readers? Will the biases in children's literature correspond to real-valued, factual properties of the real-world as was the case in the WEFAT [26]? The expectations are that children's books contain different biases than books addressed to adults. This is the case because different language is used with children and the books tend to have different content. Stories about animals or fairy tales are more common in children's literature than in adult books. Hence, the analogies that will be found are expected to be different than results in [27]. Moreover, it is expected that the current word lists of WEAT are not effective in finding bias in children's literature as they cover different vocabulary and types of stories. When it comes to the relation of bias to real-value, factual properties, it is expected that the bias is more extreme and doesn't relate to the factual property linearly as was the case in [30]. This is because findings of social science studies reveal a skewed vision of children's literature on, for example, adult occupation and children's activities [10, 23, 24].

Second, it will be researched how far the WEAT word lists can be tailored to children's books to account for the difference in language use. This leads to the second sub-question: *To what extent can the WEAT gender bias categories be adapted to children's literature?* Can hand-assembled lists tailored to children's books be constructed? Or can methods to automatically detect lists of words help in revealing biases? This could be done by changing by word-lists of the WEAT based on social science findings, expert knowledge and clustering.

Third, having studied the effectiveness of current methods and their possible adaptation to children's vocabulary, they will be used to examine change over time. Hence, the third sub-question is: *How does gender bias in children's books change over time?* An analysis of gender bias over time will enable researchers to see trends in societies and place gender bias in a larger

societal context as was done by Garg et al. [30]. In line with their findings, it is expected that gender bias decreased over time.

Fourth, *How does gender bias differ between English and German children's books?* A cross-cultural analysis with German children's books is expected to reveal that gender bias differs per culture and that language plays a vital role in this. German being a language with grammatical gender makes this analysis also interesting from an NLP perspective as the transfer of methodology from an ungendered to a gendered language can show the applicability of those methods and reveal the need for new methodology specifically tailored to gendered languages.

Answering these sub-questions will help to shed light on the types of gender stereotypes children are confronted with and help to identify the extent they are present in children's literature. By doing so, parents, educators and guardian can become aware of the issue of gender bias in children's books, which gives them the opportunity to be more selective in the readings they present to their children. Moreover, they can focus on discussing gender stereotypes with their children, hence combating hampering effects on the development of their children. In addition, raising awareness of the gender stereotypes can help authors to identify potentially stereotypical characters in their writing. Knowing which traits are stereotypical offers them the possibility to redefine characters and create unique stories that oppose rather than nurture gender bias.

Academically, the research of gender bias in children's literature can help understanding the applicability of current gender bias methods to new domains. This will reveal whether the methods are usable when a difference in topic and language is present in the texts. Moreover, adjusting the methods can enhance research in the field of gender bias by widening the scope of current methods. In addition, this research gives the opportunity to confirm or reject social science findings computationally, hence expanding the current knowledge base on gender bias in literature addressing a young audience by including more years, viewpoints and cultures.

4 DATA COLLECTION AND PRELIMINARY ANALYSIS

The research question and its sub-questions will be answered using data from several sources. First, descriptions of English books as retrieved from the Online Computer Library Centre (OCLC) will be analysed in terms of their gender bias. As the descriptions explain the main characters and their activities in the book, they are suitable for gender bias analysis [25]. Second, English full-text books from Project Gutenberg [41] will be analysed. Full-text books are hoped to reveal more fine-grained results on gender stereotypes as they contain more words to train word embeddings on. Moreover, they are used to reveal how change can occur over time since most Project Gutenberg books were published around 1800 and the descriptions cover newly published books. A cross-cultural outlook is taken with the fourth sub-research question. For this, German metadata was retrieved from the German National Library (DNB) [42]. German full-text books were taken from Project Gutenberg [41].

4.1 Data From Online Computer Library Centre

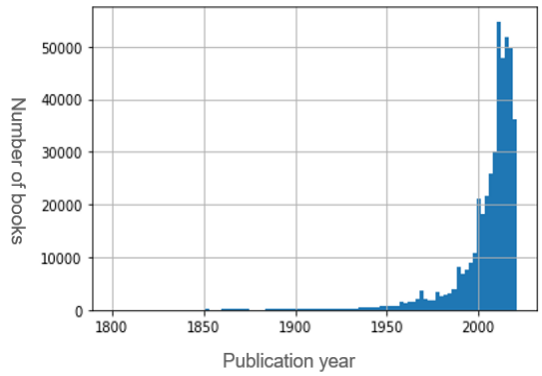
The Online Computer Library Centre (OCLC) is a nonprofit organisation dedicated to making information more accessible to the world and reducing information costs. In specific, they support libraries in cataloguing their works thereby creating an industry standard. WorldCat, a part of OCLC, is the world's biggest database of library collections. It was queried for metadata of children's books using the audience level indicator.

The English metadata retrieved from WorldCat consists of 758,518 books. Each record in the dataset contains the following data: age category of readers, title, abstract, subject, publication date and language. The age level of the audience is categorized in three groups: 1) Library of Congress's classifying scheme *E* for children up to eight years old; 2) Library of Congress's *Fic* for books intended for children older than eight years; 3) and *juvenile works* by FAST thesaurus for books with subject categories *Fic* and *juvenile*. Overall, 23,2% of the books are of the first category, 28,6% of the second and 48,2% of the third category.

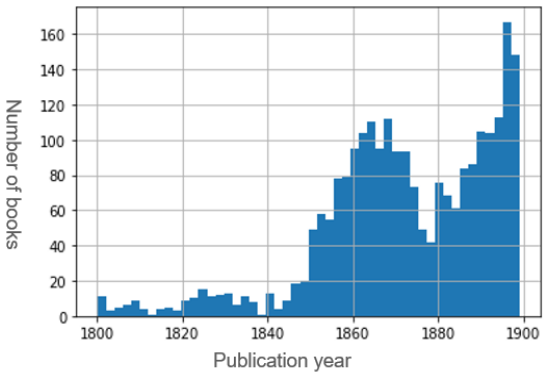
The three categories may not be exclusive and, hence, overlap may occur. When looking at the abstracts of the books, 136.406 books are present in the dataset more than once. Most of them are duplicates, but some books are in the dataset up to 286 times. The most common book is "Black Beauty" with the description: "A horse in nineteenth-century England recounts his experiences with both good and bad masters." Books with duplication in the abstract are removed from the dataset, leaving a total of 447,176 books with unique descriptions. This is done because the word embeddings will be trained on these abstracts and duplicates can bias the embeddings. After taking out duplicates, datapoints with missing values are removed. The dataset has 8,833 books with missing data in the columns subject, publication date and language. Since only the publication date is of interest for the analysis, the 2,689 books with missing data in that field are excluded from the dataset. This leaves a total of 444,487 books for analysis.

The publication dates of the books range from 0 to 9999. Since books on the extreme ends are highly likely to be mistakes in cataloguing or outliers, they will be excluded. Looking at the data per century, numerical analysis shows that only 0.54% of the books were published in

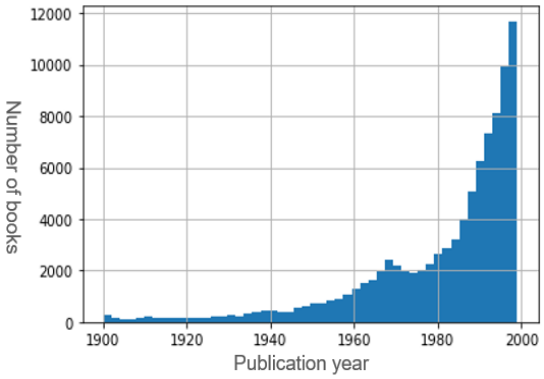
the nineteenth century (2,408 books in total), 20.47% were published in the twentieth century (90,984 books) and 78.98% in the twenty-first century (351,006 books). The distribution of the publication dates per century and overall can be found in Figure 4.1. The histograms confirm the analysis with a vast majority of books being published after 2000. Before 1950, the bars in Figure 4.1a are barely visible, indicating few books were published before that date. Given the research goal of time analysis and the small representation of books before 1900, only books published between 1900 and 2021 are kept. This leaves a total of 441,990 books in the dataset. The mean publication date is 2006, the lower percentile (25%) is in 2002, the median in 2011 and the upper percentile (75%) in 2016. This shows that 75% of the books in the dataset were published after 2002.



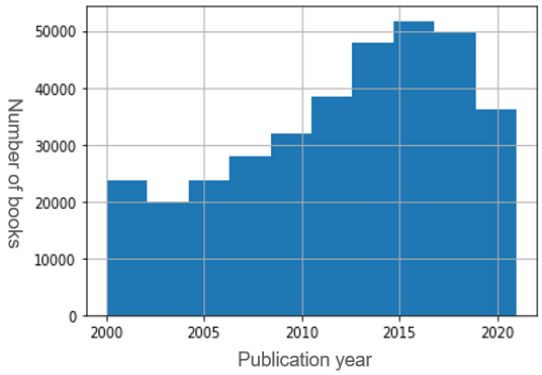
(a) Books Published Between 1800-2021



(b) Books Published Between 1800-1899



(c) Books Published Between 1900-1999



(d) Books Published Between 2000-2021

Figure 4.1: Distribution of the OCLC's Publication Dates

When relating the age level categories to the publication dates, only little differences could be found. Works for children under eight (age category *E*) span the years 1852 to 2021 with a stark increase from 1975. Books with the age category *Fic* (children above eight) were published between 1825 and 2021 with an exponential increase from 1975. The most varied in years is the last category *juvenile works* with books being published between 1800 and 2021. Also here, most books were published after 1975 which is in line with the overall trend of 75% of books being published after 2000.

For training word embeddings, the abstracts of the books will be used. Descriptive statistics show that the average description is 2.8 sentences or 48.7 words long. The amount of words per abstract varies from 0 to 1229 words with the lower percentile (25%) being at 22 words, the median at 34 words and the upper percentile (75%) at 59 words. Hence, on average there are 17.2 words per sentence.

4.2 Data From Project Gutenberg

Project Gutenberg [41] is a volunteer-run online library with over 60,000 free eBooks. It was founded with the mission of free access to literary works and includes mostly older books for which copyright has expired. Data from Project Gutenberg was retrieved using GNU's Wget [43] with filter on format *txt* and languages *en* and *de*. Each book in Project Gutenberg comes with a unique ID, making it convenient to search for specific books. The retrieved files contain the book content itself and additional information like title, author, release date of the EBook, ID and encoding of the file. It also includes the Project Gutenberg license and information on usage. An example of the files' structure can be found in Appendix A. Additional information not required for training the word embeddings was excluded by keeping only the content between *** *START OF THIS PROJECT GUTENBERG EBOOK TITLE* *** and *** *END OF THIS PROJECT GUTENBERG EBOOK TITLE* ***. These were set to mark the beginning and the end of the books' content.

To keep only children's books, Gutendex [44], a JSON web API for Project Gutenberg ebook metadata, was used to filter on children's literature. The API allows searching Project Gutenberg's database using years authors have lived in, copyright status, IDs, languages, mime type, topic and given words in author's names or book titles. For this project, queries included the language and topic condition. The latter searches for keywords in Project Gutenberg's bookshelves or a book's assigned subjects. Gutendex returns a JSON file with the matching books' IDs, titles, authors, translators, subjects, bookshelves, languages, copyright status, media type, format and download count. These were used to filter children's literature within the datasets downloaded from Gutenberg.

4.2.1 English Gutenberg Data

For English children's books, Gutendex was queried using the filters defined in Table 4.1. There was substantial overlap between the two query results, leading to a total of 7,259 eligible books. However, when searching for these books in the downloaded dataset, only 2,809 books were found. A reason for this might be that Gutendex also includes books in *html* format or audiobooks that were not downloaded using GNU's Wget.

Language	Topic	Nr. of results
en	child	3,610
en	juvenile	5,532
		7,259 unique books
		2,312 selected

Table 4.1: English Gutendex Filter

There are some duplicates in the dataset. Since duplication can lead to bias in training the word embeddings, the duplicates were removed and unique books remained in the dataset. Moreover, 795 books were missing the publication date. Since the publication date is needed to do time-axis analysis, the missing publication dates are an issue. While the files themselves include a release date for the ebook version, these are not considered as publication date since they are all dated after 2000 and the dataset includes classic books like *Alice in Wonderland* which were published much earlier. Instead, the publication date is estimated using the authors birth and death year available in the Gutendex metadata. If birth and death year were missing, they were extracted from Virtual International Authority File (VIAF) [45] when possible. For this, VIAF is queried using Python's request library [46]. A typical

query looks like this: *http://viaf.org/viaf/search?query=local.names%20all%20"NAME OF AUTHOR"&sortKeys=holdingscount&recordSchema=BriefVIAF'*. The query retrieves an HTML file with all the search results sorted in a table. The first result is taken as the best match. This can lead to errors when the best match is not correct but given a qualitative analysis of some of the results with a positive outcome, this is the best estimation of the publication date. The best match typically takes the form: *Last name, first name, birth year-death year*. From this, the birth and death year are extracted when present. Since human cataloguing leads to errors and ambiguities, query results that follow a different pattern were treated as not having found a matching result.

For books where both dates are available, the middle year is selected as the publication date, with uneven numbers being rounded up. In cases where only one of the two years is available, the publication date is calculated by adding to or subtracting from the birth or death year, respectively. The average age of the authors (69 for the present dataset) divided by two is used for this. Uneven numbers are rounded up to the next full year. If both, death and birth year, were missing, the books were excluded from the dataset.

The publication dates range from -700 to 2026. This shows that there are outliers in the dataset. Due to low numbers, books published before 1800 and after 2021 were excluded. This leaves a total of 2,312 books in the dataset. The majority of the books were published around 1850-1900. The mean publication date is 1876, the lower percentile (25%) is 1860, the median is 1885 and the upper percentile (75%) is 1904. The distribution of the estimated publication dates of the remaining books can be found in Figure 4.2. Given that the vast majority of books were published before 1950, this dataset can be used for time-axis analysis in combination with the English OCLC dataset. The latter has more than 75% of books published after 2000. Hence, the overlap in books published in the same years is minimal.

A preliminary analysis of the books reveals that, on average, the English books contain 292,768 words and 3,028 sentences. This leaves an average of 96.67 words per sentence which is considerably longer than the sentences retrieved from OCLC. The shortest book has 1,960 words with the median being 261,292 and the maximum being 2,435,174 words. Comparing this to the English OCLC data, the sentences retrieved from the full-text books are nearly twice as long as the sentences in the descriptions of the OCLC data.

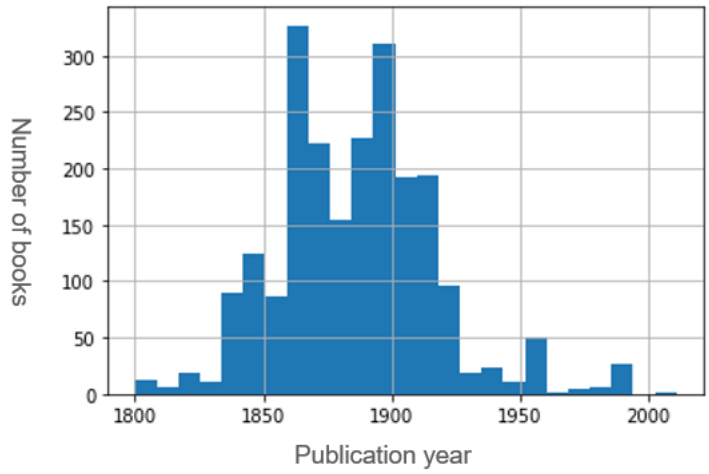


Figure 4.2: Distribution of the English Gutenberg's Publication Dates

4.2.2 German Gutenberg Data

German data was retrieved from Project Gutenberg using the Gutendex search requests as defined in Table 4.2. Combining the results from both requests and taking out overlapping

suggestions, a total of 108 books are in German and about children. Out of these, 97 books were found in the dataset downloaded using GNU'S Wget.

Language	Topic	Nr. of results
de	child	80
de	kind	57
		108 unique books
		86 selected

Table 4.2: German Gutendex Filter

Similar to the English Gutenberg set, duplicates were removed and books with missing publication dates were queried in VIAF. Again, not all authors were found in VIAF. Those books are removed. The publication dates range from 1502 to 1925. Since only few books were published before 1800, those are excluded as well. This leaves a total of 86 books in the dataset. The majority of the books were published around 1860 with the mean publication date in 1868, the lower percentile (25%) in 1851, the median in 1864, and the upper percentile (75%) in 1897. The distribution of the estimated publication dates can be seen in Figure 4.3.

In numerical analysis, it was found that the books contain 248,694 words and 1,881 sentences on average. The average sentence is 132.24 words long. The shortest book has 2,361 words, the median is at 225,893 words and the maximum is 1,690,905 words. In comparison with the English data, the German books are somewhat shorter, both regarding the number of words and sentences, but the sentences tend to be much longer than in English.

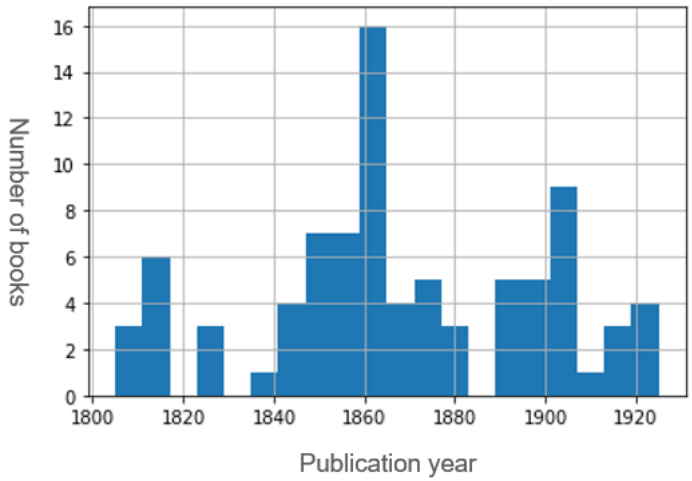


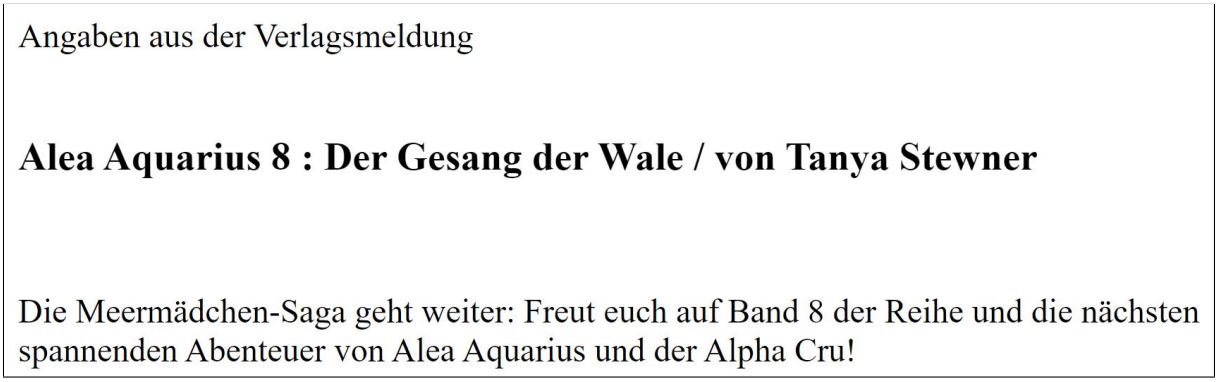
Figure 4.3: Distribution of the German Gutenberg's Publication Dates

4.3 Data From Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek (DNB) is the German National Library which keeps records of all media works published in the German language [42]. It contains a catalogue with over 39 million records including, inter alia, books, newspapers, journals, music, digitised works, maps and special collections. The catalogue was queried for metadata of children's books with the *SRU interface* [47]. Using the search query *Kinderbücher bis 11 Jahre* (children's books until 11 years) and format *MARC21-xml*, 57,695 records were found that matched the requirements. The *SRU interface* was queried using Python's Request library [46]. The queries looked as follows with the

startRecord ranging from 0 to 57600: *https://services.dnb.de/sru/dnb?version=1.1&operation=searchRetrieve&query=Kinderbücher+bis+11+Jahre&recordSchema=MARC21-xml&maximum-Records=100&startRecord=0*.

The number of returned records per query is limited to 100. Hence, 577 XML files with each up to 100 records were downloaded. From these XML files, the book ID, the publication date as well as the link to the description are retrieved. The dataset does not contain the descriptions themselves but links to the corresponding HTML files. These, in return, include a heading with the title and the author as well as the description of the book at hand. An example of such a file can be found in Figure 4.4. These are retrieved using Python library BeautifulSoup [48]. The title and the author are retrieved by splitting the heading at " / von ". The descriptions are the body of the HTML file and are taken as a whole. Only some of the records contain a link to the metadata, leaving a total of 13,123 books in the dataset.


Angaben aus der Verlagsmeldung

Alea Aquarius 8 : Der Gesang der Wale / von Tanya Stewner

Die Meermädchen-Saga geht weiter: Freut euch auf Band 8 der Reihe und die nächsten spannenden Abenteuer von Alea Aquarius und der Alpha Cru!

Figure 4.4: Example DNB Description

Upon close inspection, it could be seen that the dataset also contains books in other languages, e.g. Russian. These were excluded by applying Python's language recognition library langdetect [49]. Also, duplicates and books with missing publication date were excluded. The final dataset contains 11,960 books.

The descriptions have an average length of 92.1 words and 6.2 sentences. This makes the average sentence 14.92 words long. In comparison to the German full-text books, the description sentences are hence much shorter. The shortest description consists of one word and the longest one of 727 words. However, these are probably outliers since the lower percentile (25%) is at 64 words and the upper percentile (75%) at 110 words. Comparing the German descriptions retrieved from DNB to the English descriptions from OCLC, the former has, on average, nearly double the number of words than the latter.

The average publication date of a book in the DNB dataset is in 2017 with the whole dataset spanning 50 years from 1975 to 2025. Books having their publication date in the future are probably still planned to be launched or are a cataloguing mistake. However, they were still kept as these are just outliers and the vast majority of books was published around 2017. This dataset is less varied in terms of publication dates than the English OCLC data. The distribution of the publication dates can be found in Figure 4.5. Nevertheless, it is still suitable for time-axis analysis as it covers mostly modern books and the German Gutenberg dataset contains mostly books from 1850-1900. There is no overlap in terms of years between the two datasets. An overview of all the datasets can be found in Table 4.3.

4.4 Preprocessing

After collecting the data, it needs to be cleaned and processed so that it can be given as input to Machine Learning models to train word embeddings. Several models will be trained: 1) English

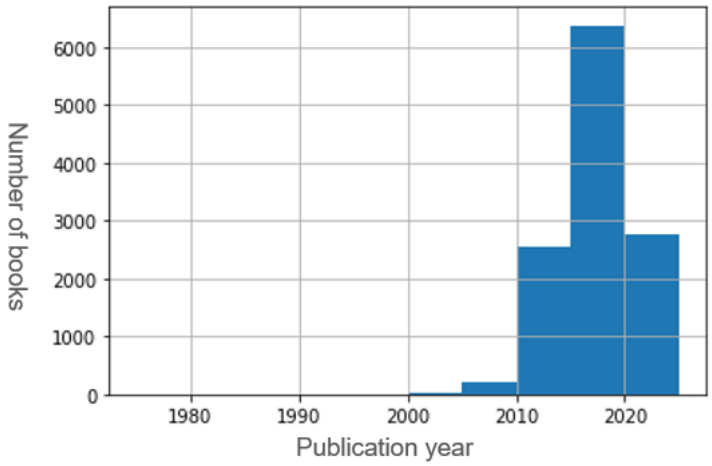


Figure 4.5: Distribution of the DNB's Publication Dates

OCLC model, 2) English Gutenberg model, 3) German Gutenberg model and 4) German DNB model. For training the OCLC and DNB model, only the abstracts of the books are being used. While these are much shorter than the full-text books, they still contain a description of the main characters and their adjectives and roles and can hence contain gender bias [25]. For the Gutenberg models, full-text books are used. From the downloaded Gutenberg files, only the content between the marked beginning and end of the book was used, not the additional information or the license agreement.

The preprocessing of the texts includes the following steps: 1) The text is split into its respective sentences using nltk.sent_tokenize(); 2) All letters are transferred to lower case; 3) Numbers and punctuation are removed; 4) Double whitespace and whitespace at the beginning and end of the text are removed; 5) The sentences are tokenised into words using nltk.word_tokenize(). This makes each book represented by a nested list with the sentences and their respective tokens. This is done because the input to training word embeddings is tokenised sentences and not tokenised documents. Hence, these tokens will be used as input to train the different word embedding models.

Dataset	Items	Pdate range	Mean pdate	Average length	Unique words
English OCLC	441,990	1900-2021	2006	48.7 words	252,886
English Gutenberg	2,312	1800-2021	1876	292,768 words	570,0674
German Gutenberg	96	1800-1925	1868	248,694 words	120,589
German DNB	11,960	1975-2025	2017	92.1 words	75,904

Table 4.3: Summary of the Datasets

5 METHODOLOGY

To answer the research question **To what extent can language models be used to examine gender bias in children's books across time and culture?**, the sub-research questions need answering. For this, a mix of methodologies derived from the literature is used. The following chapter will present an overview of the trained embeddings followed by a discussion of the sub-research questions and their methodologies. A summary of the proposed methods can be found in Figure 5.2 at the end of this chapter.

