Gender Bias in Text: Origin, Taxonomy, and Implications

Jad Doughman
American University of Beirut
Beirut, Lebanon
jad17@mail.aub.edu

Wael Khreich
American University of Beirut
Beirut, Lebanon
wk47@aub.edu.lb

Maya El Gharib
American University of Beirut, Lebanon
mme116@mail.aub.edu

Maha Wiss
American University of Beirut, Lebanon
maw16@mail.aub.edu

Zahraa Berjawi
American University of Beirut, Lebanon
zjb04@mail.aub.edu

Abstract

Gender inequality represents a considerable loss of human potential and perpetuates a culture of violence, higher gender wage gaps, and a lack of representation of women in higher and leadership positions. Applications powered by Artificial Intelligence (AI) are increasingly being used in the real world to provide critical decisions about who is going to be hired, granted a loan, admitted to college, etc. However, the main pillars of AI, Natural Language Processing (NLP) and Machine Learning (ML) have been shown to reflect and even amplify gender biases and stereotypes, which are mainly inherited from historical training data. In an effort to facilitate the identification and mitigation of gender bias in English text, we develop a comprehensive taxonomy that relies on the following gender bias types: *Generic Pronouns*, *Sexism*, *Occupational Bias*, *Exclusionary Bias*, and *Semantics*. We also provide a bottom-up overview of gender bias, from its societal origin to its spillover onto language. Finally, we link the societal implications of gender bias to their corresponding type(s) in the proposed taxonomy. The underlying motivation of our work is to help enable the technical community to identify and mitigate relevant biases from training corpora for improved fairness in NLP systems.

1 Introduction

Bias is prevalent in every aspect of our lives. We are hardwired to compartmentalize things we experience to form a plausible perception of the world around us. The process of forming these perceptions typically breeds prejudices, which allows for flagrant inequalities to shape across different demographics. The prevalence of certain biases in society, such as gender bias, can be attributed to social roles formed as a function of this compartmentalization process. According to the social role theory, the societal origin of gender stereotypes revolves around gender-typical social roles that mirror the sexual division of labor and gender hierarchy of the society (Bussey and Bandura, 1999).

The prevalence of gender bias in society is also spilled over onto language through the patriarchal worldview predominant among linguists prior to the prescriptive grammar movement in English. Bodine (1975) found that the generic use of *he* is derived from an androcentric worldview prevalent among 18th-century grammarians: "human beings were to be considered male unless proven otherwise" (Bodine, 1975). The perpetuation of bias onto language entails a negative feedback loop due to the direct impact of language on a person's perceptions (Boroditsky, 2011). Linguistic determinism, a hypothesis taken from the analytic branch of philosophy, posits that language "limits and determines human thought patterns and knowledge" (Hickmann, 2000). Hence, the recurring usage of bias in language consequently leads to a more biased perception which is fed back into our lexical (word) choice. This is even more amplified by the increased adoption of automated system based on AI, which exponentially expedites this feedback loop (as detailed in Section 2.4).

The linguistic spillover of gender bias has various direct and indirect implications on our society. The presence of gender bias in the language used by parents and in school text books causes children to develop sexist perceptions and behaviors towards other children of opposite gender and deepens the problematic outcomes of gender inequalities in society (Waxman, 2013). Additionally, sex-biased wording affects a person's perception of a career's

Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pages 34–44

attractiveness (Briere and Lanktree, 1983). Consequently, countries that adopt a gendered language tend to have disproportionate labor force participation (Gay et al., 2013). We also discuss the direct implication of hostile sexism on a person's physiological wellbeing, such as increased stress levels, anger, and elevated cardiovascular reactivity (Schneider et al., 2001). Finally, we examine the indirect implication of benevolent sexism in embedding gender inequality and intensifying its influence in the society by portraying the advantageous aspects of being a woman (deserving special treatment, care, protection, and love) (Hammond et al., 2014; Barreto and Ellemers, 2005).