5.1 Word Embeddings Development

For training the word embeddings, Gensim's Word2Vec [50] is used. Adjustable parameters are the minimum count of occurrences for words in the corpus and the vector size of the embeddings. The minimum count influences the size of the vocabulary. Minimum count and vector size influence training speed and embeddings size. The parameters are tuned based on the results of two evaluation methods provided by Word2Vec.

First, the function wv.evaluate_word_pairs() is used with the standard evaluation set *wordsim353.tsv*, which is included in the Gensim package. The function evaluates the word embeddings by computing the Pearson correlation coefficient and two-sided p-value of the embeddings with human similarity judgements. The results are reported in Table 5.1 in the column *word pairs*. It can quickly be seen that all correlations are positive (Pearson's $r > 0$) and significant. This shows that the embeddings correlate with human similarity judgements, showing their quality. The German embeddings could not be tested with this function as the test set includes only English words, causing a language barrier.

Second, the embeddings are evaluated using wv.evaluate_word_analogies(). It automatically queries the word embeddings for analogies in the format *A:B;C:?* and calculates the accuracy of the embeddings. Two different testing sets are used. The first set is the standard *questions-words.txt*. This set was established by Google and consists of 20,000 syntactic and semantic test examples. It is included in the Gensim package as well. The results of this set are reported in Table 5.1 in the column *Analogies*. Preliminary research has shown that increasing the minimum count and the vector size has lead to higher accuracy levels in this evaluation task. The English Gutenberg embeddings score around 10% higher than the OCLC embeddings. This was to be expected as the corpus is much larger and hence more training data was available. However, the Google test set includes many analogies that are irrelevant to the research proposed in this report, for example, cities in U.S. States or currencies of countries. Hence, the test set is adjusted to include only analogies connected to family and grammar. The results of this set are reported in Table 5.1 in the column *Analogies adj.* It is easily observed that the accuracy on the adjusted test set is higher for both embeddings. This makes sense as the corpora underlying the embeddings are specific in the topics they cover and do not coincide with the general Google test set.

Considering these results, a minimum count of 10 and a vector size of 300 were chosen as parameters as they seem to give the best results for the English embeddings. With those parameters, the correlation and the accuracy are rather high and the vocabulary size is still big

enough. When the minimum count is increased further the vocabulary size becomes very small.

The German embeddings are not evaluated with the Google or adjusted test set, due to the language barrier. Instead, parts of the adjusted test set are translated. This is done for the family section and some of the grammar sections. Grammar sections are only included where this made sense. For example, one section of the test set focused on gerund vs. present tense but the German language does not have a gerund form. This section is then not included. Both, the categories of the English and the German adjusted sets can be found in Appendix C. The accuracy for the German embeddings on the adjusted test set is much lower than for the English sets. One possible reason for this are mistakes in the automatic translation. Given preliminary research results, a minimum count of 10 is considered too high for the German embeddings, as the vocabulary size becomes very small, e.g. the vocabulary size for the DNB embeddings reduced to 8,000 words. Instead, a minimum count of 5 is chosen with a vector size of 300.

Dataset	**Min count**	**Vocabulary size**	**Vector size**	**Evaluation**		
				Word pairs	Analogies	Analogies adj.
English OCLC	10	41,258	300	**0.49**	0.29	0.37
English Gutenberg	10	83,365	300	**0.49**	0.45	0.57
German Gutenberg**	5	28,646	300			0.09
German DNB**	5	14,534	300			0.04

* bold indicates significance with $alpha = 0.05$

** No word pair and analogy evaluation was done for the German models

Table 5.1: Word Embeddings Summary

5.2 Sub-Research Question 1: Applicability of Current Methods to New Domains

The first sub-research question (Sub-RQ1) is: *How effective are the methods listed in Table 2.2 to study gender bias in children's books?* This question is answered using a selection of the methods in Table 2.2. They are fitted on the OCLC word embeddings (OCLC WE). This is the case because the original methods were designed for English texts and to test their applicability to a new domain, this new domain should be in English as well. Moreover, the corpus underlying the OCLC WE stems from the same period as the texts used in current studies. Since the effectiveness of each method to discover gender bias in children's literature is measured in comparison to the results of the papers establishing the respective methods, a comparison is facilitated when using word embeddings with the same language and time period. By comparing the results of the thesis to the literature, it can be seen whether the methods are finding similar bias in children's literature as in adult's texts.

5.2.1 Ranked Lists

At first, gender bias is measured using Ranked Lists. Ranked lists are used by Bolukbasi et al. [27] to study the most extreme occupations projected on the *she-he* axis. The Ranked lists were also calculated by Garg et al. [30] with gendered group vectors, which take into account several female and male words to establish the gender direction. This approach seems more encompassing and is used in this research. The following lists are used to establish the gender direction:

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

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

The vectors of the words in each list are summed and averaged. Then, the two average vectors are subtracted from each other and normalised. This results in a gender direction that does not only take into account the *she-he* axis but the entirety of gendered words. This gender direction vector is set in relation to five types of word lists, which are to shed light on the most male and female associated words. The following lists are tested:

1. Professions as provided by Bolukbasi et al. [27],
2. Nouns from [51],
3. Adjectives assembled by [51],
4. Verbs taken from the power and agency framework of Sap et al. [39],
5. Animals provided by [52].

The similarity between the words in the lists and each of the gendered group vectors is measured using *cosine similarity*: $cos(u, v) = \frac{u \cdot v}{||u|| \, ||v||}$. Since the vectors are normalized, the cosine similarity corresponds to $cos(\vec{w_1}, \vec{w_2}) = \vec{w_1} \cdot \vec{w_2}$.

These types of word lists are chosen as the top ten adjectives describe how each gender is seen and portrayed by the authors of children's books. The verbs describe which activities the children engage in and the occupation bias show which gender is more likely to follow which professions. The top ten nouns show which item and concepts are most connected to the genders while the animals are expected to shed light on the genders the animals in children's books have.

5.2.2 Analogies

Moreover, analogies, as introduced by Bolukbasi et al. [27], are used to measure bias. For this, the word embeddings are queried with gendered pairs for matching neutral pairs. The neutral pairs have a similar relationship as the gendered seed pair. For example, given the gender pair *woman-man*, a pair of words x and y is searched for, where *woman* to x as *man* to y. This outputs pairs of words considered analogous to the seed pair (f, m).

The gender-pair *she-he* is used as seed-pair in line with the approach of Bolukbasi et al. [27]. This pair determines the *seed direction* $\vec{f} - \vec{m}$, i.e. the normalized difference between the two seed vectors. The following metric by Bolukbasi et al. [27, equation (1)] is used to assign a gender bias score to each pair of words x, y:

$$S_{(f,m)}(x, y) = \begin{cases} cos(\vec{f} - \vec{m}, \vec{x} - \vec{y}) & \text{if } ||\vec{x} - \vec{y}|| \leq \delta \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

where δ is a threshold for similarity. Equation 5.1 outputs pairs that are close to parallel to the seed direction and semantically coherent. The latter is enforced through the semantic similarity threshold δ, which ensures that the neutral words are not too far apart. Based on the logic of Bolukbasi et al. [27], $\delta = 1$ is chosen. "Since all embeddings are normalized, this threshold corresponds to an angle $\leq \pi/3$, indicating that the two words are closer to each other than they are to the origin. In practice, it means that the two words forming the analogy are significantly closer together than two random embedding vectors" [27, p. 4]. The output consists of the pairs with the largest positive $S_{(f,m)}$ scores, which do not share the same word x.

5.2.3 Word-Embedding Association Test

Also Caliskan et al's [26] Word-Embedding Association Test (WEAT) is tested on its potential to reveal gender biases in children's literature. The WEAT is based on the IAT and uses two sets of target words (lists that are suspected to be biased, e.g. `programmer, engineer, scientist`; and `nurse, teacher, librarian`) and two sets of attribute words (gendered lists, e.g. `man, male, he`; and `woman, female, she`). The null hypothesis H_0 assumes no difference in the two sets of targets words in respect to their relative similarity to the two sets of attribute words. To test this, a permutation test is done. It computes the probability of a random permutation producing the same (or greater) difference in sample means and hence the (un)likelihood of H_0.

The target sets X and Y of equal size are put in relation to the attribute sets M and F using cosine similarity. The test statistic is $s(X, Y, M, F)$, which measures the difference in association of the two sets of target words with the two sets of attribute words:

$$s(X, Y, M, F) = \sum_{x \in X} s(x, M, F) - \sum_{y \in Y} s(y, M, F) \tag{5.2}$$

where $s(w, M, F)$ measures the association of a word w with the attributes and hence the gender bias:

$$s(w, M, F) = \text{mean}_{m \in M} cos(\vec{w}, \vec{m}) - \text{mean}_{f \in F} cos(\vec{w}, \vec{f}) \tag{5.3}$$

The permutation test then calculates the probability of a random permutation giving the same (or greater) results. This is tested using (X_i, Y_i), the partitions of $X \cup Y$ into two sets of equal size and their difference in test statistic to X and Y. The one-sided p-value of the test is:

$$p = \text{Pr}_i[s(X_i, Y_i, M, F) > s(X, Y, M, F)] \tag{5.4}$$

$$\text{effect size} = \frac{\text{mean}_{x \in X} s(x, M, F) - \text{mean}_{y \in Y} s(y, M, F)}{\text{std\_dev}_{w \in X \cup Y} s(w, M, F)} \tag{5.5}$$

The existing WEAT attribute and target lists are tested on the OCLC embeddings using an adapted version [53]. Since this research is focused on gender bias, only WEAT 6 - 8 by Caliskan et al. [26] are used: career vs. family, maths vs. arts, and science vs arts. Besides that, the additional WEATs of Garg et al. [30], focused on intelligence vs. appearance and strength vs. weakness, are tested. However, the lists are not reported in full in their paper. Moreover, [26] and [30] use a different list of male or female names and male or female terms for each WEAT, making a comparison across WEATs difficult. In consequence, the lists of Chaloner and Maldonado [31] are used instead since they encompass the WEATs of [26] and [30]. Also, the attribute list is the same for all sets of target words, facilitating cross-WEAT comparison. In conclusion, the six WEATs by Chaloner and Maldonado [31] are used as summarised in Table 5.2 (See Appendix B for the exact lists of attribute and target words).

Name	Target words		
W1	career	vs.	family
W2	maths	vs.	arts
W3	science	vs.	arts
W4	intelligence	vs.	appearance
W5	strength	vs.	weakness

Table 5.2: Summary of the WEATs to Be Used

5.2.4 Word-Embedding Factual Association Test

Next, Caliskan et al's [26] Word-Embedding Factual Association Test (WEFAT) is used on the corpus. The WEFAT measures in how far word embeddings can capture real-valued, factual properties of the world. A set of target concepts W (in this case occupations) with a property p_w associated with each word $w \in W$ (in this case the percentage of female workers) is put in relation to two sets of attributes M and F. The test statistics associated with each word in the target set is:

$$s(w, M, F) = \frac{\text{mean}_{m \in M} cos(\vec{w}, \vec{m}) - \text{mean}_{f \in F} cos(\vec{w}, \vec{f})}{\text{std\_dev}_{x \in M \cup F} cos(\vec{w}, \vec{x})} \tag{5.6}$$

The test statistic measures the normalized association of the word vector with the attribute. H_0 assumes no relationship between $s(w, M, F)$ and p_w. This is tested using linear regression with the test statistic being the independent and the property being the dependent variable. For each profession in the profession list provided by [54], the effect size is calculated using the above equation. The gender bias scores of the professions word are then set in relation to 2015 occupation data from the U.S. Bureau of Labor Statistics (BLS). This data is also used for the original findings of [26] and is derived from [54]. Each BLS data point includes an occupation with several tags. The percentage of women in an occupation is calculated as the mean percentage of woman for all BLS data points that include the profession from the profession list as a tag. Subsequently, the effect sizes of the professions are plotted against the percentage of female workers in the respective professions. The Pearson's correlation coefficient and p-value are analysed to understand the relationship between word embeddings and factual property. The code inspiring this research question is taken from [55] and [26].

Method	**Embeddings**	**Metric**
Ranked Lists	OCLC model English Gutenberg embeddings	$cos(\vec{w_1}, \vec{w_2}) = \vec{w_1} \cdot \vec{w_2}$
Analogies	OCLC model English Gutenberg embeddings	$S_{(a,b)}(x, y) = cos(\vec{a} - \vec{b}, \vec{x} - \vec{y})$ if $\|\|\vec{x} - \vec{y}\|\| \leq \delta$, 0 else
WEAT	OCLC model English Gutenberg embeddings	$s(X, Y, M, F) = \sum_{x \in X} s(x, M, F) - \sum_{y \in Y} s(y, M, F)$
WEFAT	OCLC model English Gutenberg embeddings	$s(w, M, F) = \frac{mean_{m \in M} cos(\vec{w}, \vec{m}) - mean_{f \in F} cos(\vec{w}, \vec{f})}{std\_dev_{x \in M \cup F} cos(\vec{w}, \vec{x})}$

Table 5.3: Summary of the Methods for Sub-RQ1

5.2.5 Expectations and Limitations

Table 5.3 gives an overview of all the methods that will be used to answer the first sub-research question. For the ranked lists, differences in gender association are expected in all categories. Girls and boys are probably described with different adjectives. According to [10, 23, 24], there is a difference in the type of activities boys and girls engage in. Boys tend to undertake active and dominant actions, whereas girls are taking part in passive or dependent activities. Ranked lists targeted to verbs are hence expected to yield a difference for the genders. The top ten occupations are also expected to differ between the genders. According to Weitzman et al. [10], women tend to be depicted as housewives whereas men are engaged in various occupations. This observation is expected to be reflected in the word embeddings by showing differences in ranked occupations.

When it comes to the analogies, output pairs are expected to have some similarity with the results produced by Bolukbasi et al. [27]. The occupations of adults in the children's literature

are highly stereotyped [10], so the analogies will likely produce stereotyped occupation pairs. Also when it comes to activities or adjectives, the analogies are expected to show gender bias in alignment with the results of literature [27]. There might also be an overlap between the output pairs and the ranked lists.

Applying the WEAT to children's books will establish whether the current lists of WEAT are appropriate to discover gender bias in literature addressing young readers. It is expected that many of the words in the WEAT lists are out of vocabulary words for the embeddings trained on children's literature or do not occur very often. This is because children's literature uses a different vocabulary than adult literature and covers different topics and stories. This leads to a shift in the vocabulary used in the books. For example, Caliskan et al. [26] use the following target lists for measuring gender bias towards family and career attributes respectively: 1) `home, parents, children, family, cousins, marriage, wedding, relatives`; 2) `executive, management, professional, corporation, salary, office, business, career`. While the words from the former list are likely to occur in many children's books, the vocabulary from the latter is not often encountered in children's literature. Salaries are not something that children are concerned with and parents do not tend to discuss this with their children, so this word is not likely to be in stories addressed to a young audience.

While the WEFAT results of Calsikan et al. [26] showed that bias in word embeddings linearly correlates with the percentage of females in the real world, the expectations are that this will not be the case for children's literature. Social science findings suggest that the vision of children's books on occupation is very much skewed [10, 24]. Women are mostly portrayed as mothers, homemakers or in a typically female profession like teacher or nurse. In return, men are portrayed pursuing a wide variety of careers.

5.3 Sub-Research Question 2: Adaption of Methods to Children's Literature

The second sub-research question (Sub-RQ2) is: *To what extent can the WEAT gender bias categories be adapted to children's literature?* The question will be answered through the use of several different methods. On the one hand, methods from Table 2.2 for automatic WEAT detection are used. On the other hand, hand-made lists are made and tested on the OCLC word embeddings.

5.3.1 Automatic WEAT Detection

New WEAT bias categories are created using automatic detection. For this, an adaptation of the Unsupervised Bias Enumeration (UBE) algorithm of Swinger et al. [34] is used. The UBE is an extension of the clustering method of Chaloner and Maldonado [31]. The latter is simply using clustering to come up with new WEAT target lists, and those clusters are not very semantically coherent [31]. The UBE algorithm, in return, uses Voronoi partitions on the clusters to create more coherent categories. However, the UBE algorithm needs some adjustments to the task at hand, since the attributes are not a list of names in this research but two sets of gendered terms M and F. Instead of clustering attribute names, the attribute lists of the WEAT test [31] are used. When it comes to the target words, the steps of the UBE algorithm are followed: 1) clustering of the M most-frequent words in the embeddings using K-means++ clustering yielding categories $A_1, ..., A_m$; 2) partitioning of the words in the clusters into Voronoi sets $V_{ij} \subseteq A_j$ depending on the proximity to the gender attributes $\overline{G_i}$.

$$V_{ij} = \left\{ w \in A_j \mid i = \underset{i' \in [n]}{\arg\max}\, \overline{w} \cdot \overline{G_{i'}} \right\} \tag{5.7}$$

The output of the UBE are A_{ij} defined as the t words maximising:

$$\max_{w \in V_{ij}} (\overline{G_i} - \mu) \cdot (\overline{w} - \overline{A_i}) \tag{5.8}$$

The parameter t was chosen to be eight as that was also the default number of words per WEAT target list in the research of Chaloner and Maldonado [31]. Clusters that were smaller or did not yield eight male and female associated words were excluded. The hypothesis test is adapted to the approach of [31]. Instead of running the UBE rotational hypothesis test, the WEAT permutation test with a limit on the iterations is used ([31] use a limit of 1,000 iterations). Table 5.4 shows the chosen as well as the default UBE input parameters.

Variable	Meaning	Default	Chosen
WE	Word embeddings	w2v	w2v
X	Attribute list	Names	Genders
n	Number of attribute groups	12	2
m	Number of categories	64	64
M	Number of frequent words	30.000	30.000
t	Number of words per WEAT	3	8
α	False discovery rate	0.05	0.05

Table 5.4: UBE Inputs

This algorithm is fitted on the OCLC embeddings to cluster and partition the words into WEAT lists. However, preliminary analysis showed that this did not result in semantically coherent WEAT lists - an issue that Swinger et al. [34] and Chaloner and Maldonado [31] also faced. To increase the quality of the word embeddings and hence clustering performance, the OCLC embeddings are postprocessed using the methodology proposed by Mu et al. [1]. As described in Figure 5.1, from each word vector $v(w)$, the mean vector μ of all words in the embeddings is subtracted. The remaining vectors $\tilde{v}(w)$ are subject to a Principal Component Analysis (PCA), keeping only the top $D = 3$ components. These are subsequently subtracted from the vectors $\tilde{v}(w)$, which don't include the mean anymore. The algorithm outputs the postprocessed vectors $v'(w)$ which are stripped of the influence of the mean and the top 3 components. This is done to extract the influence of common words like stop words on the embeddings, which co-occur with many words but do not carry semantic meaning.

Algorithm 1: Postprocessing algorithm on word representations.

Input : Word representations $\{v(w), w \in \mathcal{V}\}$, a threshold parameter D,

1 Compute the mean of $\{v(w), w \in \mathcal{V}\}$, $\mu \leftarrow \frac{1}{|\mathcal{V}|} \sum_{w \in \mathcal{V}} v(w), \tilde{v}(w) \leftarrow v(w) - \mu$

2 Compute the PCA components: $u_1, ..., u_d \leftarrow \mathrm{PCA}(\{\tilde{v}(w), w \in \mathcal{V}\})$.

3 Preprocess the representations: $v'(w) \leftarrow \tilde{v}(w) - \sum_{i=1}^{D} \left(u_i^\top v(w)\right) u_i$

Output : Processed representations $v'(w)$.

Figure 5.1: Mu et al.'s [1] Algorithm Used for WE Postprocessing

Subsequently, the postprocessed embeddings are evaluated as was done in Section 5.1. The results of the embeddings evaluation of the postprocessed WE can be found in Table 5.5. As can be seen, the performance slightly increased for adjusted analogies corpus, remained the same for the larger analogies corpus and stayed significant for the word pairs test. While this does not indicate great increase in performance, the quality of the word embeddings seemed to

have increased when evaluating the created clusters. These postprocessed word embeddings are hence used to automatically detect WEAT lists.

Dataset	Min count	Vocabulary size	Vector size	Evaluation		
				Word pairs	Analogies	Analogies adj.
English OCLC	10	41,258	300	**0.49**	0.29	0.37
English OCLC postprocessed	10	41,258	300	**0.48**	0.29	0.38

* bold indicates significance with $\alpha = 0.05$

Table 5.5: Evaluation of Postprocessed WE

5.3.2 Hand-Made Weat Lists

Next to automatically generated WEATS, new WEAT lists are constructed manually. The lists are based on knowledge gained from the social science literature, the findings of Sub-RQ1 and the results of the automatically created WEAT lists. For example, the type of activities children engage in or the professions the characters occupy are used as inspiration. Moreover, the results from the ranked lists and analogies are useful in creating new WEAT bias categories. This way, the WEAT lists are tailored to children's books to account for difference in language use.

5.4 Sub-Research Question 3: Change of Gender Bias Over Time

The third sub-research question (Sub-RQ3) is about the possibility of word embeddings changing and adapting to modern times: *How does gender bias in children's books change over time?* For this, the English Gutenberg embeddings are drawn upon, to display the difference in language over time. As the mean publication date for books in the English Gutenberg corpus is 1876, it stands in contrast to the OCLC corpus with its mean publication date of 2006. In addition, extra word embeddings are trained per decade on the English Gutenberg corpora. The Sub-RQ3 is answered using a combination of different methods.

5.4.1 Ranked Lists Over Time

At first, Ranked Lists of the most biased words in the English Gutenberg embeddings are calculated using the method of [27]. As was the case for Sub-RQ1 (Section 5.2.1), the lists reflect the ten adjectives, occupations, nouns, verbs and animals most associated with each gender. Not only are the lists expected to differ per gender, but also in comparison to the OCLC embeddings. Garg et al. [30] describe a shift in adjectives describing women from charming and sweet in the 1910s to maternal, protective and emotional in the 1960s. Similar results, tailored to girls and boys, are expected as results of the research.

5.4.2 Analogies Over Time

Subsequently, gender bias in the English Gutenberg embeddings is measured using the analogies task from [27] as was done in Sub-RQ1. The analogies for the seed pair *she-he* are expected to differ from the results in Sub-RQ1 to the point that they may include more old-fashioned words like *gown* or *master* instead of the modern equivalents *dress* and *mister*.

5.4.3 WEAT Over Time

Next, the WEATs are tested on the English Gutenberg embeddings. First, the WEATs as used in Sub-RQ1 (see Chapter 6) are tested on the embeddings along with the hand-made WEATs as established in Sub-RQ2 (see Chapter 7). The results are compared to the relevant literature and the findings of the OCLC embeddings.

Second, embeddings are trained per decade of the English Gutenberg corpus. The gender bias score (equation 5.3) of the adjective target list in respect to the gendered attribute lists is calculated. The bias scores of the adjectives for each decade are then set in relation to each other using the Pearson correlation coefficient, replicating the approach of Garg et al. [30]. In line with their findings, it is expected that the decades closer to each other have a higher correlation than decades further apart. Furthermore, historical events that could heavily influence gender bias like the women's movement in the 1960s are expected to be shown in the correlation matrix.

5.4.4 WEFAT Over Time

Lastly, the WEFAT is repeated on the English Gutenberg embeddings. However, a different set of occupation data is needed to compare the gender bias to as the data of BLS 2015 data [54] is not applicable to a corpus with a mean publication date of 1876. For this, the 1851 Census Report [56] is drawn upon. In the report, occupational data per county of the United Kingdom (UK) is collected. The data includes the amount of males and females working in 458 professions separated into age groups with a range of five years. To get an overview of the UK as a whole, the number of male and female workers are summed across all counties and age groups, resulting in a dataset that gives the total number of males and females working in a certain professions. The bias scores of professions are then set in relation to the percentage of women in the respective professions, with the expectation of a roughly linear relationship between the two variables. The results are expected to be of similar relationship as the OCLC embeddings and the BLS data.

5.4.5 Expectations and Reasoning

To sum up, the methods are expected to produce different results for the OCLC embeddings and English Gutenberg embeddings. On the one hand, they will differ because of the types of texts they include. The OCLC embeddings were trained on short descriptions of books, which contain less information than the full-text books from the English Gutenberg embeddings. Especially the OCLC embeddings are expected to have many out-of-vocabulary words for the WEAT as it is trained on the descriptions of children's books only. Those are very short and limited in their vocabulary use. On the other hand, the two language models will differ in their gender bias as they cover different periods. The Gutenberg embeddings contain mostly books from 1850-1925 whereas the OCLC embeddings are mostly trained on books that were published after 2000. As established in Chapter 3, gender bias is expected to change over time. Hence, the two embeddings are likely to contain different kinds of biases.

5.5 Sub-Research Question 4: Gender Bias Across Cultures

For sub-research question 4 (Sub-RQ4), a cross-cultural approach is being taken: *How does gender bias differ between English and German children's books?* Inspired by the work of Kurpicz-Briki [33], who established that the WEAT is suitable to find gender bias in German embeddings, the question is answered using a selection of methods from Sub-RQ1 and the German datasets DNB and German Gutenberg. The results from the analysis are compared

with the results from Sub-RQ1. The analysis includes: 1) ranked lists of the top ten nouns, adjectives, verbs and animals; and 2) the translated WEATs from Table 5.2 as well as the manually crafted WEATs from Sub-RQ2. For the former, the gender direction is established by averaging and subtracting the vectors of gendered words as was done in Sub-RQ1. The following lists were used to establish the gender direction:

- male = er, sohn, sein, ihm, vater, mann, junge, männlich, bruder, söhne, väter, männer, brüder, onkel, neffe, neffen, cousin, cousins, papa, papas
- female = sie, tochter, ihres, ihr, mutter, frau, mädchen, weiblich, schwester, töchter, mütter, frauen, schwestern, tante, nichte, nichten, cousine, cousinen, mama, mamas

This gender direction is set in relation to nouns, adjectives, verbs and animals lists. For this, the lists from Sub-RQ1 are translated using DeepL [57] and corrected by hand where possible. Alternatives for the verb list were found on [58]. Similarly, for the WEAT lists, the English target and attribute lists were translated using DeepL [57] and manually corrected. Translation was also used by Kurpicz-Briki [33] to find gender and origin bias in German and French embeddings. The translated lists can be found in Appendix G.

It is expected that there is a difference between the results from the English and the German embeddings. On the one hand, gender bias differs per culture. With language being a vital part of a culture, the former is anticipated to capture societal gender bias. Hence, a difference between the two languages is expected. On the other hand, German is a language with grammatical gender. Most occupations have a male and female version, making them naturally gender-biased. At the same time, the male version tends to be used more often than the female version and when the plural is used, it is usually the male plural, even when addressing a group of male and female professionals. Hence, it could be that some gendered occupations still show more bias to one of the genders.

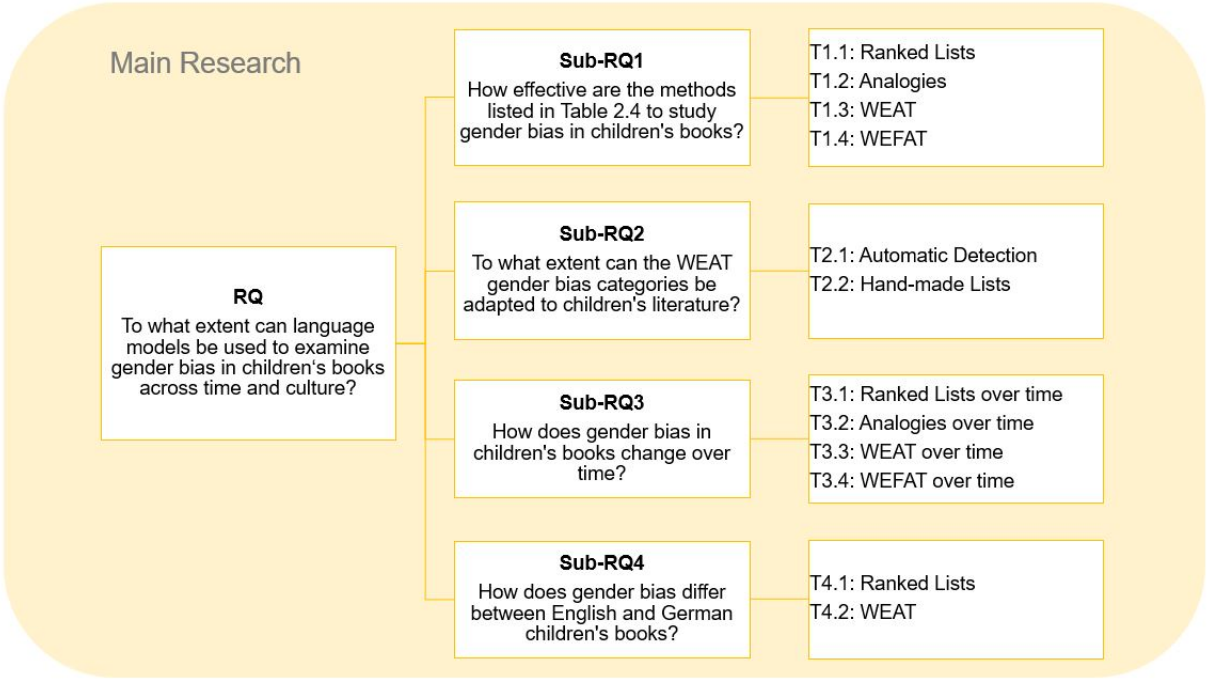


Figure 5.2: Summary of the Methodology

6 APPLICABILITY OF CURRENT METHODS TO NEW DOMAINS

The first sub-research question (Sub-RQ1) is: *How effective are the methods listed in Table 2.2 to study gender bias in children's books?* It is answered using Ranked Lists, Analogies, WEAT and WEFAT. The methods are applied to the OCLC word embeddings (OCLC WE), as these are comparable in language and time to the literature establishing the methods. In the following, the results of each of the methods will be presented and discussed.

6.1 Ranked Lists

First, the children's books are analysed in regard to their most biased words using Ranked Lists. The Ranked Lists include the top ten adjectives, verbs, occupations, nouns and animals connected to each gender. These word lists were set in relation to the gendered group vector consisting of gendered pronouns and family relations, the so-called gender-direction. For each word list, the top 15 biased words for the OCLC embeddings were calculated. These are presented in the following sections. While Bolukbasi et al. [27] excluded gender appropriate associations like `fireman` or `businesswoman`, this is not done for the results of this thesis. This is the case because excluding gender appropriate associations gives a skewed vision on the word embeddings, giving the impression that they are highly biased. However, by leaving them in, a more fair picture can be painted of how biased the embeddings are.

6.1.1 Professions

Table 6.1 shows the results of the projection of professions on the gender direction for the OCLC embeddings as well as the results of Bolukbasi et al. [27]. The list of Bolukbasi et al. [27] was filtered and does not contain gender appropriate associations, which limits possibilities for comparison. Moreover, the words in the list were set in relation to the *she-he* axis rather than the overall gendered direction. This can lead to different words being associated with each gender.

General trends for female professions in the embeddings of [27] seem to be home-related (`homemaker` and `housekeeper`), care-related (`nurse` or `nanny`), fashion-related (`socialite`, `hairdresser` and `stylist`) or of organisational nature (`receptionist`, `librarian` and `bookkeeper`). The OCLC embeddings, in return, show most association of art-related professions (`ballerina, dancer, actress, choreographer, singer, understudy` and `pianist`) with the female direction. Besides that, home- and care-related professions (`maid, nanny, housekeeper`) show association with the female direction. Also the fashion-related professions from [27] can be found back in the OCLC embeddings' associations.

When it comes to the association of professions with the male direction, Bolukbasi et al.'s [27] results do not show such clear themes as was the case with the female axis. There are marine-related (`skipper` and `captain`) and arts-related professions (`maestro` and `magician`). Besides that, the professions can hardly be grouped. For the OCLC results, several directions can be identified. Many of the professions relate to agriculture and food provision

(farmer, fisherman, butcher) and to war/army (commander, colonel, lieutenant, sergeant, soldier). Some relate to marine professions (sailor and captain), law enforcement (ranger and policeman) and religion (preacher). However, even these professions are much more varied and form only loose groups.

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

(a) Results of Bolukbasi et al. [27]

Extreme *female*		Bias	Extreme *male*		Bias
1.	ballerina	0.410	1.	fireman	-0.374
2.	maid	0.305	2.	ranger	-0.317
3.	dancer	0.302	3.	farmer	-0.315
4.	actress	0.282	4.	commander	-0.304
5.	hairdresser	0.232	5.	policeman	-0.290
6.	socialite	0.209	6.	fisherman	-0.288
7.	nanny	0.204	7.	colonel	-0.287
8.	stylist	0.203	8.	preacher	-0.281
9.	valedictorian	0.198	9.	lieutenant	-0.279
10.	planner	0.181	10.	captain	-0.275
11.	choreographer	0.176	11.	sailor	-0.268
12.	singer	0.166	12.	sergeant	-0.267
13.	understudy	0.155	13.	butcher	-0.264
14.	housekeeper	0.147	14.	soldier	-0.258
15.	pianist	0.141	15.	foreman	-0.252

(b) Results of the OCLC WE

Table 6.1: The Most Extreme Professions in the OCLC Embeddings as Projected on the Gender Direction

Overall, the trend can be observed that females fulfill roles that are about nurturing others and caring for home and the people surrounding them, while males occupy professions related to providing, enforcing and protecting. It is interesting to notice that the results coincide with findings from the social science literature. Weitzman et al. [10] and Shahnaz, Fatima and Qadir [24] found males engaging in a wide variety of professions while females were depicted mostly as wives, mothers and homemakers. This can also be seen in the most associated professions with the gender direction. For the female occupations, trends can clearly be identified while for the male ones, this is harder, indicating a wider variety of professions for males. Another interesting observation is that the bias scores are higher for the male direction than the female direction. This shows that the association of the professions with the male direction are, in general, higher than for the female direction. This connects to the previous point, that men tend to be ones working in the children's stories, while women are simply mothers and wives, roles that are not seen as professions and hence not included in the profession lists.

6.1.2 Nouns

In Table 6.2, the nouns most associated with the gender direction in the OCLC embeddings can be found. The female-associated words are not very diverse and give rise to only two clusters. A general theme in the most female associated words is related to fashion and gossip (magazine, drama, designer, skirt and dress). Besides that, there are many gender appropriate words for females like miss, lady, girl, she and queen.

The words associated most with male direction in the OCLC embeddings are much more diverse than the female direction as was the case with the professions. Within the ranked list, themes can only hardly be found. There are transport-related words (tank, truck, engine and gas) and war-related words (sir, tank and general). However, these seem to form more lose groups. Another difference worth pointing out is the fact that the female direction has quite some gender appropriate association, while this is less the case for the male direction.

Extreme *female*		Bias	Extreme *male*		Bias
1.	magazine	0.344	1.	sir	-0.437
2.	drama	0.303	2.	tank	-0.390
3.	girl	0.301	3.	uncle	-0.364
4.	queen	0.300	4.	bill	-0.350
5.	designer	0.293	5.	buddy	-0.347
6.	skirt	0.291	6.	truck	-0.328
7.	dress	0.286	7.	son	-0.316
8.	lady	0.282	8.	farmer	-0.315
9.	wedding	0.275	9.	engine	-0.306
10.	miss	0.272	10.	chip	-0.305
11.	sweet	0.260	11.	hook	-0.294
12.	beautiful	0.248	12.	dump	-0.285
13.	flower	0.247	13.	baseball	-0.282
14.	formal	0.240	14.	general	-0.281
15.	she	0.238	15.	gas	-0.279

Table 6.2: The Most Extreme Nouns in the OCLC Embeddings as Projected on the Gender Direction

6.1.3 Adjectives

The results for the most associated adjectives can be found in Table 6.3. In the OCLC embeddings, the female direction is associated with adjectives such as `glamorous`, `graceful`, `fancy`, `enchanting`, `fabulous`, `gorgeous`, `dazzling` and `elegant`. These adjectives present the female as something to long for and look up to. At the same time, they are superficial, focusing only on the appearance of girls and women. This can be seen as an objectification of females. This might give children the impression that all females are valued for their looks and demeanor rather than their skills.

For the adjectives associated with the male direction in the OCLC embeddings, there are many different adjectives and trends can only be identified with difficulty. Males are seen as `swift`, but also as being in a bad mood (`growling`, `miserly`, `frank` and `snarling`) and cannot be trusted (`slippery`). On the other hand, they are `trusty` themselves as well as `nifty` and `jolly`. Hence, there are great contradictions in the associations. The adjectives could stem from the portrayal of males as fathers in the children's books. As pointed out in the previous section on professions, the men tend to be the ones working outside of the house and they hence may be in a bad mood when coming home, complaining about their workday. At the same time, fathers are also the ones to play ball with as they are also `swift`, `big` and `jolly`. The image of males as being in a bad mood could also come from villains. The *bad guy* is, as the phrase indicates, usually male. So many negative adjectives associated with the male direction could come from this association.

All in all, the trend continues that the female direction shows a theme in its association while the male direction is characterised by diversity. This may be restricting for children. Especially girls might feel restricted in their development as characters as there is no diversity in their expected gender adjectives. Also they are mostly expected to pay attention to their looks rather than develop their skills. If the interests of a girl do not include her looks but lie in a different area, she can feel pressured to spend more time on clothes, hair and makeup even though that is not enjoyable to her. She might also encounter obstacles when not following this gender role description which can lead to frustrations. In return, the male direction shows more diversity in characteristics, allowing boys a greater variety in development. However, boys can also

be restricted by this portrayal of gender roles as being nurturing and loving are considered female features. This can lead to anxiety when trying to show emotions and over-expression of masculinity [6].

Extreme *female*		**Bias**	**Extreme *male***		**Bias**
1.	glamorous	0.431	1.	swift	-0.381
2.	graceful	0.289	2.	general	-0.281
3.	bubbly	0.286	3.	growling	-0.280
4.	fancy	0.284	4.	trusty	-0.278
5.	pink	0.283	5.	chief	-0.270
6.	enchanting	0.282	6.	giant	-0.264
7.	fabulous	0.280	7.	snarling	-0.253
8.	sparkling	0.275	8.	miserly	-0.249
9.	gorgeous	0.272	9.	frank	-0.248
10.	glittering	0.261	10.	mammoth	-0.240
11.	spirited	0.261	11.	jolly	-0.230
12.	sweet	0.260	12.	grizzled	-0.229
13.	dazzling	0.252	13.	nifty	-0.225
14.	elegant	0.250	14.	slippery	-0.218
15.	lovely	0.248	15.	flat	-0.216

Table 6.3: The Most Extreme Adjectives in the OCLC Embeddings as Projected on the Gender Direction

6.1.4 Verbs

According to the verbs most associated with the female direction in the OCLC embeddings, females are still associated with being caring (`embraces, adores` and `charms`). Moreover, they are related to fashion (`dresses, sparkles, shines` and `threads`). This is in line with the findings from the adjectives list, where females are objectified and reduced to their looks. As Sap et al. [39] researched, women use low-agency verbs which attributes them an aesthetic role rather than one contributing to the plot of story. However, the verbs are more diverse than the previous word categories and also include negative actions such as `violates, resents` and `pricks`. Considering the power and agency framework of [39], some verbs in the OCLC embeddings give the female agency and power over others such as `juggles` and `violates`, diminishing the effects of objectification. However, it is unclear whether the women are the ones violating or being violated. When females are the subject of violation, the power is reversed and lies with the violator, rather than the victim.

On the male direction, many of the verbs relate to physical actions such as `bangs, bites, shovels, fumbles` and `smashes`. Some also related to marine actions again (`ships` and `sails`), as was the case for the professions. A difference between male and female association is that males have physical actions requiring strength while females have actions that are about (physical) caring or that are non-physical, e.g. `immersing` and `shining` (see Table 6.4). This paints a picture of the strong and protecting male and the caring and good-looking female.

6.1.5 Animals

Given the findings of social science literature on the great disparity of the genders in animal characters, a list with animals was tested as well. In the OCLC embeddings, mostly birds were associated with the female direction (`swan, quetzal` and `nightingale`). Another group of

Extreme *female*		Bias	Extreme *male*		Bias
1.	dresses	0.306	1.	associates	-0.332
2.	sparkles	0.276	2.	bones	-0.302
3.	embraces	0.231	3.	bangs	-0.267
4.	violates	0.219	4.	bites	-0.266
5.	adores	0.216	5.	rockets	-0.257
6.	chimes	0.207	6.	tires	-0.255
7.	charms	0.205	7.	rounds	-0.252
8.	resents	0.200	8.	huffs	-0.242
9.	twirls	0.195	9.	shovels	-0.238
10.	shines	0.188	10.	zooms	-0.237
11.	juggles	0.186	11.	ships	-0.236
12.	immerses	0.185	12.	retires	-0.231
13.	pricks	0.185	13.	sails	-0.230
14.	whirls	0.180	14.	fumbles	-0.230
15.	threads	0.176	15.	smashes	-0.227

Table 6.4: The Most Extreme Verbs in the OCLC Embeddings as Projected on the Gender Direction

animals are insects (`ladybird` and `butterfly`) and cat and dog breeds (`poodle`, `ragdoll`, `pekingese`, `persian` and `siamese`).

It is noteworthy that two of the animals are also female names: `molly` is a small fish and `tiffany`, a long-haired cat breed. This overlap could explain the high association scores for these two animals. Moreover, social science literature noted that birds tend to be presented as mothers, as in the book *Have you seen my duckling?* for instance, where a duck is looking for her child. This would explain the association of birds with the female direction. In addition, `coral` might be associated with the female gender because it is worn as jewellery a lot, rather than it being a female animal in children's books. Similarly, `ladybird` might be associated with females as the word *lady* is in it.

On the male direction of the OCLC embeddings, it is even harder to identify clear groups and the most associated animals seem rather random. However, the association with the male direction is much higher than with the female direction. One reason for this could be that most animals in the stories are male as indicated by social science literature [10, 23, 25]. This great disparity in animal characters could be reflected by these association lists.

When setting these results in relation to the results of the adjective Ranked Lists, it can be seen that results seem to coincide. On the one hand, birds are usually small and are seen as beautiful, adjectives that are also used to describe females in Table 6.3. On the other hand, males are associated with animals that are big or dangerous, adjectives that are associated the male direction.

6.2 Analogies

Second, the descriptions from the OCLC data were analysed using the Analogies task from Bolukbasi et al. [27]. The presented results were taken from the top 100 generated analogies and exclude analogies containing names or words from the gender lists (pronouns and family relations). Analogies containing names caused a great amount of noise in the results as they represented the majority of the generated analogies. Although they are gender appropriate they do not contribute to the knowledge base about gender bias in children's literature, making it hard to find more relevant analogies. The results from OCLC were set in relation to the results

Extreme *female*		Bias	Extreme *male*		Bias
1.	tiffany	0.221	1.	raccoon	-0.370
2.	swan	0.207	2.	beaver	-0.349
3.	ladybird	0.200	3.	camel	-0.347
4.	butterfly	0.135	4.	catfish	-0.324
5.	nightingale	0.134	5.	rat	-0.321
6.	quetzal	0.130	6.	bullfrog	-0.314
7.	molly	0.093	7.	moose	-0.311
8.	coral	0.082	8.	mule	-0.297
9.	poodle	0.051	9.	coyote	-0.296
10.	ragdoll	0.048	10.	lizard	-0.295
11.	pekingese	0.039	11.	buffalo	-0.292
12.	persian	0.037	12.	hyena	-0.291
13.	horse	0.036	13.	snake	-0.290
14.	quokka	0.035	14.	crocodile	-0.288
15.	siamese	0.034	15.	squid	-0.286

Table 6.5: The Most Extreme Animals in the OCLC Embeddings as Projected on the Gender Direction

from Bolukbasi et al. [27]. Yet, comparability may be limited as they [27] filtered the analogies to contain only stereotypical ones, whereas this is not the case for the results of this research.

When looking at the results in Table 6.6, it can be seen that many gender stereotypical analogies are about social relations. Females relate to `bffs` (best friends forever) the way males relate to `girlfriends`, `teammates` and `partners`. When girls have `exboyfriends`, boys have `archenemies`. When it comes to their social roles as teenagers, girls are `cheerleaders` and boys are `troublemakers`. Another group of gender stereotypical analogies is about sports: females do `gymnastics` and play `netball` while males play `football` and `baseball`. While this might give rise to concern, one must also note that there are many gender appropriate analogies generated through the embeddings, e.g. `girlhood-boyhood`, `cowgirl-cowboy` and `queen-king`. This shows that the embeddings are also able to find appropriate relations.

When comparing the analogies generated from OCLC embeddings to the results of Bolukbasi et al. [27], it can be seen that the topics that are covered are different. The analogies generated by Bolukbasi et al. [27] contain many professions which are not covered by the results produced in this thesis. This could come from the difference in texts underlying the OCLC embeddings and the Google News embeddings, which were used by Bolukbasi et al. [27]. The latter is likely to contain many professions as news stories tend to contain the jobs of the people they are about, while the former is about children's stories, which are usually not about jobs as children are not concerned with this topic. Nevertheless, the difference between sports is reflected in both embeddings given the analogies `volleyball-football` and `softball-baseball`. Hence, to some extent, the results of Bolukbasi et al. [27] could be reproduced.

6.3 Word-Embedding Association Test

Third, the Word-Embedding Association Test was fitted on the children's books of the OCLC. The results in Table 6.7 show the p-value and Cohen's ds for the five WEATs from Table 5.2. The results of the OCLC embeddings are reported and set in relation to the results of Caliskan et al. [54] as well as Chaloner and Maldonado [31]. P-values in bold indicate statistically significant gender bias for $p < 0.05$.

	Analogies
1.	sewing-carpentry
2.	nurse-surgeon
3.	blond-burly
4.	giggle-chuckle
5.	sassy-snappy
6.	volleyball-football
7.	registered nurse-physician
8.	interior designer-architect
9.	feminism-conservatism
10.	vocalist-guitarist
11.	diva-superstar
12.	cupcakes-pizzas
13.	housewife-shopkeeper
14.	softball-baseball
15.	cosmetics-pharmaceuticals
16.	petite-lanky
17.	charming-affable
18.	lovely-brilliant

(a) Analogies from Bolukbasi et al. [27]

	Analogies	Score
1.	girlhood-boyhood	0.565
2.	bff-girlfriend	0.496
3.	bffs-teammates	0.474
4.	ladyinwaiting-valet	0.473
5.	cowgirl-cowboy	0.466
6.	resigns-redeems	0.464
7.	queen-king	0.455
8.	gymnastics-football	0.447
9.	netball-baseball	0.428
10.	hen-rooster	0.428
11.	skirt-shirt	0.427
12.	witch-wizard	0.427
13.	heiress-schoolboy	0.426
14.	empress-squire	0.422
15.	womanhood-manhood	0.412
16.	heroine-hero	0.410
17.	exboyfriend-archenemy	0.410
18.	cheerleader-troublemaker	0.408

(b) Analogies from the OCLC WE

Table 6.6: Analogies for the *She-He* Axis

The OCLC embeddings shows significant gender bias for W1, W2 and W3 only. This means there is a significant relation between `career`, `maths`, `science` and `male` and between `family`, `arts` and `female`. The conventional small, medium and large values of Cohen's d are 0.2, 0.5 and 0.8, respectively. Keeping this classification in mind, the effect sizes are large according to their d values. However, upon close inspection it can be seen that W1 is barely significant and its Cohen's d is much lower than for the other models. This might indicate a weaker relationship between career and male as well as family and female. One reason for this might be that the career words are not used a lot in the corpus, hence the relationship is not that strong. Yet, given the conventional interpretation of Cohen's d, a value of 0.839 is still large. Test W4 and W5 do not show significant gender bias, indicating that `intelligence, appearance, strength` and `weakness` do not have a stronger association with one of the genders. This might be because the biases in W1 - W3 are more deeply rooted in society than the biases targeted in W4 and W5.

	Caliskan [54]		Chaloner and Maldonado [31]		OCLC	
WEAT category	*p*	*d*	*p*	*d*	*p*	*d*
W1: career vs. family	**0.001**	1.81	**0.001**	1.37	**0.045**	0.839
W2: maths vs. arts	**0.018**	1.06	**0.017**	1.02	**0.012**	1.078
W3: science vs. arts	**0.010**	1.24	**0.004**	1.25	**0.000**	1.446
W4: intelligence vs. appearance			**0.000**	0.98	0.370	0.114
W5: strength vs. weakness			**0.006**	0.89	0.643	-0.152

Table 6.7: Results of the WEAT Tests on the OCLC WE. P-Values in Bold Indicate Statistically Significant Gender Bias ($p < 0.05$).

6.4 Word-Embedding Factual Association Test

Fourth, the OCLC embeddings were tested on their gender bias using the Word-Embedding Factual Association Test. The gender biases of 50 professions were calculated and set in relation to the percentage of female workers in that occupation. When plotting the association score of the tested professions against the percentage of females in the workforce, a linear correlation can be seen (see Figure 6.1). When the percentage of women is below 50%, then the effect size tends to be negative. In return, when the percentage of females in the occupation is above 50%, the effect size is positive. At equal percentages of men and women, the effect size tends to be around the origin. This shows the validity of the results, as they are not skewed. Given the conventional interpretation of the Pearson's correlation coefficient[1], the correlation is largely positively linear with $r = 0.689$ and a significant p-value of $p < 10^{-6}$.

Nonetheless, there are also outliers. For example, `hygienist` has 96.4% female workers but the gender bias score is -0.532 (dot on the right lower corner of Figure 6.1). This would indicate a strong male bias, yet the professions is mainly done by women. Similarly, `clerk` has 69.5% female workers though the gender bias is -0.87.

Figure 6.1: Occupation-Gender Association in the OCLC Embeddings. Pearson's Correlation Coefficient $r = 0.689$ With $p < 10^{-6}$.

6.5 Conclusion and Discussion

In conclusion, the methods from Table 5.3 can largely be used to find gender bias in children's descriptions as they revealed interesting and significant results for the OCLC embeddings. The Ranked Lists showed varying association for the genders with clear and identifiable trends. Females were continuously associated with being caring and nurturing while a focus on their appearance remained prominent. This focus on inter-personal relationships and looks can be restricting for girls who are more independent and skills-oriented. These characteristics are more male-associated and can lead to obstacles for girls trying to display more masculine characteristics and behaviours. In return, the male direction was shaped by diversity, allowing boys more variety in characteristics to adopt. Little overlap can be seen between the two genders directions, clearly marking certain behaviours as male and female. Although variety is present in the male direction, boys trying to showcase female attributes might receive negative feedback from their environment, confining them to over-expressions of masculinity [6].