Gender bias in NLP presents itself in many stages along the design and development process. It can be found in the training data, the pre-trained models, and the algorithms themselves. The propagation of bias from text to features and algorithms leads to real-world consequences when integrated into AI systems and are used in critical decision-making applications. In particular, discriminatory decisions occur when these systems assist humans in critical decisions (Dressel and Farid, 2018). These prejudiced decisions could entail allocational or representational harms (Blodgett et al., 2020b). As mentioned previously, discriminating algorithms accelerate the unavoidable feedback loop, which increases the degree and volume of bias against females and other gender minority groups, especially in online media content. Automated NLP-based decision-making algorithms will re-consume this increasingly growing biased content to update their models, and so on. This feedback loop contributes to an increased gender bias and further discrimination.

Several works in NLP revolving around bias focused on the projection of word embedding vectors on a gender direction (he - she) to detect and mitigate bias in a pre-trained model, without a clear link to the implications on society and their underlying applications (Blodgett et al., 2020b). There has been previous attempts that address bias at the sentence level and provide an initial categorization of gender bias types (Hitti et al., 2019). We build on their work and provide a more comprehensive understanding of the various forms of gender bias while linking to several real world implications on society.

In this paper, we develop a comprehensive taxonomy to identify various types of gender bias. We also provide a bottom-up overview of gender bias, from its societal origin to its spillover onto language. We then link between the psycho-social implications of gender bias and the corresponding type(s) in the proposed taxonomy. Our underlying motivation is to enable the the technical community working on gender bias in NLP to focus on the identification and mitigation of relevant biases for improved fairness in NLP systems. We also hope that by addressing and linking the sources and implications of gender bias in text, we encourage the community to further push the research in this direction and raise more awareness on bias and discrimination in NLP systems.

2 Gender Bias

2.1 Definition

We define gender bias in text as being an exclusionary, implicitly prejudicial, or generalized representation of a specific gender as a function of various societal stereotypes. The sections below provide a bottom-up overview of gender bias, from its societal origin to its spillover onto language while highlighting its perceptual and societal implications.

2.2 Social Role Theory

The social role theory posits that gender stereotypes are rooted in the distinct social roles designated to women and men (Bussey and Bandura, 1999). Historically, men and women have maintained diverse social roles: Men have been more likely to engage in tasks that require "speed, strength, and the possibility of being away from home for long periods of time", while women have been more likely to "stay home and engage in family tasks, such as child-rearing" (Eagly et al., 2000). This dispersion comes with various consequences. Firstly, men are perceived as, and expected to be agentic, particularly, active, independent, and resolute, whereas women are perceived as, and expected to be, communal, namely, kind, helpful, and benevolent (Eagly et al., 2000). Secondly, women and men become more inclined to acquire particular skills linked to successful role performance and by adapting their social behavior to role requirements (Eagly et al., 2000). Essentially, both actors and observers are inclined to inherit traits from observed behaviors in their specific social roles (Steffens et al., 2015). This creates an unavoidable negative feedback loop that continuously perpetuates

gender bias in society by segregating each gender into a specific social role and actively promotes the divergence through the pursuit of successful role performance.

2.3 Linguistic Spillover

Gender stereotypes in society also found their way into language, tunneling through a patriarchal worldview adopted by grammarians prior to the prescriptive grammar movement. Bodine (1975) found that the generic use of *he* is derived from an androcentric worldview prevalent among 18th-century grammarians: "human beings were to be considered male unless proven otherwise" (Bodine, 1975). This is also supported by the limited role of women in forming and shaping the English language (Kramarae, 1981). Feminist scholars maintain that the *generic he* and similar words "not only reflect a history of male domination" but also "actively encourage its perpetuation" (Sniezek and Jazwinski, 1986). The *generic he* has also intensified sexist behaviors and attitudes in a subtler psychological and perceptual manner. The foundation of this argument is in the Sapir-Whorf hypothesis: "our grammar shapes our thought" (Whorf, 1956). Blaubergs (1980) applies this hypothesis to sexist words and phrases in the English language, including the *generic he*. She maintains that regardless of its origins,"Sexist language by its existence reinforces and socializes sexist thinking and practices" (Blaubergs, 1980). Consequently, the recurring usage of biased language leads to a more biased perception which is fed back into our lexical (word) choice.