Nevertheless, the Ranked Lists also exhibit certain shortcomings in their methodology and implementation. One such issue stems from the target lists used in this research. The dif-

[1] $r = 0$ means no correlation, $r = 1$ means perfect, positive, linear correlation and $r = -1$ means perfect, negative, linear correlation

ferent lists were not exclusive and contained questionable words. For example, the words `beautiful, neat` and `soft` are in the nouns and adjective lists, even though those are not nouns. This lead to overlap in the results between the lists and somewhat unreliable results. Connected to this, issues with the Part-of-Speech (POS) of words emerged. For example, in the Ranked Lists of nouns and adjectives, females were associated with the word `sweet`. However, probably females are mostly associated with sweet as an adjective and not as a noun. Yet, this difference in POS cannot be taken into account with the Ranked Lists.

The Analogies showed that many of the associations are gender appropriate, however some can also be classified as stereotypical. This gives an idea of how biased the embeddings are and which biases have the highest score. Nonetheless, the results also included a lot of noise in the form of names. Roughly 80% of the generated analogies were male and female names. While these made sense, they do not contribute much to the knowledge base about gender bias in children's books and cause noise. This then requires manual inspection and sorting of the results, which is time consuming.

When it comes to the WEATs, W1-3 were significant, indicating gender bias when it comes to *career*, *family*, *maths*, *arts* and *science* word lists. When taking a closer look at the test, it could be seen that some words were unknown in the embeddings. Out of vocabulary words were `computation` from the `maths` list, `Shakespeare` from the `arts` list, `NASA` and `Einstein` from the `science` list, `judicious` from the `intelligence` list and `blushing` and `voluptuous` from the `appearance` list. Whether a list has out-of-vocabulary words did not seem to affect the test, as both significant and non-significant tests had words that were not in the model. However, it can make the results less reliable as not all the information can be taken into account, limiting possibilities for comparison between embeddings. Moreover, out of vocabulary words can be seen as bias as well if there is a pattern in their absence.

Another shortcoming of this methodology is the overlap between W2 and W3. Both use the arts list as the female target. When arts are closely correlated with the female direction, it is more likely the difference between the two target groups is significant. Moreover, there is a difference in the target list size between W1 - W3 and W4 as well as W5. The former have only eight words per target list, whereas the latter have 25 and 15 words, respectively. This may affect the significance levels of the tests as the clusters of words are bigger.

Looking at the last test, the bias scores of professions correlated with percentages of females in the WEFAT. This shows measurable bias in the OCLC embeddings along real-valued data. Appendix D gives a detailed overview of the tested professions, the out of vocabulary words, the percentage of females in the occupations, their effect sizes and their p-values. Occupations that could not be found in the vocabulary of the embeddings were excluded. It is worthy to point out, that only very few WEFATs were significant and that there were many professions not included in the OCLC embedding vocabulary. This gives rise to doubts about the full applicability of the method to this new corpus. Adjusting the professions list based on literature or automatic clustering might reveal more clear correlations for a greater number of professions.

Overall, the WEAT is only partially able to find gender bias in the OCLC embeddings as only some of the WEATs are significant. Given the results from the Ranked Lists, there is a difference in the adjectives used to describe males and females which is expected to show in the WEAT. Since this is not the case, the method is only partially applicable and needs readjusting to the corpus. An overarching issue seems to be difference in language use between adult's and children's literature. This can be combated by adapting the methods to the vocabulary used in literature addressing young audience, which is explored in the following chapter.

7 ADAPTATION OF METHODS TO CHILDREN'S LITERATURE

Given the limitations of current methods to find gender bias in children's literature, the second sub-research question (Sub-RQ2) asks: *To what extent can the WEAT gender bias categories be adapted to children's literature?* This question is answered using automatic detection and manual creation of WEAT lists. In this chapter, the results of these methods are discussed.

7.1 Automatic WEAT Detection

To automatically detect WEAT lists, the OCLC word embeddings were clustered and partitioned to yield WEAT lists using an adaption of the Unsupervised Bias Enumeration Algorithm (UBE). A selection of the most semantically coherent lists is presented in Table 7.1, while the whole list can be found in Appendix E.

Overall, it can be observed that all clusters yielded statistically significant WEATs with very low p-values ($p < 0.000$). It is interesting to note that the effect sizes were all very large with an average Cohen's d of 1.89. When looking at the clusters' contents, cluster 7 seems to be about celebrations, where the X Targets are about Christmas and the Y Targets about a child's birthday party. Within the target groups, the words seem to belong to one topic, however, it is more difficult to see the connection between the two target groups. Nevertheless, the association of the X targets with male and of the Y targets with female seems to make sense. `Santa Claus` is male and so is `Rudolph the red nosed Reindeer`. Since the words in the list are all connected to `Santa`, it seems logical for the list to be male-associated. However, this is not necessarily a biased finding. The association of the Y targets with females, in return, is stereotypical. None of the words in the list are inherently female, yet they are female-associated. This list corresponds to my personal stereotypes as well, with girls generally playing with `dolls` a lot, women being the ones receiving `flowers` and baking `cupcakes`.

The Y targets of cluster 15 are very similar, including the singulars `doll` and `cupcake`. Moreover, the list features female pieces of clothing like `dresses` and `tiaras`. Hence, this list seems only partially biased. However, the X targets are quite stereotypical as they only contain two inherently male words: `donald` and `postman`. Besides that, the X targets are about vehicles and war related items. The male bias could be explained through the historical case of men being the ones going to war as e.g. tank drivers and jet pilots. Also boys are the ones stereotypically playing with tanks and guns, while girls are thought to play with dolls. Hence, the results of this cluster are rather biased.

Cluster 22 seems to be about fantasy stories, which are often read by teenagers. The men in these stories are `vikings`, `warlords` and `generals`, while women are `witches`, `fairies` and `goddesses`. This cluster is an interesting one as it is not immediately apparent why it is stereotypical. In fantasy literature and games, men tend to be portrayed as powerful warriors which are victorious but also cruel and violent. The female equivalent, however, are usually wicked and use witchcraft and deception to achieve their goals. This can give the false impression that men are honorable and honest while women are treacherous and false.

In cluster 25, the actions of each gender are portrayed. They seem to correspond with the findings of social science literature, where men and boys are active while women are passive [10]. Similarly, cluster 30 paints a picture of a brave and victorious males while females are delicate and pretty. This is in line with the findings from social science literature and the results from Sub-RQ1.

WEAT	X Targets	Y Targets	p-value	Cohen's d
7	reindeer, santas, nuts, elves, snacks, nest, toys, carrots	dolls, sweet, flowers, cupcakes, hearts, celebrating, baking, decorations	0.000	1.871
15	tank, donald, jet, gun, postman, truck, bulldozer, pickup	cupcake, gown, doll, dresses, sparkly, tiara, pink, necklace	0.000	1.932
22	commander, general, lair, viking, warlord, ferocious, master, savage	witch, fae, faerie, turmoil, sorceress, goddess, rebellion, wicked	0.000	1.897
25	crashes, chases, builds, drives, flies, catches, saves, went	wears, throws, holds, reveals, appears, threatens, introduces, makes	0.000	1.878
30	conquers, bravery, overcomes, humility, testing, endurance, neverending, boredom	glamour, blossoming, bittersweet, delicate, embraces, social, fragile, cultural	0.000	1.918

Table 7.1: Most Coherent Automatic WEAT Lists

7.2 Hand-Made WEAT Lists

Building upon the results of Sub-RQ1 and the automatically detected WEAT lists, new lists were created manually. They take into account the aforementioned results, literature findings and my personal stereotypes. Given $\alpha = 0.05$, six out of nine manually created WEAT lists were significant. The results can be found in Table 7.2.

W6 tested whether males are more associated with the outdoors while females are more connected to indoors. This idea was taken from Weitzman et al. [10], however, the test did not yield a significant difference between the two genders. W7 is about typically male and female toys. According to the literature, the clusters, in particular cluster 7 and 62, and my personal bias, boys tend to play with cars and other vehicles and like to play war, whereas girls play with dolls, pony's and like to dress up. This difference is significant with a very high Cohen's d of 1.737. It is interesting to note, that it was much harder to create the WEAT list of female toys than the list of male toys. One reason for this could be the lacking variety of toys for girls, similarly to the lacking variety of occupations for women. W8 makes a difference between typically male and female sports. These were mostly taken from the ranked lists generated in Sub-RQ1 and yield a significant test with a Cohen's d of 1.550. Hence, this difference in sport preferences is reflected in the OCLC corpus. For W9, the activities and games described in [10] were used as an inspiration to create the lists of more active and more quiet games. The hypothesis was that boys are active and like running around while girls are taught to prefer more quiet and "civilized" activities like reading and drawing. Also this difference is significant with a Cohen's d of 1.259.

In W10, supposedly male and female adjectives from Weitzman et al. [10] were tested. Boys are described as loud, rough and brave, while girls are more tamed, passive and emotional.

However, the test did not yield a significant difference between the two genders, contradicting previous findings from the Ranked Lists. The male and female professions in W11 were taken from the WEFAT dataset. The eight professions with the least percentage of females in the workforce were used as basis for the X target set and the eight professions with the highest percentage of females in the workforce were set as Y targets. In line with the linear relationship of female participation in occupations and female bias in word embeddings, W11 showed significant difference in the association of the targets lists with the attribute lists. Moving to more results from the social science literature, the supposedly male and female adjectives according to Kortenhaus and Demarest [23] were tested in W12. They are similar to the ones from W10 in so far as boys are active and clever and females are passive and emotional. Once again, the test was not significant, confirming the results of W10. In W13, the hypothesis was tested that males tend to be more dominant and instrumental while females are passive and obedient. With a p-value just below 0.05, this difference is reflected in the OCLC word embeddings. However, the effect size for this relation is much lower than for the other WEATs with a Cohen's d of 0.837. Nevertheless, this is still a large value according to the conventional interpretation of Cohen's d. Lastly, the association of traditionally male and female school subjects with the genders is tested in W14. This test was inspired by Kurpicz-Briki [33], who tested for a difference in study choices between males and females in German embeddings. It was hypothesized that boys are more drawn to the sciences and girls to the humanities. Also this test was significant with a high effect size of 1.101.

Together, the hand-made WEAT lists help to understand the gender bias present in the OCLC embeddings. Children specific topics like toys, sports, games and school subjects showed a significant difference between the genders. Males preferred rough playing and sports, as well as active and loud activities. They are dominant and enjoy science subjects. Girls, on the other hand, play with dolls, enjoy aesthetics sports like dancing and are drawn to more quiet activities such as reading and writing. In terms of characteristics, girls are associated with obedient behaviour, creativity and the humanities. The difference in adjectives that was discovered by Weitzman et al. [10] and Kortenhaus and Demarest [23] could not be confirmed by W6, W10 and W12. Hence, these social science findings were not present in the OCLC embeddings. This contradicts previous findings from the Ranked Lists and also W13 on dominance and obedience. One reason for this might be the specific mix of words in the target lists. Some of the words might be clearly female or male associate, while others are not. Joining them in one list could lead to non-significant results. Permutating the adjectives and trying different combinations might yield significant differences after all.

7.3 Conclusion and Discussion

In summary, the automatic detection of WEAT lists using the adapted UBE algorithm [34] yielded significant tests for nearly each cluster with large effect sizes. While these results would indicate a success of the automatic detection of WEAT lists, they need to be examined more closely. Often, the lists are not very semantically coherent within their own target group or across target groups. This can mean the target groups themselves are not coherent or target group X being thematically coherent but seemingly unconnected to the theme of target group Y from the same cluster. The issue remained even though Voronoi partitioning and postprocessing of the word embeddings was done. This leaves questions about the coherence of the underlying clusters and difficult to interpret WEATs. Another limitation of the UBE is the evaluation of produced clusters. While qualitative assessment the automated clusters was done, another researcher might come to different conclusion on the quality of the results. Quantitative measures of evaluation are missing. Related to that, it is not always apparent to the human eye why the target groups are inherently male or female. One reason for this could be the issue of coherence within and across target groups. Another reason could be that they don't correspond to my

personal stereotypes. Moreover, the theory of direct and indirect bias of Bolukbasi et al. [27] might be connected to that. It could be that words form a clusters because they re indirectly biased towards a gender, meaning they are closer to words that high gender bias. For example, `bookkeeper` and `kindergarden` are associated because they are both close to `softball`, a sport that is female associated.

Using the knowledge gained from the literature, Sub-RQ1 and the automatically detected WEATs, new WEATs were drawn up by hand. Many of the manually crafted WEATs were significant with large Cohen's ds, confirming the biases through computation. Drawing up WEATs specifically for a certain corpus prevents the issue of out-of-vocabulary words that was encountered for the standard WEATs. If several words of a WEAT list (usual length is eight words) are unknown in a corpus, the results can get less statistically reliable as less information can be taken into account. Moreover, the presence and absence of words can also yield information on gender bias. For example, if the word `waitress` is present in the corpus but the word `waiter` is not, than this already shows a female bias in the corpus. Moreover, adapting WEATs to a specific corpus helps in finding more relevant biases according to the effect size. The Cohen's ds of the automatically detected WEATs were very high (average size 1.89), but even for the manually created lists, the effect sizes increased in comparison to the standard WEATs of Caliskan et al. [26], Garg et al. [30] and Chaloner and Maldonado [31]. Lastly, these lists served as a computational confirmation or rejection of the biases discussed in social sciences literature and my personal stereotypes. Being able to find biases computationally does not only help to confirm or reject knowledge about gender bias in children's literature but also enhances this knowledge given the large scale of the corpus. The social science findings can be generalized to the larger corpus of children's literature with more confidence with the findings of this research.

WEAT		Targets	p value	Cohens d
W6: outdoor vs. indoor	X	outdoor, outside, nature, garden, tree, backyard, lake, mountain	0.332	0.205
	Y	indoor, inside, kitchen, household, home, sofa, bedroom, bathroom		
W7: male vs. female toys	X	ball, bat, truck, car, gun, bicycle, soldier, blue	0.000	1.737
	Y	doll, barbie, makeup, ballerina, jewellery, pony, dollhouse, pink		
W8: male vs. female sports	X	football, basketball, baseball, soccer, wrestling, rugby, boxing, cycling	0.001	1.550
	Y	volleyball, gymnastics, netball, softball, cheerleader, dance, skating, lacrosse		
W9: active vs. quiet games	X	flies, drives, jumps, climbs, swims, slides, drives, skips	0.004	1.259
	Y	reads, watches, hides, listens, draws, paints, sketches, writes		
W10: male vs. female adjective	X	dirty, untidy, loud, rough, fearless, active, achieving, brave	0.159	0.516
	Y	clean, neat, quiet, gentle, fearful, passive, emotional, caring		
W11: male vs. female professions	X	plumber, mechanic, carpenter, machinist, engineer, programmer, architect, officer	0.000	1.640
	Y	hygienist, hairdresser, nurse, librarian, planner, therapist, practitioner, teacher		
W12: male vs. female adjectives (Kortenhaus and Demarest [23])	X	competent, instrumental, achieving, motivated, clever, adventure, earning, master	0.153	0.537
	Y	nurturing, dependent, obedient, incompetent, passive, victim, unsuccessful, invisible		
W13: dominant vs. obedient	X	dominant, ruling, oppressive, controlling, commanding, superior, authority, instrumental	0.047	0.837
	Y	obedient, willing, attentive, considerate, wellbehaved, polite, forced, cooperative		
W14: male vs. female school subjects	X	mathematics, physics, science, chemistry, computing, engineering, sports, technology	0.017	1.101
	Y	humanities, arts, education, biology, medicine, language, music, english		

Table 7.2: Hand-Made WEAT Lists and Their Test Values

8 CHANGE OF GENDER BIAS OVER TIME

In Sub-research question 3, the effect of time on gender bias in children's books is researched. The research question is: *How does gender bias in children's books change over time?* Similarly to Sub-RQ1, the question is answered using Ranked Lists, Analogies, WEAT and WEFAT. However, these methods are fitted on embeddings trained on the English Gutenberg corpus. This is the case because of the temporal difference between the OCLC and the English Gutenberg embeddings. While the books underlying the OCLC embeddings have a mean publication date of 2006, the books in the English Gutenberg corpus were published in 1876, on average. By comparing the two embeddings, a change in gender bias over time can be analysed.

8.1 Ranked Lists

At first, Ranked Lists were extracted from the English Gutenberg embeddings concerning the most biased professions, nouns, adjectives, verbs and animals. The test lists were queried in relation to the gendered group vectors as established by Garg et al. [30]. Again, gender-appropriate results were not excluded to get an image of how biased the embeddings actually are. Comparison is drawn in regard to the results of the OCLC embeddings as reported in Section 6.1.

8.1.1 Professions

Looking at the results in Table 8.1, it can be seen that there is great overlap between the results of the English Gutenberg embeddings and the OCLC embeddings. Both show high association of the female direction with the home- and care-related professions (`maid`, `nurse`, `housewife`, `housekeeper` and `nanny`). Moreover, both show high correlation of females with art-professions like `actress`, `singer`, `soloist`, `pianist`, and `dancer`. In addition, the female direction is correlated with educational professions such as `teacher` and `dean`. It is worth mentioning that the areas of profession are linked. Home- and education-related professions also contain care aspects, giving the female professions the overarching theme of care.

A difference can be found in fashion-related professions. Those can be found back in the OCLC embeddings and the results of Bolukbasi et al. [27] but not in the English Gutenberg ones. This difference between the English Gutenberg and the OCLC results in regard to fashion-related professions might be due to the mean publication date of the data underlying the models, 1876 and 2006 respectively. The OCLC model is hence close to the GoogleNews corpus which was used by Bolukbasi et al. [27]. As the fashion industry has mostly grown with the world wars in the twentieth century, it was simply not as present in previous centuries [59].

When looking at the professions associated with the male direction in the English Gutenberg embeddings (see Table 8.2), it is clear that the dominating theme in the Ranked List is military. Many of the professions are military ranks like `deputy`, `marshal`, `lieutenant`, `sergeant` and `commander`. Besides that, there are professions connected to law enforcement (`constable` and `ranger`), religion (`rabbi`) and sports (`boxer`). While the OCLC embeddings also showed association of males with the military, there were also other groups present like agriculture and

	Extreme *female:* OCLC	Bias		Extreme *female:* EN Gutenberg	Bias
1.	ballerina	0.410	1.	maid	0.325
2.	maid	0.305	2.	nurse	0.319
3.	dancer	0.302	3.	housewife	0.219
4.	actress	0.282	4.	waitress	0.218
5.	hairdresser	0.232	5.	actress	0.214
6.	socialite	0.209	6.	housekeeper	0.197
7.	nanny	0.204	7.	nun	0.139
8.	stylist	0.203	8.	singer	0.132
9.	valedictorian	0.198	9.	pianist	0.106
10.	planner	0.181	10.	soloist	0.106
11.	choreographer	0.176	11.	comic	0.102
12.	singer	0.166	12.	teacher	0.097
13.	understudy	0.155	13.	dancer	0.096
14.	housekeeper	0.147	14.	nanny	0.066
15.	pianist	0.141	15.	dean	0.059

Table 8.1: Comparison of the Most Extreme Professions as Projected on the Female Gender Direction of the OCLC and the EN Gutenberg Embeddings

food provision. This focus on military ranks in the English Gutenberg embeddings is likely due to the mean publication dates. War and military-dominated cultures much more in the past than was the case after the Second World War. Also, children's stories were evolving around strong and dapper soldiers, which is not the case that much anymore in recent years.

	Extreme *male:* OCLC	Bias		Extreme *male:* EN Gutenberg	Bias
1.	fireman	-0.374	1.	deputy	-0.366
2.	ranger	-0.317	2.	foreman	-0.355
3.	farmer	-0.315	3.	marshal	-0.354
4.	commander	-0.304	4.	carpenter	-0.347
5.	policeman	-0.290	5.	lieutenant	-0.335
6.	fisherman	-0.288	6.	constable	-0.327
7.	colonel	-0.287	7.	provost	-0.318
8.	preacher	-0.281	8.	sergeant	-0.312
9.	lieutenant	-0.279	9.	rabbi	-0.311
10.	captain	-0.275	10.	commander	-0.309
11.	sailor	-0.268	11.	boss	-0.309
12.	sergeant	-0.267	12.	boxer	-0.306
13.	butcher	-0.264	13.	bodyguard	-0.303
14.	soldier	-0.258	14.	surveyor	-0.291
15.	foreman	-0.252	15.	ranger	-0.290

Table 8.2: Comparison of the Most Extreme Professions as Projected on the Male Gender Direction of the OCLC and the EN Gutenberg Embeddings

8.1.2 Nouns

In Table 8.3, the most extreme nouns of the English Gutenberg embeddings as projected on the female gender direction can be found. It quickly becomes apparent that the female direction includes many gender appropriate words like `miss`, `lady`, `girl`, `she`, `woman` and `queen`. This is the same as was the case with the OCLC embeddings (see Table 6.2). Nevertheless, there is also a difference between the two Ranked Lists. The English Gutenberg embeddings contain association of nouns describing a `sweet`, `neat` and `soft` woman with nicknames like `honey` and `dear`. This is in line with the findings of Garg et al. [30] and the expectations. This image of a quiet and polite woman changes a bit when looking at nouns associated with the OCLC embeddings. There, females are associated with `magazines` and `drama`, describing their interest in fashion and tabloids. It must be noted that the words `sweet`, `beautiful`, `neat` and `soft` are ambiguous in their word form and can also be seen as adjectives instead of nouns. This can change the picture that is being painted as a focus on adjectives remains.

	Extreme *female:* OCLC	**Bias**		**Extreme *female:* EN Gutenberg**	**Bias**
1.	magazine	0.344	1.	sweet	0.372
2.	drama	0.303	2.	miss	0.369
3.	girl	0.301	3.	lady	0.347
4.	queen	0.300	4.	girl	0.333
5.	designer	0.293	5.	nurse	0.319
6.	skirt	0.291	6.	beautiful	0.309
7.	dress	0.286	7.	she	0.307
8.	lady	0.282	8.	woman	0.279
9.	wedding	0.275	9.	baby	0.262
10.	miss	0.272	10.	queen	0.256
11.	sweet	0.260	11.	neat	0.246
12.	beautiful	0.248	12.	soft	0.242
13.	flower	0.247	13.	honey	0.237
14.	formal	0.240	14.	dress	0.235
15.	she	0.238	15.	dear	0.235

Table 8.3: Comparison of the Most Extreme Nouns as Projected on the Female Gender Direction of the OCLC and the EN Gutenberg Embeddings

For the male direction, there are great differences between the OCLC and the English Gutenberg embeddings. The latter contains many gender-appropriate words like `uncle`, `brother` and `son`, but also military ranks (`major` and `officer`) and some work-related words (`boss`, `employer`, `master`, `leader`, `staff`). In return, the OCLC Ranked Lists contains hardly identifiable trends and the biggest group was about vehicles. This shows, that over time, the nouns associated with males become much more diverse and the focus on the military decreased, as was the case for the professions.

8.1.3 Adjectives

The results of the most female associated adjectives in the English Gutenberg embeddings can be found in Table 8.5. The most female-biased words describe a `lovely`, `sweet`, `adorable` and `charming` girl or woman. While this is similar to the adjectives used to describe females in the OCLC embeddings, a slight difference can be found. The most female adjectives in the OCLC embeddings focus on the looks of women and describe a beautiful female to look up to and long

	Extreme *male:* OCLC	Bias		Extreme *male:* EN Gutenberg	Bias
1.	sir	-0.437	1.	master	-0.409
2.	tank	-0.390	2.	leader	-0.366
3.	uncle	-0.364	3.	staff	-0.352
4.	bill	-0.350	4.	engineer	-0.321
5.	buddy	-0.347	5.	horse	-0.321
6.	truck	-0.328	6.	mate	-0.319
7.	son	-0.316	7.	employer	-0.315
8.	farmer	-0.315	8.	uncle	-0.314
9.	engine	-0.306	9.	boss	-0.309
10.	chip	-0.305	10.	son	-0.308
11.	hook	-0.294	11.	brother	-0.299
12.	dump	-0.285	12.	major	-0.296
13.	baseball	-0.282	13.	officer	-0.289
14.	general	-0.281	14.	king	-0.289
15.	gas	-0.279	15.	mark	-0.284

Table 8.4: Comparison of the Most Extreme Nouns as Projected on the Male Gender Direction of the OCLC and the EN Gutenberg Embeddings

for, yet objectifying her. The adjectives in the English Gutenberg are not glamorous. The girls and women are `sweet` and `rosy`. They are nice to others and tidy, making them pleasant and obedient people. Yet, they are not to be looked up to. Rather the opposite position is awarded, where females are seen as small and something to look down to. At the same time, they are also seen as `maternal`, being `dimpled`, `feminine` and `motherly`. This prescribes them the clear role of a wife and mother, which was not so clearly present in the OCLC embeddings.

	Extreme *female:* OCLC	Bias		Extreme *female:* EN Gutenberg	Bias
1.	glamorous	0.431	1.	pink	0.399
2.	graceful	0.289	2.	lovely	0.394
3.	bubbly	0.286	3.	dimples	0.385
4.	fancy	0.284	4.	feminine	0.379
5.	pink	0.283	5.	motherly	0.376
6.	enchanting	0.282	6.	sweet	0.372
7.	fabulous	0.280	7.	adorable	0.364
8.	sparkling	0.275	8.	charming	0.364
9.	gorgeous	0.272	9.	rosy	0.346
10.	glittering	0.261	10.	violet	0.341
11.	spirited	0.261	11.	vivacious	0.330
12.	sweet	0.260	12.	tidy	0.318
13.	dazzling	0.252	13.	frilly	0.316
14.	elegant	0.250	14.	fluffy	0.315
15.	lovely	0.248	15.	beautiful	0.309

Table 8.5: Comparison of the Most Extreme Adjectives as Projected on the Female Gender Direction of the OCLC and the EN Gutenberg Embeddings

On the male direction (see Table 8.6), there are also differences between the two embed-

dings. While it was difficult to establish a trend for the OCLC embeddings and males were portrayed as being in a bad mood, the English Gutenberg embeddings describe males as big (major, jumbo, giant, burly) and as adventurers (intrepid and victorious). This difference is hard to explain and does not seem to result from the temporal difference between the two corpora. Still, one reason might be the ambiguity of words like major, which can mean big but are also a military rank. Since males are closely associated with the military in the English Gutenberg embeddings, the association of major could be related. The association of males with big forms a stark contrast to the association of females with adjectives that make them small like lovely or sweet. This could be based on physical difference between the genders, whether factual or wished for, or point towards a dominance of men as big things are usually more impressive and powerful than small things.

Extreme *male:* OCLC		**Bias**	**Extreme *male:* EN Gutenberg**		**Bias**
1.	swift	-0.381	1.	trusty	-0.385
2.	general	-0.281	2.	chief	-0.337
3.	growling	-0.280	3.	frank	-0.324
4.	trusty	-0.278	4.	major	-0.296
5.	chief	-0.270	5.	jumbo	-0.273
6.	giant	-0.264	6.	giant	-0.273
7.	snarling	-0.253	7.	burly	-0.270
8.	miserly	-0.249	8.	intrepid	-0.247
9.	frank	-0.248	9.	victorious	-0.222
10.	mammoth	-0.240	10.	automatic	-0.210
11.	jolly	-0.230	11.	second	-0.209
12.	grizzled	-0.229	12.	enraged	-0.209
13.	nifty	-0.225	13.	official	-0.209
14.	slippery	-0.218	14.	powerful	-0.208
15.	flat	-0.216	15.	growling	-0.207

Table 8.6: Comparison of the Most Extreme Adjectives as Projected on the Male Gender Direction of the OCLC and the EN Gutenberg Embeddings

8.1.4 Verbs

Table 8.7 displays the most female associated verbs in the OCLC and English Gutenberg embeddings. Many words are related to home and care activities (knits, dresses, tucks, nurses, bakes, tidies, ruffles, sews and weaves) but also loving interactions (smiles, kisses, adores and flirts). It is noteworthy that most of them relate to homemaking and childcare, historically stereotypical female domains. This is in line with the findings from the nouns and adjectives list, where women were associated with being tidy and motherly. This is somewhat different to the OCLC embeddings where only some of the verbs relate to caring and loving interaction (embraces, adores and charms). The majority of the verbs was fashion-related (dresses, sparkles, shines and threads), which is less the case in the English Gutenberg embeddings. Also this is in line with previous findings, where professions in the OCLC were fashion-related, while the occupations in the English Gutenberg embeddings were not.

On the male direction (see Table 8.8), there is also a great difference between the two embeddings. The English Gutenberg corpus seems to mainly focus on law enforcement, war or physical actions of attack or protection (outlaws, guards, rams, defends, orders, flanks,

	Extreme *female:* OCLC	Bias		Extreme *female:* EN Gutenberg	Bias
1.	dresses	0.306	1.	knits	0.346
2.	sparkles	0.276	2.	dresses	0.325
3.	embraces	0.231	3.	tucks	0.307
4.	violates	0.219	4.	nurses	0.291
5.	adores	0.216	5.	weeps	0.287
6.	chimes	0.207	6.	bakes	0.282
7.	charms	0.205	7.	adores	0.272
8.	resents	0.200	8.	flirts	0.272
9.	twirls	0.195	9.	tidies	0.271
10.	shines	0.188	10.	sues	0.253
11.	juggles	0.186	11.	ruffles	0.252
12.	immerses	0.185	12.	sews	0.243
13.	pricks	0.185	13.	weaves	0.243
14.	whirls	0.180	14.	smiles	0.239
15.	threads	0.176	15.	kisses	0.235

Table 8.7: Comparison of the Most Extreme Verbs as Projected on the Female Gender Direction of the OCLC and the EN Gutenberg Embeddings

`sears`, `invades`, `surrenders` and `ventures`). This is in line with the professions that were associated with males in these embeddings, which also mainly consisted of military ranks. The OCLC embeddings, in return, did not have such a clear theme and included physical actions and bodily functions (`bangs`, `bites`, `shovels`, `fumbles`, and `smashes`) but it was not inherently military or law enforcing. Again, this shows a shift over time away from the military focus.

	Extreme *male:* OCLC	Bias		Extreme *male:* EN Gutenberg	Bias
1.	associates	-0.332	1.	outlaws	-0.320
2.	bones	-0.302	2.	guards	-0.289
3.	bangs	-0.267	3.	spurs	-0.287
4.	bites	-0.266	4.	shuffles	-0.278
5.	rockets	-0.257	5.	rams	-0.277
6.	tires	-0.255	6.	defends	-0.267
7.	rounds	-0.252	7.	associates	-0.248
8.	huffs	-0.242	8.	hurls	-0.247
9.	shovels	-0.238	9.	orders	-0.236
10.	zooms	-0.237	10.	flanks	-0.230
11.	ships	-0.236	11.	sears	-0.227
12.	retires	-0.231	12.	invades	-0.221
13.	sails	-0.230	13.	surrenders	-0.221
14.	fumbles	-0.230	14.	bellows	-0.218
15.	smashes	-0.227	15.	ventures	-0.214

Table 8.8: Comparison of the Most Extreme Verbs as Projected on the Male Gender Direction of the OCLC and the EN Gutenberg Embeddings

8.1.5 Animals

According to the animals most associated with the female direction in the English Gutenberg embeddings, females tend to be portrayed or associated with birds (`dodo`, `goose`, `chicken`, `kiwi` and `puffin`) and insects (`butterfly`, `cockroach`, `moth` and `ladybird`) (see Table 8.9). They are small and harmless but also beautiful and sweet - adjectives also used to describe women. Moreover, chickens are female birds, so this association is appropriate. The OCLC embeddings, in return, showed association of females with birds and dogs as well as cat breeds. Common issues are the association of `molly` and `ladybird` with the female gender direction, which can be explained by other factors than the common adjectives. Furthermore, `sponge` is probably associated with the female direction because of the females association with cleaning and doing the dishes, rather than it being meant as the animal.

Extreme *female:* OCLC		**Bias**	**Extreme *female:* EN Gutenberg**		**Bias**
1.	tiffany	0.221	1.	molly	0.355
2.	swan	0.207	2.	butterfly	0.208
3.	ladybird	0.200	3.	chinchilla	0.170
4.	butterfly	0.135	4.	dodo	0.160
5.	nightingale	0.134	5.	cockroach	0.156
6.	quetzal	0.130	6.	coral	0.135
7.	molly	0.093	7.	goose	0.128
8.	coral	0.082	8.	maltese	0.126
9.	poodle	0.051	9.	moth	0.115
10.	ragdoll	0.048	10.	chicken	0.105
11.	pekingese	0.039	11.	ladybird	0.104
12.	persian	0.037	12.	sponge	0.087
13.	horse	0.036	13.	kiwi	0.075
14.	quokka	0.035	14.	guppy	0.074
15.	siamese	0.034	15.	puffin	0.071

Table 8.9: Comparison of the Most Extreme Animals as Projected on the Female Gender Direction of the OCLC and the EN Gutenberg Embeddings

On the male direction (see Table 8.10), most associated animals are predators (`lion`, `badger`, `panther`, `bobcat`, `tiger` and `rattlesnake`) or large and aggressive-looking dog breeds like `bulldog` and `mastiff`. This fits the image of the fighting male, painted by the most associated nouns, adjectives and verbs as well. Males `guard`, `defend`, `surrender` and `bellow` while being `victorious`, `powerful` and `growling`. Another theme is the one of being big. `Pikes` are big fish, `horses`, `mules` and `elephants` are large mammals. As males are associated with adjectives surrounding being big, it makes sense to see many large animals being associated with males. The animals associated with the male direction in the OCLC embeddings also follow this idea of being dangerous or big to some extent (`coyote`, `hyena`, `snake` and `crocodile`), even if it is just being big relative to other types within the same species. For example, frogs are generally not very large, but the `bullfrog` is big compared to other frog types. However, this trend is not as prominent as in the English Gutenberg embeddings and the OCLC embeddings are much more diverse in the animals associated with the male direction.

Extreme *male:* OCLC		Bias	Extreme *male:* EN Gutenberg		Bias
1.	raccoon	-0.370	1.	pike	-0.328
2.	beaver	-0.349	2.	horse	-0.321
3.	camel	-0.347	3.	bulldog	-0.315
4.	catfish	-0.324	4.	mule	-0.311
5.	rat	-0.321	5.	lion	-0.296
6.	bullfrog	-0.314	6.	badger	-0.295
7.	moose	-0.311	7.	gibbon	-0.291
8.	mule	-0.297	8.	panther	-0.285
9.	coyote	-0.296	9.	burmese	-0.275
10.	lizard	-0.295	10.	lizard	-0.263
11.	buffalo	-0.292	11.	mastiff	-0.257
12.	hyena	-0.291	12.	bobcat	-0.255
13.	snake	-0.290	13.	tiger	-0.255
14.	crocodile	-0.288	14.	elephant	-0.251
15.	squid	-0.286	15.	rattlesnake	-0.249

Table 8.10: Comparison of the Most Extreme Animals as Projected on the Male Gender Direction of the OCLC and the EN Gutenberg Embeddings

8.2 Analogies

After exploring the most male and female associated words and their change over time, the English Gutenberg embeddings were queried for analogies matching the seed pair *she-he*. From the 100 analogies with the highest scores, the top 20 were reported that did not include names or words from the gendered lists used to make the gendered group vectors.

Overall, the produced analogies include many gender-appropriate relations like `mistress-master`, `girlhood-boyhood`, and `lady-gentleman` (see Table 8.11). This is similar to the OCLC embeddings, which also included gender-appropriate analogies like `cowgirl-cowboy` or `queen-king`. These terms address the criticism of Nissim, van Noord and van der Groot [28], who said a distinction needs to be made between analogies that have a logical fourth term and analogies that do not. The resulting gender-appropriate pairs have a logical fourth term since they form a logical pair that only differs in the gender direction. This is the case because the embeddings are queried for words matching the seed pair *she-he* with the dividing direction being gender.

Nevertheless, the analogies also produced gender-stereotypical pairs. Some are describing different ways of uttering for the genders. While females `giggle`, males `grin`, women `murmur` while men `mutter`, when girls `scream`, boys `yell` and when mothers `wail`, fathers `growl`. Coming back to the criticism of a logical fourth term, the fourth logical term for the analogy `she:giggle - he:?` is not immediately apparent. However, the pair `giggle-grin` both describe the action of smiling, so they are semantically coherent. Their difference hence seems to stem from their association with the genders, making them gender stereotypical. Another group of gender stereotypical analogies include clothing for men and women. Females wear `frocks`, `shawls`, `gowns` while males wear `shirts`, `blankets` and `coats`. Again, the pairs are semantically coherent while remaining a main difference in the gender direction. Both, `frocks` and `gowns` are old-fashioned words for dress, a traditionally female piece of clothing. On the other hand, they are associated with `shirts` and `coats`. The former is a traditionally male item, underlining the difference in the gender direction between these two terms. However, coats can be worn by both genders, making this analogy highly biased.

It is interesting to note the difference between the results of the English Gutenberg and the OCLC embeddings. Due to their different mean publication dates, their analogies also differ. The English Gutenberg includes old-fashioned words like `lady-gentleman`, `countess-baron` and `ladyship-lordship`, that are barely used in modern language. In return, the OCLC model talks about `bff-girlfriend`, `bff-teammates` and `cheerleader-troublemaker`, terms, that did not exist yet when books in the Gutenberg model were published. Moreover, the analogies in the English Gutenberg embeddings have higher association scores than the ones from the OCLC embeddings. This trend could also be observed with the bias scores in the Ranked Lists at times and probably stems from the larger corpus size underlying the English Gutenberg embeddings.

	Analogies	Score
1.	mistress-master	0.670
2.	girl-fellow	0.657
3.	frock-shirt	0.611
4.	girlhood-boyhood	0.609
5.	mrs-mr	0.603
6.	lady-gentleman	0.598
7.	ladylike-gentlemanly	0.596
8.	countess-baron	0.593
9.	ladyship-lordship	0.581
10.	giggled-grinned	0.580
11.	murmured-muttered	0.578
12.	girls-cadets	0.577
13.	governess-tutor	0.574
14.	screamed-yelled	0.573
15.	niece-employer	0.567
16.	shawl-blanket	0.563
17.	heroine-hero	0.556
18.	gown-coat	0.555
19.	boudoir-stateroom	0.549
20.	flounced-strode	0.546

Table 8.11: Analogies for the *She-He* Axis of the EN Gutenberg Embeddings

8.3 Word-Embedding Association Test

8.3.1 Testing Current WEATs

Next to Ranked Lists and Analogies, the WEAT was used to test gender bias in the English Gutenberg embeddings. Table 8.12 shows the results of the five WEATs from Table 5.2 and the handmade lists as established in Sub-Research Question 2. Values in bold indicate statistically significant gender bias for $p < 0.05$.

The English Gutenberg embeddings show significant bias in all of the original WEATs established by Caliskan et al. [26] and Garg et al. [30]. This means that the target lists have significantly different mean association with the genders. Given the WEATs, `career` is hence closer associated with `male` and `family` with `female`, `math` and `science` relate to `male` while `arts` relate to `female`, and `intelligence` as well as `strength` are `male` biased while `appearance` and `weakness` are `female` biased. Given the conventional small, medium and large values of Cohen's d, the effect sizes of the English Gutenberg embeddings are very large. They range

from 0.82 (W4) to 1.76 (W1), showing highest relation for W1 and W3.

When taking a closer look at the lists, it can be seen that some words were out of vocabulary for the embeddings. In the `arts` list, the word `Shakespeare` could not be found, in the `science` list words `NASA` and `Einstein` were not present and for the `weakness` list the word `wispy` was out of vocabulary. It is interesting to notice that mostly proper nouns are out of vocabulary for the embeddings. Nevertheless, this did not affect the tests as previously predicted.

The results of the English Gutenberg embeddings are comparable to the ones produced by Caliskan et al. [54] and Chaloner and Maldonado [31] with Cohen's d being very similar. It is interesting to note that the literature and the English Gutenberg embeddings report significant biases for all original WEATs, however, the OCLC embeddings only have significant biases in W1 - W3. Moreover, the effect sizes for W1 vary greatly. The OCLC embeddings seem to be the outlier with a much smaller effect size than the other embeddings are reporting for W1. This, as well as the difference in significance for W4 and W5, might be due to the nature of texts underlying the different embeddings. The corpus used for the results of Caliskan et al. [26] as well as Chaloner and Maldonado [31] was the GoogleNews corpus, consisting of news items. The English Gutenberg embeddings were trained on full-text English children books. The OCLC embeddings were trained on short descriptions of books with an average length of 49 words per description. This is much shorter than the other texts. While descriptions contain gender bias as suggested by social science literature and seen in the results presented in Chapter 6, the extent might be limited at parts. Moreover, the corpus retrieved from OCLC is smaller than the other corpora, which can play a role in training embeddings and finding significant bias.

When it comes to the hand-made lists, the English Gutenberg embeddings show significant results for W7 - W9, W11, and W13 - W14. These are the same WEATs the OCLC embeddings were significant on. The tests have large Cohen's ds in both embeddings according to the conventional interpretation. Yet, the effect sizes were larger for the OCLC embeddings across all hand-made lists. One reason for this might be that the WEATs were subconsciously tailored to the OCLC embeddings in terms of the language used and topics discussed. This makes it harder to apply them to children's literature at large. Playing into this could also be the time difference between the OCLC and the English Gutenberg embeddings. The hand-made lists might be too modern to be applied to older literature. This can also be seen in the amount of out-of-vocabulary words of the hand-made WEAT lists in the English Gutenberg embeddings. The following words were not in the vocabulary:

- Female toys: `ballerina` and `barbie`
- female and male sports: `soccer`, `netball`, `volleyball`, `softball` and `cheerleader`
- male adjectives: `charismatic`
- male and female professions: `programmer`, `planner`, `therapist` and `hygienist`
- male and female adjectives according to Kortenhaus and Demarest [23]: `motivated` and `nurturing`

As it can be seen, the out-of-vocabulary words mostly include modern words like `barbie` or `programmer`, which did not exist when the books of the English Gutenberg corpus were published. Hence, it is logical that they are not in the corpus, but it shows that temporal differences need to be taken into account when drawing up WEAT lists manually.

8.3.2 Change of Bias Scores Over Time

Besides analysing the English Gutenberg embeddings with the WEATs from literature and manual crafting, the approach of Garg et al. [30] was used to measure a change in gender bias over

		Caliskan [54]	Chaloner and Maldonado [31]	OCLC	English Gutenberg
	WEAT category	*d*	*d*	*d*	*d*
W1:	career vs. family	**1.81**	**1.37**	**0.839**	**1.755**
W2:	maths vs. arts	**1.06**	**1.02**	**1.078**	**1.052**
W3:	science vs. arts	**1.24**	**1.25**	**1.446**	**1.743**
W4:	intelligence vs. appearance		**0.98**	0.114	**0.809**
W5:	strength vs. weakness		**0.89**	-0.152	**0.820**
W6:	outdoor vs. indoor			0.185	**1.258**
W7:	male vs. female toys			**1.737**	**1.302**
W8:	male vs. female sports			**1.550**	**0.715**
W9:	active vs. quiet games			**1.259**	**1.215**
W10:	male vs. female adjective			-0.949	-0.898
W11:	male vs. female professions			**1.640**	**1.304**
W12:	male vs. female adjectives [23]			0.537	0.102
W13:	dominant vs. obedient			**0.837**	**1.282**
W14:	male vs. female school subjects			**1.101**	**0.829**

Table 8.12: Results of the WEAT Tests. Values in Bold Indicate Statistically Significant Gender Bias ($p < 0.05$).

time. This approach includes the comparison of gender bias scores of adjectives across embeddings trained on different decades. Embeddings per decade were trained on the English Gutenberg corpus. However, not all decades were represented equally in the corpus, leaving some decades with as little as two books to represent them. As a consequence, only the decades ranging from 1820 to 1950 were used in the analysis as they counted a sufficient number of books. Each decade is represented by embeddings trained on books published in that decade. Table 8.13 gives an overview of the number of books per decade, the minimum word count chosen for the embeddings and the resulting vocabulary size of the embeddings. Decades in grey were excluded from the analysis.

For each decade, the association of the adjectives in the RL target list with the gendered attribute lists was calculated. These embedding bias scores are then set in relation to each other and the correlation coefficient Pearson's r is calculated for each combination of decades. The results of this analysis can be found in Figure 8.1. The diagonal represents the correlation with itself and is hence 1.0. The darker the colouring of the cell, the higher the correlation between the decades.

It can quickly be seen that the decades 1840 to 1920 seem to form a block, meaning they correlate more with each other than with the decades outside of the block. A likely explanation for this is the large number of books in these decades. These decades have much more books than earlier and later decades do, giving them more data to make higher quality embeddings and calculate embedding bias scores.

Other events that could have influenced the break between the decades before and after 1830 are the crowning of Queen Victoria in 1837 and the women's rights movement in the United States (U.S.) in the 1830s and 1840s. When Queen Victoria took the throne in 1837, women were highly disadvantaged and discriminated against. They had no right to vote and, once married, were seen as the property of their husbands, including their possessions, wages and body [60]. Yet, the crowning of Victoria as Queen might have sparked stories and children's books about her highness, influencing the published literature. Moreover, in the 1830s and 40s, women started to speak out for their rights for the first time in U.S. history. Many women were active in the abolition movement, fighting against slavery. However, this was not welcomed by men, even those supporting the abolition cause. As a result, a women's rights movement

Decade	Nr. of books	min word count	vocabulary size
1800	13	3	11,955
1810	6	1	15,359
1820	26	3	16,521
1830	34	3	18,058
1840	172	10	21,891
1850	97	5	28,936
1860	361	10	35,085
1870	259	10	29,034
1880	227	10	28,495
1890	348	10	33,278
1900	240	10	27,141
1910	209	10	23,575
1920	92	5	23,992
1930	17	3	9,483
1940	25	3	12,000
1950	52	5	11,611
1960	2	1	9,661
1970	6	1	12,326
1980	28	3	11,308
1990	4	1	9,081

* decades in grey were not included in the analysis

Table 8.13: Summary of the Embeddings Trained per Decade

started and “the climate began to change when a number of bold, outspoken women championed diverse social reforms of prostitution, capital punishment, prisons, war, alcohol, and, most significantly, slavery” [61]. This movement probably affected the literature written for upcoming generations of children, including more active and outspoken women.

On the other side of the temporal scale, there is a break between the 1920s and the 1930s. The most likely cause for this is probably the small number of books present in the later decades, leading to much smaller language models. Another explanation could stem from World War 1 and World War 2, which took place in the early twentieth century. During times of war, few books are written or published as all hands are needed to support the military and other crucial sectors. Moreover, the stories written in this time are likely to play in a war setting, while books written in peaceful times are likely to tell stories about such times.

8.4 Word-Embedding Factual Association Test

Lastly, the WEFAT was fitted on the English Gutenberg embeddings to reveal gender bias and possible changes over time. In the original WEFAT, the embedding bias is compared to the percentage of women in the workforce for 50 selected professions. The results of this relation can be found in Figure 8.2a. As can be seen, the relationship between the effect size and the factual property is weak with a Pearson’s $r = 0.394$. Even when the percentage of females in the profession rises, gender bias does not. Overall, most occupations have a negative and hence a male bias. A reason for this is the temporal difference between the English Gutenberg dataset and the occupation data [54]. The former has a mean publication date of 1876 while the latter is from 2015. As women tended to stay at home with the children and household chores, they did not work. This would explain the male bias of most professions. In addition, the percentage of women in the jobs as well as the use of language have changed over time

	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950
1820	1.0	0.467	0.249	0.27	0.11	0.085	0.193	0.088	0.127	0.156	0.26	0.191	0.353	0.244
1830	0.467	1.0	0.286	0.244	0.197	0.168	0.28	0.174	0.129	0.242	0.272	0.117	0.312	0.237
1840	0.249	0.286	1.0	0.694	0.703	0.56	0.703	0.654	0.623	0.612	0.559	-0.041	0.128	0.268
1850	0.27	0.244	0.694	1.0	0.718	0.494	0.735	0.566	0.611	0.566	0.615	0.053	0.259	0.359
1860	0.11	0.197	0.703	0.718	1.0	0.628	0.785	0.738	0.706	0.608	0.55	-0.085	0.081	0.299
1870	0.085	0.168	0.56	0.494	0.628	1.0	0.644	0.691	0.6	0.574	0.442	-0.119	-0.093	0.061
1880	0.193	0.28	0.703	0.735	0.785	0.644	1.0	0.726	0.698	0.716	0.576	-0.06	0.163	0.348
1890	0.088	0.174	0.654	0.566	0.738	0.691	0.726	1.0	0.717	0.689	0.497	-0.133	-0.065	0.233
1900	0.127	0.129	0.623	0.611	0.706	0.6	0.698	0.717	1.0	0.629	0.502	-0.026	0.035	0.341
1910	0.156	0.242	0.612	0.566	0.608	0.574	0.716	0.689	0.629	1.0	0.525	-0.001	0.13	0.325
1920	0.26	0.272	0.559	0.615	0.55	0.442	0.576	0.497	0.502	0.525	1.0	0.018	0.238	0.166
1930	0.191	0.117	-0.041	0.053	-0.085	-0.119	-0.06	-0.133	-0.026	-0.001	0.018	1.0	0.354	0.145
1940	0.353	0.312	0.128	0.259	0.081	-0.093	0.163	-0.065	0.035	0.13	0.238	0.354	1.0	0.376
1950	0.244	0.237	0.268	0.359	0.299	0.061	0.348	0.233	0.341	0.325	0.166	0.145	0.376	1.0

Figure 8.1: Pearson Correlation in Embedding Bias Scores for Adjectives Over Time

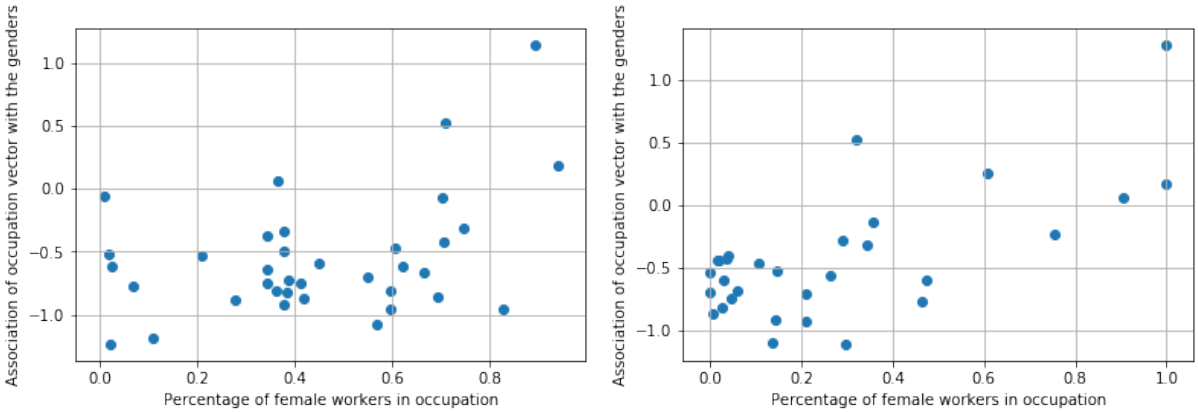
and therefore cannot be compared over time. Moreover, there is a difference in vocabulary due to the time difference. Many of the modern occupations such as programmer or technician did not exist when the books in the English Gutenberg corpus were written. Hence, many of the professions were out-of-vocabulary for the embeddings.