2.4 Bias in NLP

Natural Language Processing (NLP) and Machine Learning (ML) techniques, the main pillars of narrow or practical AI, are designed to learn from data and try to generalize the learned concepts to unseen data. However, they are prone to inherit, reflect, and amplify biases and stereotyped-associations that are present in historical data provided for training. Manifestations of different kinds of biases have been shown to exist in various components used to develop NLP and ML systems, from training data to pre-trained models to algorithms and resources (Olteanu et al., 2016; Tolan, 2018; Danks and London, 2017; Mehrabi et al., 2019; Sun et al., 2019; Blodgett et al., 2020b; Hovy et al., 2020; Hitti et al., 2019).

Word embeddings is a family of techniques that learn word representation from texts, such that words with similar meaning have a similar representation (Mikolov et al., 2013b,a). Since their inception, word embeddings have become the predominant representation of text features and an integral part of NLP applications. However, most research on gender bias in NLP has focused on the projection of word embedding vectors on a gender direction (he - she) to detect and mitigate bias in a pre-trained model. For example, occupational gender bias in word embedding models is typically measured by comparing the distances between gendered word vectors and occupational terms. The bias scores resulting from the manipulation of word vectors in a pre-trained word embedding are strictly dependent on the corpus utilized to train that model. Using such models to detect whether new sentences are biased will not only project the biases of the model but also misconstrue its origin (Blodgett et al., 2020b).

The key existing solutions to mitigate these biases focused on modifying the training data, imposing constraints on the word embeddings objective function, or applying post-processing techniques to reduce the bias in word embedding models including word2vec (Mikolov et al., 2013b,a), and GloVe (Bojanowski et al., 2017), and more recently in contextual word embedding models such as ELMO (Hoffman et al., 2010), BERT (Devlin et al., 2018), and ALBERT (Lan et al., 2019). Although several other papers discussed different methodologies to debias word embedding model, these techniques have been scrutinized on several occasions (Blodgett et al., 2020a). In addition, the majority of research did not focus on the impact of gender bias in real-word applications (Blodgett et al., 2020a). Automatic detection of gender bias beyond the word level requires an understanding of the semantics of written human language, which remains an open problem and successful approaches are restricted to specific domains and tasks. In an effort to redirect the focus to the linguistic forms of bias and their societal implications, Section 3 contains a comprehensive breakdown of the various gender bias types and their subsequent subtypes, while the next section will be geared towards their societal implications.

2.5 Implications

Gender bias leaks into some of the fundamental life aspects and tends to jeopardize the normal func-

tioning of the affected gender group (Fraser, 2000). The sections below describe the negative implications of gender bias on children's mental imagery, career attractiveness, labor force participation, and human behavior.

2.5.1 Children's Mental Imagery

Gender bias can manifest itself at an early age in one's life and thus can have a more profound impact on one's attitude and behavior. Children and even infants can be exposed to gender bias presented in language and can also be affected by it, through the process of categorization (Waxman, 2013). The process of category learning begins early on in a person's life and is perceived as a building block for children's lexical acquisition (Waxman, 2013). However, this process could promote stereotypical beliefs and gender biases in children's cognition and perception about individuals, especially if the language used in this process is a gendered language (Bigler and Leaper, 2015). A gendered language which makes gender salient, tends to treat gender as a major attribute upon which children will rely on, to classify and make inferences about others (Hilliard and Liben, 2010). Therefore, the learnt categorizations will promote and perpetuate several forms of gender bias, such as in-group favoritism (Arthur et al., 2008; Bigler and Liben, 2006; Leaper and Bigler, 2004). In-group favoritism can be reflected in children's behavior where a child would prefer to play with another child of the same gender rather than a child with an opposite gender (Fagot et al., 1986).