As a result, data is needed that is comparable in time and the professions list needs to be adjusted so that the vocabulary fits the dataset. Hence, the 1851 Census Report [56] is used instead. After summing the number of males and females across counties and age groups, the dataset gives an overview of the percentage of males and females working in certain professions. Trying the WEFAT with the original list of 50 professions and the adjusted dataset, it quickly became apparent that this approach is not sustainable as many of the original professions were either not in the English Gutenberg corpus or not in the adjusted dataset. Thus, the list of professions was adjusted. Using the most male and female associated professions as found in the Ranked Lists or Analogies did not yield better results. Basing the professions list on the occupations in the dataset was a better approach. This lead to a list of 86 professions of which two had female participation of 0% (`architect` and `banker`) and two had female percentage of 100% (`wife` and `Queen`). While `wife` is generally not considered a profession, in the 1851 Census report, it was listed as a profession to be able to account for most people's daily tasks. The extreme male and female professions were included to test for the validity of the embeddings as a profession with 0% women is expected to have great male bias and the other way around. The remaining 82 professions had both males and females working in the profession.

The results of the analysis are presented in Figure 8.2b. The relationship between the embedding bias score and the percentage of women in the 1851 Census Report seems to be linear. The Pearson's $r = 0.675$ with $p < 0.00$ is comparable to the results of SubRQ1 (see Section 6.4). The association between the two variables is hence largely linearly positive, meaning that the more women work in a profession, the higher the female gender bias and vice versa.

Nevertheless, there is an issue with the skewness of the data. Many of the professions have

large percentages of males working in them and only little have a majority of female workers. This can also be confirmed when looking at the individual WEFAT tests for each profession as shown in Appendix F. It indicates the tested professions, the out of vocabulary words, the percentage of females in the occupations, their effect sizes and their p-values. Only five of the professions in the list have a percentage of female workers above 50%. This skewness influences the generalisability of the results, as it mostly gives information about male professions having male bias. Little can be said about female professions having female bias. When taking a closer look at the individual data points in Figure 8.2b, this suspicion can be somewhat confirmed as there are data points that have more than 70% of women working in them, but their bias is around 0. This would indicate a lack of female bias, even when the professions are mostly occupied by females. Further data is needed on professions with high percentages of women to confirm this hypothesis.



(a) EN Gutenberg Embeddings With Original WEFAT Data. Pearson's Correlation Coefficient $r = 0.395$ With $p < 0.017$.

(b) EN Gutenberg Embeddings With Appropriate WEFAT Data. Pearson's Correlation Coefficient $r = 0.675$ With $p < 0.000$.

Figure 8.2: Occupation-Gender Association in the EN Gutenberg Embeddings

8.5 Conclusion and Discussion

In this chapter, the question was asked *How does gender bias in children's books change over time?* It was answered using a variety of methods including Ranked Lists, Analogies, WEAT and the WEFAT.

Overall, the Ranked Lists paint a rather steady picture of females with only a few changes over time. Home- and care-related professions remain the main occupations together with art professions. Females are still associated with being `sweet` and `beautiful`. They are `lovely`, `adorable` and `charming` while their actions focus on caring for others and making the home. They are associated with birds and insects, animals that are small and beautiful. One of the few differences between the two sets of embeddings is the focus on fashion. As time progressed, the association of females with fashion, gossip and tabloids increased. More modern associations can be found in the Ranked Lists such as `designer`, `magazine` and `sparkles`. There is a slight shift away from being `dimpled`, `feminine` and `motherly` to being `glamorous`, `enchanting` and `dazzling`.

For the males, however, associations changed over time with the English Gutenberg results consisting mostly of military ranks and professions related to law enforcement. The focus on agriculture and food provision that was seen in the OCLC embeddings seems to be a new notion that cannot be found back in the English Gutenberg embeddings. Also in the other ranked lists, a decrease in association with military is found, which is replaced by diversity. Instead, men

are related to vehicles (trucks, engine and gas) and sports (baseball). Males went from being major, jumbo and giant to being swift, growling and trusty, characteristics that are hard to group. Only the target list on animals did not show a great shift, as males remain to be associated with big and dangerous animals.

When it comes to the analogies, both the OCLC and the English Gutenberg embeddings produced gender-appropriate results like girlhood-boyhood. However, the terms changed over time. In the English Gutenberg embeddings, old-fashioned phrases like mistress-master and ladyship-lordship were found, whereas the OCLC embeddings contained words that are still frequently used to this day like cowgirl-cowboy and witch-wizard. Another difference is the focus on sports and social relations in the OCLC embeddings that cannot be traced back to the English Gutenberg embeddings.

Looking at the results of the WEATs, the English Gutenberg embeddings contain comparable amounts of gender bias to relevant literature and the OCLC embeddings. A change can only be seen in a slight decrease in the effect sizes for the hand-made lists, which can be attributed to the difference in vocabulary use. Similarly, the linear relationship between percentages of female workers and gender bias scores remained when comparing the English Gutenberg embeddings to relevant factual data.

Issues in the methodology remain similar to Chapter 6. The overlap in target lists between word types leads to ambiguous results with POS-taggings of words being ignored by the embeddings. The analogies contain a great amount of noise through names and the WEAT contains overlap between the target lists in W2 and W3. Nevertheless, the WEFAT also exposed limitations and dependence on the data used. Embeddings can only be set in relation to data from the same country and time frame. Finding appropriate data is time-consuming, especially for historical data, and even then the quality of the data might not be sufficient. In this case, the 1851 UK Census report [56] was used, which included questionable professions like daughter or butcher's wife. Moreover, the dataset showed great bias towards males with a majority of the professions being occupied by males only. The imbalance between male and female-dominated professions made it difficult to relate the embedding bias linearly. In addition, new datasets require new profession lists to be tested. In this research, there was only partial overlap between the embeddings' vocabulary and the professions in the dataset, limiting the comparability of the two.

Comparing embedding bias scores over time yielded a break in the 1840s and the 1930s, with high association scores between these two dates. While real-time events might have caused these breaks, a more likely explanation is the vocabulary size of the embeddings underlying the method. To analyse changing embedding bias scores over time, a great deal of data is necessary. Although data were available from 1800 - 1925, only the decades 1820 - 1950 had a sufficient amount of data to train embeddings and even then there were great differences between the decades. Moreover, it is computationally expensive to implement this method, while interpretability is limited.

All in all, gender bias is just as present in the old books of the English Gutenberg corpus as it is present now in the modern books of the OCLC corpus. Changes occurred in large parts in the use of language, switching old-fashioned words like flounce and strode with modern words like netball and baseball. While male bias seems to have shifted away from association with the military to diversity, female bias remained largely the same. Having established a change of gender bias over time, a change across cultures and languages can be analysed.

9 GENDER BIAS ACROSS CULTURES

In the last sub-research question, the question was posed: *How does gender bias differ between English and German children's books?* To analyse gender bias across cultures, two data sets were drawn upon: descriptions from the German National Library (DNB) and full-text books from Gutenberg (German Gutenberg). Similarly to the English sets, the descriptions are mostly from modern books with a mean publication date of 2017 and the Gutenberg books are older with a mean publication date in 1868. This way, gender bias cannot only be compared across cultures but simultaneously also across time.

9.1 Ranked Lists

Like before, Ranked Lists were used to measure the most male and female-biased words. As the previously used target lists were in English, translation to German was required. This was done using the application of DeepL [57]. The list of professions was not translated as professions are gendered in German, meaning there is a grammatical male and female version of each profession. However, words are only translated to one version of the profession, according to the bias of the translator. For example, the profession `secretary` is translated to `Sekretärin` (female) and `doctor` is translated to `Arzt` (male). This way, association with the genders is not based on stereotypes but grammar. For the adjective, nouns and animals list, the DeepL translator [57] was used. The resulting lists were checked by hand for translation errors and possible gendered words. When a list included a profession, it was either removed (as was the case for the adjectives list) or it was duplicated and both the male and the female versions were included in the list (as was the case for the nouns list). For the verb list, translation was very faulty and a German list of around 8.000 verbs was used instead, drawn upon from [58]. The verbs are in the third form singular, as was the case for the English verbs.

9.1.1 Nouns

In Table 9.1, the most male and female associated nouns of the DNB and the German Gutenberg embeddings can be found. The list is sorted based on the highest bias score of the German word with the German attribute lists, but an English translation is included as well. It is interesting to note the extremely high association scores, especially for the male direction. These scores were much lower in the English models (max around |0.3|), and the difference between male and female direction is very significant. A reason for this might be the size of the corpora. A smaller corpus leads to higher co-occurrences as the vocabulary is much smaller. The higher the co-occurrences, the higher the association between words.

Moreover, there seems to be a higher quality in male associations as those include gender-appropriate results like `grandfather`, `brother`, `male friend` and `guy`. The female direction, in return, includes only appropriate results in the German Gutenberg embeddings, which are `female friend` and `cousin` (in this case translated to the female form `cousine`) and `friends` (female plural `freundinnen`). The DNB embeddings do not show any gender-appropriate association with the female direction. One reason for this might be the lack of females in titles and

main characters as was discovered by Weitzman et al. [10], Kortenhaus and Demarest [23] as well as McCabe et al. [25]. Since the DNB corpus consists only of titles and abstracts, which usually describe the main character and main plotline, it stands to reason that females are less often named in this corpus and hence few gender-appropriate associations were found.

When looking at the remaining results, the DNB seems to associate females with challenges, missions and tasks, which are interesting results. They would indicate that females face many struggles or have ambitions, which was not a theme in the English embeddings. Moreover, women are connected to flows of information through the association with terms as language, solution, ask and information. In the German Gutenberg embeddings, in return, females are more associated with arts and entertainment (surprise, excitement, music and entertainment), but also the home and social relations (apartment, friendship and group). This shows some overlap with the female association in English Gutenberg embeddings, where females were portrayed as dancers, singers and actresses.

On the male gender direction, association with a few loose groups can be found. Surprisingly, males are associated with animals (horse and dog). This trend continues in the German Gutenberg embeddings with males being associated with dogs, birds and fish. It might be that this association stems from animals frequently being represented as male in stories, with horses and dogs being very familiar animals for children and hence used a lot in stories. Another very familiar animal are cats, however, those tend to be represented as females. Yet, those are not in the most female associated words, which might contradict this reasoning. Another explanation could be the association of males with sports and outdoor activities such as hunting, fishing and horse riding. The books of the German Gutenberg embeddings are very old and in the time of publication of most books, it was common that men were the ones hunting game or birds, for which they use hunting dogs and horses. Moreover, men were the ones riding horses, whether for transportation purposes or sports, while women were more seen in carriages.

Other male-associated groups in the DNB embeddings are thinking (desire, focus, dream and memory) and spaces (room and restaurant). This seems to indicate that males are the ones that have hopes and dreams, which is surprising, as these are generally more considered female traits as males are connected to ambition rather than dreaming. The connection to spaces is an interesting one, as females were also associated with apartments and the home. However, a difference can be seen between those two groups with the female words being related to home whereas the male ones are outside the home (restaurant). The words estate was translated wrongly and corresponds more to wealth than to real estate.

It is worth mentioning that the German language has three articles for nouns instead of just the English article “the”. These are: “die” for female nouns, “das” for neutral nouns, and “der” for male nouns. Yet, they must not be confused with the gender of the noun, as also objects like “bus” or “door” have grammatical gender. The former is “der Bus” (the male bus) and the latter is “die Tür” (the female door). Keeping this classification in mind, it can be observed that all nouns associated with the female direction have the grammatical gender female, while all items on the male direction have the grammatical gender neutral or male. This might have influenced the association of nouns with the genders as language models cannot make a distinction between grammatical and syntactic gender.

9.1.2 Adjectives

The adjectives used in the target lists were translated from the English adjective list. The list was checked by hand for translation mistakes and gendered words, which were corrected or removed from the list. Issues in translation included adjectives like first and second as well as dearest, beautiful, easy-going, beloved and amazing being translated to the female noun, instead of the adjective. For example, beautiful was translated as “schöne”, which means “the beautiful” (female). It was corrected to schön, meaning beautiful. The resulting

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	kosten	cost	0.484	1.	großvater	grandfather	-0.602
2.	herausforderung	challenge	0.445	2.	hemd	shirt	-0.544
3.	mission	mission	0.425	3.	bruder	brother	-0.538
4.	aufgabe	task	0.425	4.	pferd	horse	-0.528
5.	sprache	language	0.382	5.	zimmer	room	-0.516
6.	chance	chance	0.381	6.	wunsch	desire	-0.512
7.	lösung	solution	0.379	7.	vater	father	-0.511
8.	region	region	0.377	8.	fokus	focus	-0.510
9.	fragen	ask	0.377	9.	traum	dream	-0.506
10.	dimension	dimension	0.374	10.	freund	male friend	-0.496
11.	gruppe	group	0.372	11.	hund	dog	-0.487
12.	kraft	force	0.369	12.	gedächtnis	memory	-0.478
13.	bühne	stage	0.364	13.	restaurant	restaurant	-0.477
14.	liebe	love	0.361	14.	kumpel	buddy	-0.473
15.	informationen	information	0.356	15.	vermögen	estate	-0.452

(a) The Most Extreme Nouns in the DNB Embeddings

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	sachen	stuff	0.358	1.	name	name	-0.630
2.	freundin	female friend	0.353	2.	hund	dog	-0.618
3.	cousine	cousin	0.326	3.	vogel	bird	-0.608
4.	überraschung	surprise	0.316	4.	fisch	fish	-0.591
5.	aufregung	excitement	0.305	5.	kerl	guy	-0.584
6.	musik	music	0.299	6.	spaß	fun	-0.578
7.	wohnung	apartment	0.299	7.	großvater	grandfather	-0.574
8.	freundinnen	friends	0.295	8.	gast	guest	-0.562
9.	anleitung	guidance	0.293	9.	freund	male friend	-0.560
10.	tätigkeit	activity	0.288	10.	traum	dream	-0.560
11.	freundschaft	friendship	0.286	11.	fall	case	-0.546
12.	möglichkeit	possibility	0.279	12.	ball	ball	-0.542
13.	gruppe	group	0.276	13.	wunsch	desire	-0.528
14.	pflicht	duty	0.275	14.	bruder	brother	-0.526
15.	unterhaltung	entertainment	0.275	15.	teufel	devil	-0.525

(b) The Most Extreme Nouns in the German Gutenberg Embeddings

Table 9.1: The Most Extreme Nouns as Projected on the Gender Direction

list was set in relation to the gender direction based on the German attribute list. The results can be found in Table 9.2. It is striking to see the difference in association scores between the two gender directions. The male scores are much higher than the female scores, especially for the German Gutenberg embeddings.

On the female direction, the DNB embeddings show association with `clever`, `wise` and `experienced`, which is quite opposing to the image that the English embeddings have painted. The latter focused on the appearance of females, leaving little to no attention to skills. Also, the German Gutenberg embeddings show an association of the female gender direction with `wise` but, the bias score is very low, indicating a weak relationship. However, the embeddings also include stereotypical associations like females being `clean`, `utilized`, `silent`, `thin`

and shameful. They fit this image of women being passive and obedient, that Weitzman et al. [10] described. Overall, the adjectives used to describe females in German embeddings are substantially different from the ones used in English embeddings. There seems to be less association with superficial adjectives and much more attention to personal characteristics. These characteristics include both positive as well as stereotypical associations and are very diverse. This is a positive result, as it opens up the possibilities for girls in their development, something that was criticised in English children's literature before [6].

On the male direction, an association with negative characteristics and emotions like wrathful, queasy, lost, boring, angry, ill, stale, offensive, coarse, burdensome, lumbering and agitated can be found. This is in line with the English embeddings, where males were associated with being in a bad mood. While there are also positive adjectives associated with males (useful, pure, darling, ideal, proper and super), these are definitely outnumbered. The picture of a happy man, which was prominent in English language models, cannot be found back in German embeddings. All in all, there is diversity in the adjectives associated with the male direction, but many adjectives are negative, giving boys the idea that they are allowed to or supposed to show negative emotions towards others.

9.1.3 Verbs

While the nouns and adjectives were translated from the English lists, this was not done for the verbs. This is the case because the translation lead to many errors, of which one was that the verbs in third person singular can also be the plural noun form, e.g. forces. Those were more frequently translated to the German plural noun than the third person verb, which is not the intended form. Instead, a German verbs list was taken from [58] instead, including more than 8000 verbs in their third-person conjugation as to keep the same form as the English list. The verb list was set in relation to the German attribute lists, leading to the Ranked Lists that can be found in Table 9.3. Since some of the verbs have several meanings, the translation includes the most common interpretation, which can also be nouns. Similarly to the Ranked Lists on adjectives, the Ranked Lists on verbs show significantly higher association scores for the male direction than the female direction. Yet, the difference is not as extreme as it was with the former results.

In the DNB embeddings, it is hard to group the verbs most associated with the female gender. There are actions connected to planning and secretarial activities (staples and books), to achievement and leadership (overcomes, leads, exceeds and increases), as well as human connections (connects and unites). Many of the verbs associated with the female direction in the DNB embeddings give the character agency and power, according to the Power and Agency framework of Sap et al. [39]. This is contradictory to the findings in the English embeddings, where females were confined to being caring and fashionable, contributing to the aesthetics of a story rather than the plot. The German Gutenberg embeddings, in return, are much more stereotypical. Many of the verbs relate to household activities like (pours, accommodates, feeds, and cleans). This contrast in activities might stem from the time difference between the DNB and the German Gutenberg embeddings. With a near variance of 150 years in the mean publication date between the two embeddings, it can be inferred that the German Gutenberg embeddings contain more traditional and stereotypical results.

The results on the male direction follow a similar trend in that they are hard to group and that many of the verbs are high in agency or power. However, a difference between the male and female direction is that the former includes more verbs focused on physical actions such as shakes, drives or growls. This is in line with the English embeddings, where the male direction showed a similar trend of associating males with physical actions. Besides that, males seem to be in a position of power where they steer, cause, claim, allow or order. They hence seem to inherit power over others and their actions. Although the female direction also

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	spanisch	spanish	0.421	1.	zornig	wrathful	-0.522
2.	clever	clever	0.286	2.	mulmig	queasy	-0.386
3.	weise	wise	0.271	3.	ähnlich	similar	-0.358
4.	rund	circular	0.267	4.	nützlich	useful	-0.358
5.	doppelt	double	0.219	5.	verloren	lost	-0.305
6.	bequem	comfortable	0.207	6.	langweilig	boring	-0.283
7.	sauber	clean	0.201	7.	betroffen	concerned	-0.281
8.	trotzig	defiant	0.199	8.	rein	pure	-0.279
9.	schließen	close	0.191	9.	versteckt	hidden	-0.277
10.	gleich	equal	0.181	10.	hungrig	hungry	-0.271
11.	erfahren	experienced	0.180	11.	blind	blind	-0.267
12.	alle	all	0.162	12.	wütend	angry	-0.262
13.	verwendet	utilized	0.162	13.	chef	chief	-0.258
14.	interessant	interesting	0.145	14.	krank	ill	-0.256
15.	mutig	brave	0.143	15.	schal	stale	-0.253

(a) The Most Extreme Adjectives in the DNB Embeddings.

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	kraus	curly	0.327	1.	liebling	darling	-0.562
2.	beide	both	0.210	2.	kapital	capital	-0.455
3.	schließen	close	0.187	3.	ideal	ideal	-0.441
4.	alle	all	0.136	4.	beleidigend	offensive	-0.430
5.	leise	silent	0.111	5.	grob	coarse	-0.395
6.	silbern	silver	0.103	6.	gewachsen	grown	-0.385
7.	dünne	thin	0.103	7.	gehorsam	obedient	-0.369
8.	eindringlich	haunting	0.097	8.	beschwerlich	burdensome	-0.368
9.	persönlich	personal	0.096	9.	schal	stale	-0.366
10.	beschämend	shameful	0.094	10.	ernst	grave	-0.357
11.	weise	wise	0.094	11.	richtig	proper	-0.349
12.	leichtfertig	easy-going	0.088	12.	wertvoll	precious	-0.339
13.	geschlossen	closed	0.083	13.	super	super	-0.339
14.	ehrfürchtig	awesome	0.083	14.	schwerfällig	lumbering	-0.338
15.	eifrig	eager	0.075	15.	aufgewühlt	agitated	-0.336

(b) The Most Extreme Adjectives in the German Gutenberg Embeddings

Table 9.2: The Most Extreme Adjectives as Projected on the Gender Direction

included words with high power according to the connotation framework of Sap et al. [39], their verbs were not inherently about guiding the actions of others but rather about agency. For example, `overcoming` and `exceeding` are more about the individual's actions and stance, whereas `allowing` and `ordering` are about the actions of others and the individual's control over them.

9.1.4 Animals

The German target list of animals was obtained through a translation of the English list. The association scores of animals with different gender directions can be found in Table 9.4. The most female associated animals seem to continue the trend from the English embeddings in so far as

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	enthält	contains	0.372	1.	lenkt	steers	-0.534
2.	heftet	staples	0.367	2.	kuschelt	cuddles	-0.492
3.	demoliert	demolishes	0.367	3.	bewirkt	causes	-0.452
4.	flucht	escape/swears	0.320	4.	gleicht	resembles	-0.379
5.	bucht	books	0.297	5.	schenkt	gifts	-0.363
6.	ergattert	snags	0.260	6.	schüttelt	shakes	-0.362
7.	erneut	again	0.248	7.	treibt	drives	-0.351
8.	überwindet	overcomes	0.240	8.	wünscht	wishes	-0.343
9.	führt	leads	0.235	9.	behauptet	claims	-0.332
10.	übertrifft	exceeds	0.225	10.	verpasst	misses	-0.332
11.	doppelt	double	0.219	11.	entkommt	escapes	-0.317
12.	explodiert	explodes	0.219	12.	klaut	steals	-0.317
13.	erhöht	increases	0.217	13.	guckt	looks	-0.317
14.	verbindet	connects	0.216	14.	fällt	falls	-0.309
15.	vereint	unites	0.215	15.	heult	wails	-0.302

(a) The Most Extreme Verbs in the DNB Embeddings

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	bezieht	obtains	0.254	1.	gast	guest	-0.562
2.	flucht	escape/swears	0.236	2.	begegnet	meets	-0.468
3.	furcht	fear/furrows	0.217	3.	erlaubt	allows	-0.406
4.	gießt	pours	0.205	4.	knurrt	growls	-0.401
5.	braut	bride/brews	0.201	5.	gesteht	confesses	-0.369
6.	haut	skin/hits	0.199	6.	heischt	hails	-0.361
7.	schnurrt	purrs	0.186	7.	vertilgt	eats	-0.356
8.	ereilt	befalls	0.172	8.	verspricht	promises	-0.353
9.	beherbergt	accommodates	0.168	9.	bestellt	orders	-0.351
10.	klatscht	claps	0.157	10.	passiert	happens	-0.346
11.	bindet	binds	0.156	11.	erstickt	suffocates	-0.342
12.	füttert	feeds	0.152	12.	verlobt	engaged	-0.335
13.	faltet	folds	0.148	13.	berechnet	calculates	-0.333
14.	putzt	cleans	0.147	14.	schreitet	strides	-0.331
15.	prüft	checks	0.138	15.	spiegelt	mirrors	-0.330

(b) The Most Extreme Verbs in the German Gutenberg Embeddings.

Table 9.3: The Most Extreme Verbs as Projected on the Gender Direction

that females are associated with birds of all kinds (`nightingale`, `albatross`, `magpie`, `goose` and `duck`). One reason for this might be once again the shared adjectives used to describe females as well as birds such as `beautiful`. Also the trend of female association with insects remains (`caterpillar`, `honey bee`, `wasp` and `fly`). While this relationship might also stem from common adjectives, another reason can be the grammatical gender of the insects, as all have the grammatical gender female. Besides that, there are some stereotypical associations with animals like `goat`, `snake`, `cow` and `mouse`. These are stereotypical as these animals are often used to describe females or as an insult. In German, common insults for girls and women are “Du blöde Ziege!” and “Du dumme Kuh!”, meaning “You are a stupid goat/cow”. They are exclusively used for females, which can explain the female association. Another insult is “Du

falsche Schlange!" (translation: "You are a snake in the grass!"), which is used mostly for females and describes a lying and deceiving individual. PONS Online dictionary even translates the saying as "back-stabber (woman)" [62], showing the intrinsic connection between females and snakes. The association of `mouse` with females stems from a more positive origin. The words `Maus` or `Mäusschen` are often used as nicknames for girls.

In comparison, males are associated with `horse`, `dog`, `bird` and `fish`, animals that were already in the ranked nouns list. For the last three, the association could be explained by the grammatical gender as their articles are "der". Moreover, `dog` is gendered in German and the translation refers to a male dog. Besides these, men are associated with predators like `wolf`. `fox`, `lion` and `bear`, which was also the case in the English embeddings. Yet, the trend is much less prominent in the German embeddings. Instead, the animals are quite diverse ranging from birds over mammals to amphibians.

Once more, the grammatical gender of the words could be influencing the results. On the female direction, all words besides a few exceptions, in particular, `der Ablatros` and `das Meerschweinchen`, have the grammatical gender female. On the male direction, in return, all words are of male or neutral grammatical gender, just as was the case for the nouns list. Another interesting point is the repeating observation of high association scores with the male direction and lower bias scores with the female direction.

9.2 Word-Embedding Association Test

Last, the Word-Embedding Association Test was fitted on the German embeddings. For this, the original sets from Caliskan et al. [26] as well as Chaloner and Maldonado [31] were used. Also, the hand-made lists from Chapter 7 were fitted on the sets. To do so, translation of the attribute as well as target lists was needed. These translations can be found in Appendix G, together with the out of vocabulary words for each German language model. For the profession lists, both male and female versions of each occupation were included. Both lists include both gender versions to combat the significance stemming from grammatical gender alone. While this was feasible for the target lists in the WEATs due to their small size, this was not the case for the Ranked Lists target lists due to the large list sizes. Hence, this option was only chosen for the WEAT target lists.

When looking at the results of the WEAT in Table 9.5, it can quickly be seen that the German language models have only a few significant results. For both, W1 (career vs. family) was significant with Cohen's ds comparable to the English language models tested in Chapter 6 and 8. This shows that the difference in the association of males and females with career and family seems to be widespread and even exists across cultures and languages. This might be due to the similarity in cultural history. Both languages stem from traditionally Christian countries, where men tended to work and females tended to stay at home with children and household chores. These Christian values have shaped societies for hundreds of years and remain, to this time, an important anchor of morals and tradition. However, with modernisation, this influence has shrunk as women began to gain more rights and independence with the women's movement in the 1960s. This can also be seen in the effect sizes of the language models. The embeddings incorporating older literature (German and English Gutenberg) have a higher effect size than the language model with more modern books (DNB and OCLC). Hence, this shift in society, which was also observed in social science literature, can be confirmed computationally.

Besides W1, the DNB language model only shows significant gender bias in W9, where males are associated significantly more with active games like running and riding a bike while girls are more associated with quiet games like reading and painting. Given the conventional interpretation of Cohen's d, the effect size is medium. This is much lower than the very high values recorded in the English language models which could be attributed to the difference in corpus size. The German Gutenberg model records a small effect size for this WEAT with

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	raupe	caterpillar	0.295	1.	pferd	horse	-0.528
2.	katze	cat	0.234	2.	hund	dog	-0.487
3.	ziege	goat	0.219	3.	wolf	wolf	-0.372
4.	schlange	snake	0.202	4.	mensch	human	-0.364
5.	nachtigall	nightingale	0.198	5.	schwein	pig	-0.362
6.	albatros	albatross	0.178	6.	esel	donkey	-0.350
7.	qualle	jellyfish	0.178	7.	vogel	bird	-0.341
8.	maus	mouse	0.176	8.	rotkehlchen	robin	-0.320
9.	honigbiene	honey bee	0.161	9.	eichhörnchen	squirrel	-0.314
10.	elster	magpie	0.159	10.	fuchs	fox	-0.306
11.	türkei	turkey	0.142	11.	maulwurf	mole	-0.288
12.	fledermaus	bat	0.141	12.	gürteltier	armadillo	-0.287
13.	gans	goose	0.137	13.	pinguin	penguin	-0.287
14.	hyäne	hyena	0.128	14.	löwe	lion	-0.280
15.	kuh	cow	0.125	15.	frosch	frog	-0.276

(a) The Most Extreme Animals in the DNB Embeddings

	Extreme *female*	Translation	**Bias**		**Extreme *male***	Translation	**Bias**
1.	schlange	snake	0.228	1.	hund	dog	-0.618
2.	kuh	cow	0.207	2.	esel	donkey	-0.613
3.	nachtigall	nightingale	0.194	3.	vogel	bird	-0.608
4.	wespe	wasp	0.171	4.	fisch	fish	-0.591
5.	katze	cat	0.154	5.	wolf	wolf	-0.541
6.	gans	goose	0.149	6.	papagei	parrot	-0.541
7.	fledermaus	bat	0.131	7.	löwe	lion	-0.538
8.	ente	duck	0.130	8.	igel	hedgehog	-0.531
9.	raupe	caterpillar	0.118	9.	fuchs	fox	-0.520
10.	ziege	goat	0.114	10.	mensch	human	-0.520
11.	elster	magpie	0.106	11.	hase	hare	-0.519
12.	fliege	fly	0.099	12.	bär	bear	-0.509
13.	ratte	rat	0.098	13.	frosch	frog	-0.508
14.	meerschweinchen	guinea pig	0.090	14.	schwan	swan	-0.464
15.	schnecke	snail	0.090	15.	adler	eagle	-0.443

(b) The Most Extreme Animals in the German Gutenberg Embeddings

Table 9.4: The Most Extreme Animals as Projected on the Gender Direction

a p-value just above the threshold of $\alpha = 0.05$. This shows that this difference has possibly existed for a longer period but has increased, instead of decreased, over time. A factor that could have influenced the significance of WEATs is the vocabulary. Many of the target words were not known in the German language models, however, only two words from W9 were out of vocabulary. This shows that either more common words were used in this WEAT or that the activities in W9 are more present in the corpora than the words from the other WEATs.

In relation to this, one can see that there are no results for the male and female professions for DNB as all words were out of vocabulary. In return, the German Gutenberg embeddings record a significant difference between the two groups of professions. While this might point to a difference in types of professions males and females go after, it could also be attributed to

the words that are out of vocabulary. In the X target list (hypothesized male professions), five pairs of professions and three female versions of professions were unknown. While the pairs will not influence the results, the three missing female versions can lead to a male bias of the target list as it contains more male than female-gendered professions. Similarly, the Y target list (hypothesized female profession) is missing six pairs of professions and one male version of a profession. Hence there are slightly more female versions in the Y target list, which could lead to female bias. Given the missing words in the two target lists, the significant difference between the X and Y professions could be traced back to the imbalance of male and female versions in the lists. This is also an interesting observation as to show which professions are used in the corpus at all with male versions being much more common than female ones. This shows an intrinsic bias of the text towards males in occupations.

Besides that, the German Gutenberg embeddings showed significant gender bias for strengths and weaknesses (W5) and male and female school subjects (W14). The effect sizes are comparable to the respective English Gutenberg results, with large Cohen's ds. This shows that also in German children's books, females are seen as weak while males are portrayed as strong, a result that corresponds to my personal bias as well. Moreover, there is a difference in the school subjects that children are associated with, with boys enjoying the natural sciences and girls languages and the humanities. Also, this is a difference that confirms social science literature, the results of Kurpicz-Briki [33] as well as my personal biases and experiences. In German schools and universities, boys tend to choose technical and natural science subjects, whereas girls choose humanities and law [63].

		OCLC	**English Gutenberg**	**DNB**	**German Gutenberg**
	WEAT category	*d*	*d*	*d*	*d*
W1:	career vs. family	**0.839**	**1.755**	**0.729**	**1.349**
W2:	maths vs. arts	**1.078**	**1.052**	-0.072	-1.036
W3:	science vs. arts	**1.446**	**1.743**	0.329	0.006
W4:	intelligence vs. appearance	0.114	**0.809**	-0.792	-0.615
W5:	strength vs. weakness	-0.152	**0.820**	0.174	**0.766**
W6:	outdoor vs. indoor	0.185	**1.258**	-0.271	0.291
W7:	male vs. female toys	**1.737**	**1.302**	0.850	0.804
W8:	male vs. female sports	**1.550**	**0.715**	-0.473	-0.466
W9:	active vs. quiet games	**1.259**	**1.215**	**0.520**	0.139*
W10:	male vs. female adjective	-0.949	-0.898	0.681	-0.068
W11:	male vs. female professions	**1.640**	**1.304**	**	**1.251**
W12:	male vs. female adjectives [23]	0.537	0.102	-1.193	-0.991
W13:	dominant vs. obedient	**0.837**	**1.282**	-0.326	-0.769
W14:	male vs. female school subjects	**1.101**	**0.829**	0.568	**1.238**

* Nearly significant with p-value = 0.054

** No result as all words were out of vocabulary.

Table 9.5: Results of the WEAT Tests. Values in Bold Indicate Statistically Significant Gender Bias ($p < 0.05$).

9.3 Conclusion and Discussion

To answer the question of gender bias in other cultures, German children's books were analysed. Two language models were trained, one with modern book descriptions (DNB embed-

dings) and one with older full-text books (German Gutenberg embeddings). Ranked Lists show that there is a substantial difference between German and English gender bias when it comes to the adjectives used to describe females. The latter focuses on superficial characteristics of women, paying little attention to skills and other character attributes. The German embeddings, in return, focused more on personal characteristics and skills, giving more freedom to girls in their development. On the male direction, both German and English embeddings paint a picture of men and boys with negative moods, indicating similarity in the portrays of boys in the two languages. This trend, where the female direction differs across language and the male direction remains largely similar, can also be seen in the verbs and animals target lists. In the English language models, females are portrayed as being caring and a certain focus on appearances and looks can be seen. In the German embeddings, females are granted more diversity due to a wider variety of associated words. The male direction, in return, is always characterised by diversity and males are frequently portrayed as being physical, large or powerful.

When it comes to the WEAT, German embeddings show a clear distinction between males and females and their association with career and family terms. This is in line with the expectations and similarities to English language models seem to stem from a common history in Christianity and patriarchy.

Yet, the influence of grammatical gender should not be discarded lightly. Nearly all nouns in the female direction of the Ranked Lists have the grammatical gender female while the male direction is shaped by male and neutral grammatical gender. Moreover, translation issues and specificities of languages can influence results. For example, the German language has gendered professions into male and female versions. Keeping this in mind, it does not make sense to simply compare male and female professions to their gender association, as the gender difference is inherent in the words themselves. Alternative ways of comparing gendered words in their bias are needed. In addition, it is time-consuming to translate and correct the target lists. Instead, German lists can be drawn upon right away, as was done for the verb lists. However, this might limit comparability across languages if there is little overlap in target list words. Furthermore, high-quality embeddings are needed to get reliable results that can be generalised. Especially for less common languages, finding quality data can pose a challenge. Although German is a widely used language with a long history in poetry and literature, the datasets were much smaller than the English corpora. This influenced language model development and likely also quality.

All in all, the German language seems to be less stereotypical than the English language or expose different kinds of biases. However, translation issues, language model quality and grammatical gender should be taken into consideration when analysing the results.

10 CONCLUSION AND FUTURE WORK

Having answered all sub-research questions, the conclusion of this research can be drawn. In this Chapter, the conducted research is summarised and the overarching research question is answered. Subsequently, the contributions to practice and research are pointed out, finishing with recommendations for future work.

10.1 Conclusion and Answer to Research Question

In this research, the question was posed **to what extent language models can be used to examine gender bias in children's books across time and culture?** This question was answered in four parts. First, the applicability of current methods to children's literature was investigated. This was done using Bolukbasi et al.'s [27] methods Ranked Lists and Analogies as well as Caliskan et al.'s [26] methods WEAT and WEFAT on the OCLC embeddings. This yielded interesting, significant and consistent results, showing the effectiveness of the methods on a new domain in large parts. However, out-of-vocabulary words and domain relevance limited the ability of the WEAT to find gender bias in the OCLC embeddings, leading to the need to adapt the WEAT lists.

Following up on this limitation, the UBE algorithm of Swinger et al. [34] was leveraged to find WEAT lists automatically in the OCLC language model. This method proved difficult to evaluate due to a lack of quantitative measures. Moreover, coherence within and across WEAT lists left much to be desired. As an alternative, WEAT lists were drawn up by hand, countering issues of vocabulary use and coherence. Six out of nine manually crafted lists yielded significant differences between the genders. The effect sizes were very high, establishing WEAT lists that are specific to children's literature. This helped to confirm biases that were discovered in social science literature computationally while enhancing generalisability due to the size of the dataset.

After establishing the applicability of methods to find gender bias in children's books and adapting the WEAT, the third sub-research question investigated the change of bias over time. For this, Project Gutenberg [41] was leveraged for English books covering the period of 1800-1925. Embeddings trained on this corpus were analysed through Ranked Lists, Analogies, WEAT and WEFAT. Overall, it can be said that the English Gutenberg embeddings contain gender bias comparable to the bias in the OCLC embeddings. Many differences can be attributed to a change in language as more old fashioned words are used in the English Gutenberg corpus. The male bias seems to shift away from military associations while being replaced by many different words rather than a new trend. Female gender bias, in return, remained largely the same with women and girls being associated with inter-personal relationships and appearance.

Moving from the time axis to the cultural axis, the last sub-research question examined the difference in gender bias in English and German embeddings. This was analysed using Ranked Lists and the WEATs. All in all, the German embeddings show a difference between males and females, where both are marked by much more diversity than granted by the English language models. The objectification of females seems to be less present in German children's books than in English ones. Yet, the results are likely to be heavily influenced by the presence of grammatical gender, language model quality and translation issues.

10.2 Contribution to Practice and Research

This work was motivated by the goal of shedding light on gender bias in the literature addressing a young audience. The present research, therefore, offers contributions to practice and research.

On the one hand, the research has shown that gender bias is present in English and German children's books as well as in their descriptions. Leveraging computational possibilities, a vast amount of books and descriptions could be analysed to paint a picture of gender bias in society. By making authors, parents and educational authorities more aware of the biases present in children books, they can be more sensitised to it. This gives parents and guardians the possibility to be more selective in the readings they provide to their children while being more sensitive towards the consequences of biased reading. Gender bias can then be identified as hampering a child's development and combated more effectively.

On the other hand, the research has computationally confirmed social science findings and suspicions. By immensely increasing the number of books that are studied, the scope of research could be widened. More books, years and languages have been analysed, leading to greater generalisability of results. On top of that, this work has helped to establish the applicability of current methods to find gender bias in new domains. This showed that some methods need adjusting to new domains whenever a difference in language use and topic is expected.

10.3 Recommendations and Future Work

Future work is recommended to address the limitations discussed in the relevant chapters while trying to improve the children language models. An individual issue that needs addressing is finding profession target lists for the Ranked Lists method. In this work, time and resource constraints have made it difficult to pass from ungendered to gendered word lists. As a result, no analysis was conducted on the most male and female-biased professions in the German embeddings. Future work should address the possibilities of applying gender bias methods to inherently gendered words. For example, this could be done by including both male and female versions of the profession, applying stemming or using lemmatisation. Moreover, the effect of grammatical gender on the gender bias results needs further investigation. The most female associated nouns in the German embeddings had the grammatical gender female (female article "die") and the most male associated nouns had the grammatical gender neutral or male (articles "das" and "der", respectively). This, in combination with the results being unexpected and not in line with the English results, raised questions on the influence of grammatical gender on gender direction association. Future work should look into how big this influence is and whether it can be extracted to retrieve results that are based on gender bias rather than grammatical influences.

Another field that requires further investigation is the Unsupervised Bias Enumeration Algorithm (UBE). This method, which automatically detect WEATs, yielded significant results for nearly all clusters. Yet, it was unclear to the eye why the lists were gender-biased and how they connected to each other. Therefore, work is needed on how to find more semantically coherent clusters in word embeddings, an issue that Chaloner and Maldonado [31], as well as Swinger et al. [34] also faced. Moreover, the WEATs were significant with high effect sizes, but the information they contained on gender bias was unclear. Hence, a metric for evaluation is needed to estimate the quality of the automatically generated WEAT lists and their resulting statement on gender bias.

Besides fixing issues of this research, future work can expand and confirm the information in this report by drawing on the opinion of others. To a large extent, the interpretation of gender bias results was based on my personal lens. By including expert opinion or crowd-worker evaluation, the results of this report could be interpreted and confirmed through qualitative and

quantitative measures. This could increase the impact, reliability and generalisability of the results, as several views are taken into account.

Expanding on the knowledge that children's literature contains a substantial degree of gender bias, other research could build on current knowledge on debiasing and apply methods to soften gender bias in children language models. Debiased embeddings could then be used to generate children's stories computationally, offering less biased readings to children and parents. Furthermore, future work could build applications that can point out gender bias in particular books, helping authors to identify gender bias in the writing process. This could give the opportunity to writers to adjust their children's stories to be less biased while still in the writing process.

REFERENCES

[1] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective post-processing for word representations. URL http://arxiv.org/pdf/1702.01417v2.

[2] United Nations Sustainable Development. Gender equality and women's empowerment, 15.02.2021. URL https://www.un.org/sustainabledevelopment/gender-equality/.

[3] Mino Vianello and Mary Hawkesworth, editors. *Gender and Power: Towards Equality and Democratic Governance*. SpringerLink : Bücher. Palgrave Macmillan, London, 2016. ISBN 978-1-137-51415-8.

[4] Brian A. Nosek, Mahzarin Banaji, and Anthony G. Greenwald. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101–115, 2002. ISSN 1089-2699. doi: 10.1037//1089-2699.6.1.101.

[5] Brian A. Nosek, Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, Selin Kesebir, Norbert Maliszewski, Félix Neto, Eero Olli, Jaihyun Park, Konrad Schnabel, Kimihiro Shiomura, Bogdan Tudor Tulbure, Reinout W. Wiers, Mónika Somogyi, Nazar Akrami, Bo Ekehammar, Michelangelo Vianello, Mahzarin R. Banaji, and Anthony G. Greenwald. National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10593–10597, 2009. doi: 10.1073/pnas.0809921106.

[6] Ruth E. Hartley. Sex-role pressures and the socialization of the male child. *Psychological Reports*, 5(2):457–468, 1959. doi: 10.2466/pr0.1959.5.h.457.

[7] Daniel G. Brown. Sex role preferences in young children. *Psychological Monograph*, 70 (14):1–19, 1956.

[8] E. H. Spitz. *Inside Picture Books*. Yale University Press, 1999.

[9] Timothy J. Frawley. Gender schema and prejudicial recall: How children misremember, fabricate, and distort gendered picture book information. *Journal of Research in Childhood Education*, 22(3):291–303, 2008. ISSN 0256-8543. doi: 10.1080/02568540809594628.

[10] Lenore J. Weitzman, Deborah Eifler, Elizabeth Hokada, and Catherine Ross. Sex-role socialization in picture books for preschool children. *American Journal of Society*, 77(6): 1125–1150, 1972.

[11] Nancy K. Schlossberg and Jane Goodman. A woman's place: Children's sex stereotyping of occupations. *Vocational Guidance Quarterly*, 20(4):266–270, 1972. ISSN 00427764. doi: 10.1002/j.2164-585X.1972.tb02062.x.

[12] Lisa K. Barclay. The emergence of vocational expectations in preschool children. *Journal of Vocational Behavior*, 4(1):1–14, 1974. ISSN 00018791. doi: 10.1016/0001-8791(74) 90086-4.

[13] Susan Knell and Winer, Gerald, A. Effects of reading content on occupational sex role stereotypes. *Journal of Vocational Behavior*, 14:78–87, 1979. ISSN 00018791.

[14] World Bank Blogs. Can gender equality prevent violent conflict?, 16.02.2021. URL `https://blogs.worldbank.org/dev4peace/can-gender-equality-prevent-violent-conflict`.

[15] Dominique Geißler. Research topics: A computational analysis of gender bias in children's books, 2021.

[16] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. *Proceedings of SIGCIS*, 2017.

[17] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, 2019.

[18] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, 1996. ISSN 1046-8188. doi: 10.1145/230538.230561.

[19] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485.

[20] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Soclal Psychology*, 74(6):1464–1480, 1998.

[21] S. M. Czaplinski. *Sexism in award winning picture books*. Know, Pittsburgh, PA, 1972.

[22] Anita P. Davis and Thomas R. McDaniel. Rapid research report: You've come a long way, baby: Or have you? research evaluating gender portrayal in recent caldecott-winning books. *The Reading Teacher*, 52(5):532–536, 1999.

[23] Carole M. Kortenhaus and Jack Demarest. Gender role stereotyping in children's literature: An update. *Sex Role*, 28:219–232, 1993.

[24] Ambreen Shahnaz, Syeda Tamkeen Fatima, and Samina Amin Qadir. 'the myth that children can be anything they want': gender construction in pakistani children literature. *Journal of Gender Studies*, 29(4):470–482, 2020. ISSN 0958-9236. doi: 10.1080/09589236.2020.1736529.

[25] Janice McCabe, Emily Fairchild, Liz Grauerholz, Bernice A. Pescosolido, and Daniel Tope. Gender in twentieth-century children's books. *Gender & Society*, 25(2):197–226, 2011. ISSN 0891-2432. doi: 10.1177/0891243211398358.

[26] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.)*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230.

[27] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *30th Conference on Neural Information Processing Systems*, pages 4356–4364, 2016. URL https://dl-acm-org.ezproxy2.utwente.nl/doi/pdf/10.5555/3157382.3157584.

[28] Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensationall: Man is to doctor as woman is to doctor. *Association for Computational Linguistics*, 46: 487–497, 2020.

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In Qatar Computing Research Institute Alessandro Moschitti, Google Bo Pang, and University of Antwerp Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

[30] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *CoRR*, abs/1711.08412, 2018. doi: 10.1073/pnas.1720347115.

[31] Kaytlin Chaloner and Alfredo Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 25–32, 2019. doi: 10.18653/v1/W19-3804.

[32] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguous pronouns. URL http://arxiv.org/pdf/1810.05201v1.

[33] Mascha Kurpicz-Briki. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. *Proceedings of 5th SwissText & 16th KONVENS Joint Conference 2020 (Zurich)*, 2020.

[34] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan, IV, Mark D. M. Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? *AIES 2019: the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2018. URL http://arxiv.org/pdf/1812.08769v4.

[35] Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 18–24, 2019. doi: 10.18653/v1/W19-3803.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[37] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *CoRR*, 2019.

[38] Hannah Rashkin, Sameer Singh, and Yejin Choi. Connotation frames: A data-driven investigation. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 311–321, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1030.

[39] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1247.

[40] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *ICLR*, 2016. URL http://arxiv.org/pdf/1511.02301v4.

[41] Project Gutenberg. Project gutenberg, 18.02.2021. URL https://www.gutenberg.org/.

[42] Deutsche Nationalbibliothek. Startseite, 06.04.2021. URL https://www.dnb.de/DE/Home/home_node.html.

[43] GNU Operating System. Gnu wget - free software foundation, 23.03.2021. URL http://www.gnu.org/software/wget/.

[44] Gareth Johnson. Gutendex, 2017. URL http://gutendex.com/.

[45] OCLC. Viaf, 2010-2021. URL http://viaf.org/.

[46] Kenneth Reitz. Requests: Python http for humans., 2011.

[47] Deutsche National Bibliothek. Sru-schnittstelle, 06.04.2021. URL https://www.dnb.de/EN/Professionell/Metadatendienste/Datenbezug/SRU/sru_node.html.

[48] Leonard Richardson. Beautiful soup documentation. *April*, 2007.

[49] Michal Mimino Danilak. langdetect: Language detection library ported from google's language-detection., 2014.

[50] Radim Řehůřek. Gensim: Word2vec, 2009.

[51] Hugsy. english-adjectives.txt, 27.05.2021. URL https://gist.github.com/hugsy/8910dc78d208e40de42deb29e62df913.

[52] Alberto Scorrano. animals.txt, 08.06.2021. URL https://github.com/skjorrface/animals.txt.

[53] H. L. James. Compare embedding bias, 2019. URL https://github.com/hljames/compare-embedding-bias.

[54] Aylin Caliskan. Replication data for: Wefat and weat, 2017. URL https://dataverse-harvard-edu.ezproxy2.utwente.nl/dataset.xhtml?persistentId=doi:10.7910/DVN/DX4VWP.

[55] Aditya Maini. Weat-wefat, 17.06.2021. URL https://github.com/adimaini/WEAT-WEFAT.

[56] L. Shaw-Taylor, E. A. Wrigley, and P. Kitson. 1851 census report: County occupational data.

[57] DeepL. Deepl translator, 18.08.2021. URL https://www.deepl.com/en/translator.

[58] Viorelsfetea. german-verbs-database, 11.08.2021. URL https://github.com/viorelsfetea/german-verbs-database.

[59] Eric Wilson. Evolution of the fashion industry. URL https://fashion-history.lovetoknow.com/fashion-clothing-industry/evolution-fashion-industry.

[60] Barbara Irene Kreps. The paradox of women: The legal position of early modern wives and thomas dekker's the honest whore. *ELH*, 69(1):83–102, 2002. doi: 10.1353/elh.2002.0007.

[61] ushistory.org. Women' rights: U.s. history online textbook, 2021. URL //www.ushistory.org/us/26c.asp.

[62] PONS. Online dictionary: Translation for "falsche schlange", 23.08.2021. URL https://en.pons.com/translate/german-english/falsche+Schlange.

[63] Statistisches Landesamt Rheinland-Pfalz. Girls' and boys' day: Berufs- und studienwahl junger frauen und männer, 13.09.2021. URL https://www.statistik.rlp.de/no_cache/de/gesamtwirtschaft-umwelt/verdienste-und-arbeitskosten/pressemitteilungen/einzelansicht/news/detail/News/2183/.