Gender-generic noun statements, such as "Girls are good at activity X while boys are good at activity Y" that are usually stated by parents and found in school textbooks, influence how children think about themselves. These statements also undermine children's achievements in the relevant activities given their belonging to one of the gender categories (Bigler and Leaper, 2015; Cimpian et al., 2012). In their study, Cimpian et al. (2012) discovered that when children are exposed to gender-generic statements that link their ability to perform a certain activity to a social group, they tend to perform worse on the given activity irrespective of whether the statement is positive or negative. Cimpian et al. (2012) study implies how threatening gendered generic statements can be in relation to the beliefs that children instantly create about their own capabilities and achievements.

2.5.2 Career Attractiveness

In a study to assess the contribution of biased language relating to the attractiveness of a career, Briere and Lanktree (1983) established that biased language significantly affects a subjects' perception of the attractiveness or employment in a psychology career for women (Briere and Lanktree, 1983). Generic pronouns (as detailed in Section 3.1) and masculine nouns were linked with a decline in the presumed attractiveness of a psychology career for women, with respect to a nonsexist condition (Briere and Lanktree, 1983). Consequently, the use of generic pronouns in texts could discriminatively inhibit female interest in fields they might alternatively seek out (Briere and Lanktree, 1983).

Additionally, a study conducted by Stout and Dasgupta (2011) reveals that gender-biased language in the professional field is associated with negative nonverbal emotional responses from women. Accordingly, women who are exposed to a gender exclusive language during a job interview tend to feel demotivated and socially and actively rejected by the workplace (Stout and Dasgupta, 2011). Other evidence by Vervecken et al. (2013) proposes how children's perceptions of stereotypically male jobs can be influenced by the linguistic form used to present an occupational title. For example, the generic use of masculine plural forms when describing occupations will most likely lead children to restrictive, male only associations and perceptions about stereotypically male occupations (Vervecken et al., 2013).

2.5.3 Labor Force Participation

The gender gaps between women and men in the labor market are almost present in every country, yet with varying degrees, given the cultural norms and values that play a crucial role in introducing or generating new stereotypical beliefs and resisting the existing ones as time passes and cultures change. Aside from the cultural system represented by the social norms and values, a country's adopted language system and the intensity to which it marks gender differences tends to be a very crucial variable in determining the extent to which women can participate in the socio-economic life (Gay et al., 2013). The idea that a country's language system affects women's socioeconomic participation sets off from the idea that language is a key vehicle of the cultural system (North et al., 1990). In their study, Gay et al. (2013) discovered that gendered language has a direct impact on women's socio-

economic choices and outcomes. For example, female labor force participation for the year 2000 in countries following a gender binary linguistic system, such as France and Spain, was 16% lower as compared to countries which have no gender marking or have more than three genders in its most spoken language (Gay et al., 2013).

2.5.4 Human Behavior

As stated in the social role theory suggested by Eagly et al. (2000), the gendered roles construct the societal belief system that sets the expectations of men and women, and biased language is instrumentalized to maintain the genders' distinct responsibilities (Stahlberg et al., 2007). As a result, these stereotypical beliefs would be reflected in the everyday lexical choices that refer to men or women, including prejudice or stereotypes that are based on gender or, in other words, sexism (Menegatti et al., 2017). As detailed in Section 3.2, Glick and Fiske (1996) divided sexism into hostile sexism, the typical prejudice against women, and benevolent sexism, the seemingly 'positive' sexism that enforces masculine dominance in the society through viewing women as caring, delicate, emotional, and in need of men's protection (Glick and Fiske, 1996).

Bosson et al. (2010) state that women suffer from the emotional impact of hostile sexism for a shorter period of time due to the direct anger expression that's linked to it. Moreover, the exposure to a hostile sexist language motivates women to participate in collective action to stop gender inequality, and it encourages them to socially compete with men in order to reclaim their righteous social status (Becker and Wright, 2011). Nevertheless, hostile sexist language may not have a direct impact on embedding further gendered stereotypes in society, but it has severe direct impact on women's physiological wellbeing, such as increased stress levels, anger, and elevated cardiovascular reactivity (Schneider et al., 2001).

On the other hand, there has been a research consensus on the impact of women's exposure to benevolent sexist language on embedding gender inequality and intensifying its influence in the society (Hammond et al., 2014; Barreto and Ellemers, 2005). For instance, Hammond et al. (2014) indicate that the positive attributes that benevolent sexism holds for women may impair women's opposition to the gendered stereotypes due to how this form of sexism portrays the advantageous aspects of being a woman (deserving special treatment, care, protection, and love). Another study shows that benevolent sexist language is often not identified as sexism for many people exposed to it (Barreto and Ellemers, 2005). Thus, this may keep this issue unrecognized and further maintain the acceptance of prejudicial gendered stereotypes, allowing for continuous promoting of sexism and their direct or indirect impact on women (Barreto and Ellemers, 2005).

3 Taxonomy

The first step of detecting biased language is to categorize the various forms of that bias while carefully maintaining a clear segregation between the resultant groups. The below sections develop a comprehensive taxonomy that includes a wide range of gender bias types and their subsequent subtypes. Each subsection includes the definition of a bias subtype and a couple of examples that illustrate its usage in a sentence. Table 2 provides an overview of the taxonomy, with one example pertaining to each subtype alongside its societal implication (discussed in Section 2.5).

3.1 Generic Pronouns

Given that the choice of a pronoun follows the sex of the referent, a problem arises when a pronoun is to be used with sex-indefinite antecedents (Ozieblowska, 1994). Pronouns which do not specify sex are traditionally called "generic", because generic statements about human referents discuss people in general, and therefore the sex of the referents is irrelevant (Ozieblowska, 1994). The most notable forms of generic pronouns are: *generic he*, *generic she*, and *gendered generic man*.

3.1.1 Generic He

The use of the pronoun *he* in circumstances of sex-indefinite reference overly emphasizes men over women, thereby both "re-constituting and signifying males' micro-political hegemony" (Stringer and Hopper, 1998). Thus, *generic he* occurs when the pronoun he, his and him are used as referents to nouns of no specific gender. Among the gendered generic pronouns, *his* is the most recurring sexist antecedent to most nouns. Below are some example:

- The client should receive **his** invoice in two weeks.
- A good employee knows that **he** should strive for excellence.

- A teacher is expected to be a good role model in all areas of **his** life.

3.1.2 Generic She

While the generic *he* is the most recurring form of generic pronouns, generic *she* is also excessively present in written discourse. Below are some example:

- A nurse should ensure that **she** gets adequate rest.
- A dancer should watch **her** diet carefully.
- **She** presents us diverge ways, but **she** lets us choose our path.

3.1.3 Gendered Generic Man

Gendered generic man appears when *man* is utilized as a masculine noun representation both genders. It's used not only as a noun but also as a verb. Below are some example:

- Good teachers know how to **man** the classroom.
- Effective teachers lead or **man** the students well.
- It is even more fulfilling when a teacher sees a once stubborn child who became a **man** of success and responsibilities crown with various achievements.
- All **men** are born for a reason.
- A teacher is an ordinary **man** with extraordinary roles.

3.2 Sexism

According to the ambivalent sexism theory, sexism against women is divided into an aggressive expression, or hostile sexism, and a positive (for men) expression, or benevolent sexism (Glick and Fiske, 1996). In this section, we will be discussing these two divergent forms of sexist language:

3.2.1 Hostile Sexism

Hostile sexism is the view of men as more powerful and competent than women (Becker and Wright, 2011). It views women as a threat to men's dominance through their violation to traditional gendered roles in the society (Becker and Wright, 2011; Mastari et al., 2019). In general, hostile sexism reflects men's hatred towards women (or misogyny), and it is expressed in aggressive and blatant manner (Connor et al., 2017). Men with hostile sexist mentality view women as manipulative, unintelligent, and incompetent (Jain et al., 2019). Below are some examples of hostile sexist statements:

- The people at work are childish. It's run by women and when women don't agree to something, oh man.
- Women always get more upset than men.
- Women are incompetent at work.