Appendix A GUTENBERG EXAMPLE BOOK OUTLINE

```
The Project Gutenberg EBook of TITLE, by AUTHOR

This eBook is for the use of anyone anywhere in the United States and most
other parts of the world at no cost and with almost no restrictions
whatsoever.  You may copy it, give it away or re-use it under the terms of
the Project Gutenberg License included with this eBook or online at
www.gutenberg.org.  If you are not located in the United States, you'll have
to check the laws of the country where you are located before using this ebook.

Title: EXAMPLE TITLE

Author: EXAMPLE AUTHOR

Release Date: MONTH, YEAR [EBook #ID]

Language: English

Character set encoding: ENCODING

*** START OF THIS PROJECT GUTENBERG EBOOK THE WONDERFUL WIZARD OF OZ ***


[CONTENT OF THE BOOK]


End of the Project Gutenberg EBook of TITLE, by AUTHOR

*** END OF THIS PROJECT GUTENBERG EBOOK THE WONDERFUL WIZARD OF OZ ***

***** This file should be named ID.txt or ID.zip *****
This and all associated files of various formats will be found in:
        http://www.gutenberg.org/ID/ID/

Updated editions will replace the previous one--the old editions will
be renamed.

Creating the works from print editions not protected by U.S. copyright
law means that no one owns a United States copyright in these works,
so the Foundation (and you!) can copy and distribute it in the United
States without permission and without paying copyright
royalties. Special rules, set forth in the General Terms of Use part
of this license, apply to copying and distributing Project
Gutenberg-tm electronic works to protect the PROJECT GUTENBERG-tm
concept and trademark. Project Gutenberg is a registered trademark,
and may not be used if you charge for the eBooks, unless you receive
specific permission. If you do not charge anything for copies of this
eBook, complying with the rules is very easy. You may use this eBook
for nearly any purpose such as creation of derivative works, reports,
performances and research. They may be modified and printed and given
away--you may do practically ANYTHING in the United States with eBooks
not protected by U.S. copyright law. Redistribution is subject to the
trademark license, especially commercial redistribution.

START: FULL LICENSE

THE FULL PROJECT GUTENBERG LICENSE
PLEASE READ THIS BEFORE YOU DISTRIBUTE OR USE THIS WORK

To protect the Project Gutenberg-tm mission of promoting the free
distribution of electronic works, by using or distributing this work
(or any other work associated in any way with the phrase "Project
Gutenberg"), you agree to comply with all the terms of the Full
Project Gutenberg-tm License available with this file or online at
```

Appendix B ENGLISH WEATS

	Attribute words	M	male, man, boy, brother, he, him, his, son, father, uncle, grandfather
		F	female, woman, girl, sister, she, her, hers, daughter, mother, aunt, grandmother
Target words	W1: career vs. family	X	executive, management, professional, corporation, salary, office, business, career
		Y	home, parents, children, family, cousins, marriage, wedding, relatives
	W2: math vs. arts	X	math, algebra, geometry, calculus, equations, computation, numbers, addition
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	W3: science vs. arts	X	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	W4: intelligent vs. appearance	X	precocious, resourceful, inquisitive, genius, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, apt, venerable, imaginative, shrewd, thoughtful, wise, smart, ingenious, clever, brilliant, logical, intelligent
		Y	alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong
	W5: strength vs. weakness	X	power, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, shout, dynamic, winner
		Y	weak, surrender, timid, vulnerable, weakness, wispy, withdraw, yield, failure, shy, follow, lose, fragile, afraid, loser

Appendix C ADJUSTED ANALOGY TEST SET

C.1 Adjusted analogy categories English

- family
- gram1-adjective-to-adverb
- gram2-opposite
- gram3-comparative
- gram4-superlative
- gram5-present-participle
- gram7-past-tense
- gram8-plural
- gram9-plural-verbs

C.2 Adjusted analogy categories German

- family
- gram2-opposite
- gram3-comparative
- gram8-plural

Appendix D WEFAT RESULTS SUB-RQ1

Professions that were out of vocabulary for the OCLC embeddings:

- nutritionist, pharmacist, appraiser, salesperson, bartender, electrician, paralegal, receptionist, pathologist

The following table presents the results of the WEFAT test on the OCLC embeddings. The first column shows the profession, the second column the percentage of female workers in real world data p_w, the third column shows the effect size (calculated using equation 5.6) and the fourth column the p-values. P-values in bold indicate statistically significant association with one of the genders ($p < 0.05$).

Profession	Percentage of women	Effect size	p-value
plumber	0.007	-1.114	0.994
mechanic	0.018	-1.181	0.997
carpenter	0.021	-0.821	0.975
machinist	0.067	-1.021	0.992
engineer	0.107	-0.712	0.949
programmer	0.184	-0.244	0.720
architect	0.208	-0.307	0.751
officer	0.279	-0.692	0.944
paramedic	0.329	0.190	0.330
janitor	0.343	-0.650	0.944
inspector	0.344	-1.030	0.993
lawyer	0.345	-0.595	0.918
chemist	0.361	0.099	0.412
worker	0.367	-0.141	0.621
advisor	0.379	-0.355	0.791
physician	0.379	-0.234	0.715
surgeon	0.379	-0.458	0.869
manager	0.385	-0.729	0.957
supervisor	0.386	-0.187	0.662
technician	0.403	-0.183	0.665
specialist	0.412	-0.145	0.643
scientist	0.419	-0.717	0.955
investigator	0.451	-0.021	0.524
administrator	0.549	-0.195	0.665
examiner	0.569	0.054	0.446
accountant	0.597	-0.122	0.604
veterinarian	0.605	-0.222	0.686
baker	0.608	-0.639	0.938
instructor	0.623	0.107	0.405
counselor	0.665	0.504	0.119
clerk	0.695	-0.870	0.979
psychologist	0.703	-0.009	0.535
educator	0.708	0.174	0.364
teacher	0.710	0.209	0.326
practitioner	0.748	0.135	0.353
therapist	0.767	0.309	0.243
planner	0.786	**1.068**	**0.007**
librarian	0.830	0.143	0.371
nurse	0.896	0.432	0.147
hairdresser	0.942	**1.308**	**0.001**
hygienist	0.964	-0.532	0.890

Appendix E AUTOMATICALLY CREATED WEAT LISTS

The following 64 WEAT lists were created from the postprocessed OCLC Embeddings using the top M = 30.000 words.

WEAT	X Targets	Y Targets	p value	Cohens d
0	egypt, pyramids, knights, mummies, caves, temple, dungeons, greece	ballet, princess, sparkle, princesses, cinderella, twilight, queen, sparkling	0.000	1.915
1	bites, farts, maclary, dodging, inventions, twoheaded, nosepickingly, pranks	stepsisters, entanglements, manipulative, flair, gossip, charm, romances, charms	0.000	1.919
2	wally, steve, ricky, buddy, bob, gary, jerry, trevor	zoey, alyssa, angelina, actress, isabelle, stacey, cheerleader, melody	0.000	1.931
3	nephew, cody, uncle, bill, harold, carl, johnny, hank	grace, rose, olivia, victoria, elizabeth, lynn, bedelia, sabrina	0.000	1.933
4	aboard, sailing, san, trading, deserted, uncles, bay, remote	glamorous, fashion, privileged, fancy, wedding, opera, luxurious, attending	0.000	1.844
5	wifes, error, continually, tosses, drunk, disobeys, yells, injures	resents, surrogate, marrying, strict, heavenly, casts, breakup, girlfriends	0.000	1.912
6	his, scrooge, browns, toms, hanks, mikes, jimmys, tommys	roses, elizabeths, annes, fairys, kirstys, daisys, ellas, olivias	0.000	1.904
7	reindeer, santas, nuts, elves, snacks, nest, toys, carrots	dolls, sweet, flowers, cupcakes, hearts, celebrating, baking, decorations	0.000	1.871
8	willy, himself, raccoon, tyrannosaurus, beaver, rat, billy, farmer	mermaid, butterfly, hen, bee, mrs, lamb, winnie, kitten	0.000	1.899
9	careless, spotted, chased, dirty, invented, eaten, outwitted, talking	absolutely, babysitting, heartbroken, adored, spoiled, hers, confident, thrilled	0.000	1.933
11	growling, trumpet, gull, fishes, roars, naked, swims, sniffs	rapunzel, bloom, petals, lovely, frilly, ballroom, sunshine, swan	0.000	1.921
13	seagull, gruff, python, mule, policeman, ratty, pigeon, taxi	rag, yaga, auntie, miss, spinster, madame, duchess, dolores	0.000	1.930

14	sailor, jolly, rusty, battered, hermit, sheepdog, cranky, nemesis	lady, plain, ladys, year, magick, womans, years, sixteenyear	0.000	1.813
15	tank, donald, jet, gun, postman, truck, bulldozer, pickup	cupcake, gown, doll, dresses, sparkly, tiara, pink, necklace	0.000	1.932
16	drive, catch, fly, accompany, search, help, guide, join	unravel, invite, remember, navigate, hold, share, make, sing	0.000	1.822
17	hunter, robots, masters, guards, companions, men, vikings, hunters	sapphire, barda, teens, lief, faeries, individuals, guardian, vampires	0.000	1.864
18	vehicles, jokes, antics, facts, various, hilarious, sorts, tabs	diverse, lyrical, disney, diversity, celebration, delightful, backgrounds, charming	0.000	1.892
19	rodeo, orders, varsity, races, scouts, bet, soccer, trophy	recital, babysitters, pageant, spotlight, homecoming, dancers, prom, committee	0.000	1.918
20	robot, ranger, helicopter, zombie, flash, skull, jupiter, lightning	angel, mist, event, catastrophe, spell, rumor, disaster, disastrous	0.000	1.893
21	cowboy, soldier, kid, fortune, band, ghost, owner, shelter	girl, singer, week, revolution, weeks, months, glimpse, month	0.000	1.716
22	commander, general, lair, viking, warlord, ferocious, master, savage	witch, fae, faerie, turmoil, sorceress, goddess, rebellion, wicked	0.000	1.897
23	camel, grandson, son, buck, boyhood, buffalo, fisherman, retired	ingalls, maid, women, heiress, gifted, baba, daughter, georgia	0.000	1.923
24	mighty, zeus, apollo, kong, thunder, poseidon, rushes, gates	emerald, cursed, rising, exiled, crowned, silver, sought, heaven	0.000	1.820
25	crashes, chases, builds, drives, flies, catches, saves, went	wears, throws, holds, reveals, appears, threatens, introduces, makes	0.000	1.878
27	outsmart, outwit, defeat, avoid, beat, defeated, avenge, manage	married, believed, marry, hoped, admit, struggled, thought, loved	0.000	1.876
30	conquers, bravery, overcomes, humility, testing, endurance, neverending, boredom	glamour, blossoming, bittersweet, delicate, embraces, social, fragile, cultural	0.000	1.918
31	mate, successor, compass, cheat, nerve, xray, achieves, abeke	bridesmaid, lainey, psychic, embraced, reexamine, kaylee, underestimate, qualified	0.000	1.832
32	huck, chuck, sawyer, heffley, fergus, sid, woody, mack	cara, unpopular, roommate, darrell, invalid, liza, insecure, exbest	0.000	1.910

33	spaceship, submarine, engine, fossil, skeleton, geronimo, sailors, tintin	haven, locket, guest, studio, broadway, moors, portland, invitation	0.000	1.921
34	potter, x, comic, terry, wimpy, dickens, el, w	magazine, shirley, romantic, designer, drama, blog, enchanting, romance	0.000	1.930
35	teammates, arm, leg, villagers, predators, bullies, injuries, injury	pregnancy, marriage, friendships, disasters, relationships, coping, issues, visions	0.000	1.903
38	stinky, fetch, whiskers, smelly, rotten, foul, wee, ate	elses, splendid, wearing, minute, ”, wrapped, means, am	0.000	1.919
40	army, agent, underground, enemy, allies, mission, weapon, force	society, elite, powers, agency, rebel, alliance, group, assassin	0.000	1.619
42	hell, peru, f, zero, astronaut, sidelines, miles, suspended	bffs, sixteen, graduation, mentry, redmond, internship, term, prep	0.000	1.889
43	bumps, escapes, climbs, drags, digs, wanders, creeps, journeys	weaves, welcomes, settling, thrust, whisked, steps, drawn, insight	0.000	1.890
44	aided, military, st, built, hired, accompanied, recruited, led	attended, boarding, established, written, divided, published, introduced, born	0.000	1.889
45	teggs, robotic, hawks, sonic, fu, armored, rockets, baloo	enticing, enduringly, imperfect, immerse, embrace, showcase, equestrian, define	0.000	1.923
46	tasmania, jurassic, faroff, dense, marooned, marsh, windswept, uninhabited	aurora, magnolia, starlight, stardust, lissa, carr, celestia, lennox	0.000	1.908
48	caesar, elenna, odysseus, griefer, ahab, creeper, shere, blackbeard	coven, clique, divine, maleficent, levana, blackthorn, moroi, rigid	0.000	1.929
50	fossils, trapping, appetite, heap, leaky, robbing, delivering, stash	hosting, thriving, designing, treasured, choosing, buying, leaving, discovering	0.000	1.892
51	rednosed, fe, merry, jesus, demonstrates, u, correct, relates	elegant, classical, festivals, supporting, weave, nutcracker, arranged, subject	0.000	1.884
52	repairs, clerk, bayport, spree, lifesaving, deliveries, principals, livestock	makeup, matchmaking, formal, makeover, bracelet, mardi, gala, gras	0.000	1.916
53	sir, drover, builder, swift, hammer, fireman, wily, jr	bff, diva, danvers, headstrong, leslie, donna, lu, shay	0.000	1.925

54	roblox, stevensons, seafaring, transcontinental, right-hand, homers, data, swashbuckling	feminist, austens, graceful, theatrical, regency, lgbtq, pitchperfect, glossy	0.000	1.892
57	hiccup, simba, woodman, swallow, zuko, cowardly, caspian, midas	tiana, victorias, helena, merida, celie, mulan, fiercely, keeah	0.000	1.893
58	columbus, robber, aang, augustus, crusoe, tashi, temporarily, narrates	crewe, esther, boyfriends, salem, anastasia, hetty, passionate, alicemiranda	0.000	1.878
59	trex, blaster, buddies, goofy, dino, lightyear, pluto, rod	tutu, accessories, sew, hairstyles, lace, tutus, handmade, jeans	0.000	1.907
62	wrestling, quick, tags, halfway, guinea, swimmer, skater, motions	ballerina, dancer, spirited, independent, dreamy, netball, compassionate, socially	0.000	1.748
63	goliath, robo, injun, ricotta, barncat, kojo, nephews, mcduck	prima, miri, flissa, montgomerys, blonde, vulliamy, bliss, wakefield	0.000	1.935

Appendix F WEFAT RESULTS SUB-RQ3

Professions that were excluded for the English Gutenberg embeddings because they were out of vocabulary:

- nutritionist, pharmacist, veterinarian, programmer, therapist, technician, appraiser, salesperson, planner, hygienist, paramedic, paralegal, receptionist, pathologist

The following table presents the results of the WEFAT test on the English Gutenberg embeddings. The first column shows the profession, the second column the percentage of female workers in real world data p_w, the third column shows the effect size (equation 5.6) and the fourth column the p-values. P-values in bold indicate statistically significant association with one of the genders ($p < 0.05$).

Profession	Percentage of Women	Effect size	p value
architect	0.000	-0.537	0.901
banker	0.000	-0.695	0.954
blacksmith	0.006	-0.861	0.980
miller	0.014	-0.437	0.840
druggist	0.019	-0.442	0.845
miner	0.026	-0.812	0.975
gardener	0.028	-0.604	0.924
clothier	0.035	-0.422	0.839
labourer	0.039	-0.409	0.840
dyer	0.044	-0.746	0.966
merchant	0.061	-0.678	0.944
baker	0.105	-0.466	0.858
tailor	0.137	-1.100	0.997
draper	0.143	-0.912	0.986
stationer	0.146	-0.527	0.896
grocer	0.209	-0.711	0.955
fishmonger	0.210	-0.934	0.985
collector	0.263	-0.561	0.903
maker	0.291	-0.283	0.736
butcher	0.297	-1.109	0.993
teacher	0.320	0.520	0.128
shoemaker	0.344	-0.319	0.793
confectioner	0.356	-0.132	0.602
farmer	0.462	-0.774	0.964
dealer	0.473	-0.598	0.920
shopkeeper	0.608	0.259	0.271
servant	0.754	-0.227	0.702
worker	0.905	0.059	0.450
queen	1.000	**1.278**	**0.002**
wife	1.000	0.165	0.344

Appendix G GERMAN WEAT TRANSLATIONS

In the Table below, the translated words used as German targets for the the WEAT can be found. Out of vocabulary words for the DNB embeddings were:

- Attribute words: weiblich, männlich
- W1: geschäftsführung, gehalt, professionell, management, heirat
- W2: geometrie, addition, gleichungen, algebra, berechnung, sinfonie, shakespeare
- W3: chemie, physik, sinfonie, shakespeare
- W4: scharfsinnig, anpassungsfähig, frühreif, erfinderisch, ehrwürdig, umsichtig, reflektiert, analytisch, wissbegierig, treffend, raffiniert, intuitiv, brillant, ansehnlich, schwächlich, errötend, muskulös, modisch, attraktiv, schlank, kahl, gedrungen, häuslich, mollig, verführerisch, schmächtig, hässlich
- W5: dynamisch, kühn, triumphieren, befehl, kraftvoll, anführen, dominant, kapitulieren, verletzlich, zurückziehen, wischiwaschi, versagen, nachgeben, zerbrechlich, verlierer
- W6: drinnen, schlafzimmer
- W7: soldat, lkw, pistole, schläger, schminke, ballerina, barbie
- W8: football, fahrradfahren, baseball, basketball, rugby, boxen, volleyball, netzball, turnen, lacrosse, cheerleader, softball
- W9: skizziert, hört zu
- W10: eigenwillig, störrisch, zynisch, naiv, rebellisch, charismatisch, qualifiziert, einflussreich, professionell, geachtet
- W12: instrumental, kompetent, inkompetent, pflegend, passiv, abhängig, gehorsam
- W13: instrumental, kontrollierend, autorität, herrschend, unterdrückend, befehlend, dominant, rücksichtsvoll, willig, kooperativ, gehorsam, wohlerzogen
- W14: ingenieurswesen, chemie, physik, informatik, geisteswissenschaften, biologie

For the German Gutenberg embeddings, following words were out of vocabulary:

- Attribute words: weiblich
- W1: geschäftsführung, büro, professionell, management, cousins
- W2: geometrie, mathe, gleichungen, algebra, sinfonie, shakespeare
- W3: technik, chemie, einstein, nasa, astronomie, physik, sinfonie, shakespeare

- W4: scharfsinnig, anpassungsfähig, frühreif, erfinderisch, intelligent, umsichtig, reflektiert, analytisch, einfühlsam, wissbegierig, clever, raffiniert, intuitiv, brillant, einfallsreich, ansehnlich, sinnlich, muskulös, modisch, attraktiv, mollig, sportlich, verführerisch, hässlich
- W5: durchsetzen, dynamisch, triumphieren, selbstbewusst, dominant, kapitulieren, verletzlich, wischiwaschi, zerbrechlich, verlierer
- W6: hinterhof, badezimmer
- W7: auto, fahrrad, lkw, schläger, pony, schminke, ballerina, puppenhaus, barbie
- W8: fußball, football, fahrradfahren, baseball, basketball, rugby, boxen, volleyball, netzball, lacrosse, cheerleader, schlittschuhlaufen, softball
- W9: skizziert, hört zu
- W10: eigenwillig, störrisch, zynisch, rebellisch, charismatisch, qualifiziert, einflussreich, professionell, erfolgreich
- W11: klempner, klempnerin, mechaniker, mechanikerin, schreinerin, maschinist, maschinistin, ingenieurin, programmierer, programmiererin, architekt, architektin, offizierin, hygieniker, hygienikerin, friseur, friseurin, krankenpfleger, bibliothekar, bibliothekarin, planer, planerin, therapeut, therapeutin, praktiker, praktikerin,
- W12: instrumental, kompetent, motiviert, erfolgreich, inkompetent, pflegend, passiv
- W13: instrumental, kontrollierend, herrschend, unterdrückend, dominant, kooperativ, wohlerzogen
- W14: ingenieurswesen, sport, technik, chemie, physik, informatik, geisteswissenschaften, biologie

	Attribute words	M	männlich, Mann, Junge, Bruder, er, ihm, sein, Sohn, Vater, Onkel, Großvater
		F	weiblich, Frau, Mädchen, Schwester, sie, ihre, ihr, Tochter, Mutter, Tante, Großmutter
Target words	W1: career vs. family	X	Geschäftsführung, Management, professionell, Unternehmen, Gehalt, Büro, Geschäft, Karriere
		Y	Zuhause, Eltern, Kinder, Familie, Cousins, Heirat, Hochzeit, Verwandte
	W2: math vs. arts	X	Mathe, Algebra, Geometrie, Rechnen, Gleichungen, Berechnung, Zahlen, Addition
		Y	Poesie, Kunst, Shakespeare, Tanz, Literatur, Roman, Sinfonie, Drama
	W3: science vs. arts	X	Wissenschaft, Technik, Physik, Chemie, Einstein, NASA, Experiment, Astronomie
		Y	Poesie, Kunst, Shakespeare, Tanz, Literatur, Roman, Sinfonie, Drama
	W4: intelligent vs. appearance	X	frühreif, einfallsreich, wissbegierig, genial, erfinderisch, scharfsinnig, anpassungsfähig, reflektiert, einfühlsam, intuitiv, neugierig, umsichtig, analytisch, treffend, ehrwürdig, einfallsreich, scharfsinnig, aufmerksam, weise, klug, raffiniert, clever, brillant, logisch, intelligent
		Y	verführerisch, üppig, errötend, häuslich, mollig, sinnlich, hinreißend, schlank, kahl, sportlich, modisch, gedrungen, hässlich, muskulös, schmächtig, schwächlich, ansehnlich, gesund, attraktiv, fett, schwach, dünn, hübsch, schön, stark
	W5: strength vs. weakness	X	Macht, stark, selbstbewusst, dominant, kraftvoll, Befehl, durchsetzen, laut, kühn, gelingen, triumphieren, anführen, schreien, dynamisch, gewinnen
		Y	schwach, kapitulieren, ängstlich, verletzlich, Schwäche, wischiwaschi, zurückziehen, nachgeben, versagen, schüchtern, folgen, verlieren, zerbrechlich, ängstlich, Verlierer
	W6: outdoor vs. indoor	X	draußen, außen, Natur, Garten, Baum, Hinterhof, See, Berg
		Y	innen, drinnen, Küche, Haushalt, Zuhause, Sofa, schlafzimmer, Badezimmer
	W7: male vs. female toys	X	Ball, Schläger, Lkw, Auto, Fahrrad, Pistole, Soldat, blau
		Y	Puppe, Puppenhaus, barbie, Schminke, Ballerina, schmuck, Pony, rosa

	Attribute words	M	männlich, Mann, Junge, Bruder, er, ihm, sein, Sohn, Vater, Onkel, Großvater
		F	weiblich, Frau, Mädchen, Schwester, sie, ihre, ihr, Tochter, Mutter, Tante, Großmutter
Target words	W8: male vs. female sports	X	Football, Basketball, Baseball, Fußball, Ringen, Rugby, Boxen, Fahrradfahren
		Y	Volleyball, Turnen, Netzball, Softball, Cheerleader, Tanzen, Schlittschuhlaufen, Lacrosse
	W9: active vs. silent games	X	fliegt, fährt, springt, klettert, schwimmt, rutscht, taucht, hüpft
		Y	liest, schreibt, beobachtet, versteckt, hört zu, zeichnet, malt, skizziert
	W10: male vs. female adjectives	X	rebellisch, begabt, störrisch, eigenwillig, charismatisch, zynisch, mürrisch, naiv
		Y	qualifiziert, verehrt, geachtet, einflussreich, professionell, bekannt, begabt, erfolgreich
	W11: male vs. female professions	X	Klempner, Klempnerin, Mechaniker, Mechanikerin, Schreiner, Schreinerin, Maschinist, Maschinistin, Ingenieur, Ingenieurin, Programmierer, Programmiererin, Architekt, Architektin, Offizier, Offizierin
		Y	Hygieniker, Hygienikerin, Friseur, Friseurin, Krankenschwester, Krankenpfleger, Bibliothekar, Bibliothekarin, Planer, Planerin, Therapeut, Therapeutin, Praktiker, Praktikerin, Lehrer, Lehrerin
	W12: male vs. female qualities [23]	X	kompetent, instrumental, erfolgreich, motiviert, klug, Abenteuer, verdienen, meistern
		Y	pflegend, abhängig, gehorsam, inkompetent, passiv, Opfer, erfolglos, unsichtbar
	W13: dominant vs. obedient	X	dominant, herrschend, unterdrückend, kontrollierend, befehlend, überlegen, Autorität, instrumental
		Y	gehorsam, willig, aufmerksam, rücksichtsvoll, wohlerzogen, höflich, gezwungen, kooperativ
	W14: male vs. female subjects	X	Mathematik, Physik, Wissenschaft, Chemie, Informatik, Ingenieurswesen, Sport, Technik
		Y	Geisteswissenschaften, Kunst, Bildung, Biologie, Medizin, Sprache, Englisch, Musik