3.2.2 Benevolent Sexism

Benevolent sexism is a softer form of sexism that expresses male dominance in a more chivalrous tone (Becker and Wright, 2011). It expresses affection and care for women in return for their acceptance to their limited gendered roles (Becker and Wright, 2011; Mastari et al., 2019). Benevolent sexism describes women as caring, innocent, and in need of men's protection, and these stereotypical notions are used to reinforce women's subordinate position (Connor et al., 2017). This form of sexism explains how women complete men's chivalry, power, and intelligence with their delicate characteristics (Cross and Overall, 2018). Below are some examples of benevolent sexist statements:

- They're probably surprised at how smart you are, for a girl.
- No man succeeds without a good woman besides him. Wife or mother. If it is both, he is twice as blessed.
- I am not exploiting women: I love, protect, and care for them.

3.3 Occupational Bias

As discussed in Section 2.2, the societal origin of gender stereotypes revolves around gender-typical social roles and thus reflect the sexual division of labor and gender hierarchy of the society (Eagly et al., 2000). The resultant social roles lead to gendered occupational bias, which is a form of generalization that occurs when an occupation or role/duty is generalized onto a specific gender. This section will illustrate both the gendered division of labor and gendered roles/duties.

3.3.1 Gendered Division of Labor

Below are some examples that illustrate how certain jobs are seen as only appropriately and exclusively held by either women or men:

- **Professors** are men and elementary teachers are women.
- **Politicians** are men and women are wives.
- **Housework** is the duty of women and an option or out of question for men.
- **Scientists** are men and secretaries are women.
- **Doctors** are men and nurses are women.

3.3.2 Gendered Roles/Duties

In the first example below, the speaker's sales assistant is referred to as a girl, which diminishes the status of the role. In the second, the sales assistant is referred to by job title, which indicates that gender is not an important prerequisite for the role that the sales assistant plays.

1. I'll have my girl get you a cup of coffee.
2. I'll ask my assistant to get you a cup of coffee.

3.4 Exclusionary Bias

3.4.1 Explicit Marking of Sex

Explicit marking of sex occurs when an unknown gender-neutral entity is referred to using gender-exclusive term(s). Table 1 provides proposed corrections of some exclusionary terms.

Example	Proposed Corrections
Mankind	Humanity; human beings
Chairman	Chairperson; chair
Businessman	Business manager
Manpower	Workforce
Cameraman	Camera operator
Policeman	Police officer
Manhood	Adulthood
Brotherhood	Solidarity

Table 1: Proposed solutions to some exclusionary terms

3.4.2 Gender-based Neologisms

Neologisms are newly coined words/expressions that may be in the process of mainstream adoption, but have not yet been fully accepted. Gender-based neologisms are gendered coinages that could have underlying stereotypical tendencies (Foubert and Lemmens, 2018). Below are some examples:

- **Man-bread:** bread that is baked so big that it will take a grown man a whole week to eat it, having 4 slices a day.
- **Man-sip:** a man sized sip of a beer or drink, one can finish a beer in 4 or 5 Man-sips. For a female or light weight, it borders on chugging the drink, but for a man it is merely a sip.
- **Mantini:** a martini or alcoholic beverage that appeals to a man's palate. "My boyfriend prefers his mantini straight up which is just too strong for my tastes."

3.4.3 Gendered Word Ordering

Gendered word ordering is tendency for the male version to come first in binomials such as "men and women", "brothers and sisters", "boys and girls", or "Mr and Mrs". Many words that incorporate the word "man", such as "man-made", "mankind", "manpower", have perfectly acceptable gender-neutral alternatives: for example, "artificial" or "synthetic", "humankind", and "workforce".

3.5 Semantics

Gender bias in semantics appears when utilizing words and sentences that are demeaning in their semantic meaning (Umera-Okeke, 2012). The implicit meaning behind sexist jokes, proverbs, or even using specific non-human terms to refer to women, consciously or unconsciously, deepens the existing bias and projects it onto new generations (Umera-Okeke, 2012). The current study suggests three types of semantic gender bias: metaphors, gendered attributes, and old sayings.

3.5.1 Metaphors

People tend to express a part of the world's reality through metaphors, which contributes to ingraining their culture and beliefs. By looking into the window of metaphors, several biases of society are revealed (Rodriguez, 2009). Masculinity and bias against females are represented in metaphoric words that describe women as a non-human comparing females to food, animals, plants (Martín, 2011; Lan and Jingxia, 2019). Below are some examples of English metaphoric words that describe woman as food and animal:

- "Cookie": lovely woman
- "Old Hen": middle aged women who love to talk to each other

Type	Subtype	Example	Implication
Generic Pronouns	Generic He	The client should receive his invoice in two weeks.	Biased Mental Imagery
	Generic She	A nurse should ensure that she gets adequate rest.	Biased Mental Imagery
	Gendered Generic Man	Good teachers know how to man the classroom.	Biased Mental Imagery
Sexism	Hostile Sexism	Women are incompetent at work.	Aggressive Behavior
	Benevolent Sexism	They're probably surprised at how smart you are, for a girl.	Representational Harms
Occupational Bias	Gendered Division of Labor	Professors are men and elementary teachers are women.	Labor Force Participation
	Gendered Roles & Duties	I'll have my girl get you a cup of coffee.	Labor Force Participation
Exclusionary Bias	Explicit Marking of Sex	Chairman, Businessman, Manpower, Cameraman...	Representational Harms
	Gender-based Neologisms	Man-bread, Man-sip...	Representational Harms
	Gendered Word Ordering	"Men and Women", "Brothers and Sisters"...	Representational Harms
Semantics	Metaphors	"Cookie": lovely woman.	Bias Propagation
	Gendered Attributes	An unmarried male (bachelor) is a "personal choice". An unmarried female (spinster) is derogatorily an "old maid".	Bias Propagation
	Old Sayings	A woman's tongue three inches long can kill a man six feet high.	Bias Propagation

Table 2: Overview of the taxonomy and link to societal implications

3.5.2 Gendered Attributes

Societal ideologies revolving around each gender role, preferences, interests, and characteristics were originally created due to many historical conditions and various lifestyles, and are conveyed to language in which reflects sexist stereotypes, which might presents invisible limitations for women. Lan and Jingxia (2019) suggest that placing men in a leading position and women as subordinates is the main cause of creating gendered stereotypes (Lan and Jingxia, 2019). Researchers noted that commendatory or complementary terms are used as male words while the corresponding female words are derogatory (e.g. wizard/ witch, spinster / bachelor , governor / governess) (Lan and Jingxia, 2019). Associating positive meaning with male and negative meaning with female represents semantic derogation and disparagement. Here are sentences show the derogatory meaning of some female words:

- An unmarried male (bachelor) is a "personal choice". An unmarried female (spinster) is derogatorily an "old maid".
- A "strict male manager" is described as a responsibility taker. A "strict female manager" is described as hard to work with.

3.5.3 Old Sayings

Biased old sayings come in various forms including: proverbs, set-phrases, and formulaic expressions that present a source of stereotype against women. Those sayings are culturally seen as axioms and absolute truth, which affect people behavior to adapt them as moral standards (Martín, 2011). Below are sentences exemplifying implicit sexism in proverbs:

- A woman's tongue three inches long can kill a man six feet high
- Bad words make a woman worse
- When you see an old man, sit down and take a lesson; when you see an old woman, throw a stone

4 Conclusion

In this paper, we propose a comprehensive gender bias taxonomy that distinguishes between the various forms of gender biases in English text. The taxonomy includes various exclusionary, implicitly prejudicial, and generalized forms of biased gender representations in text. Our work also provides a bottom-up understanding of gender bias, highlighting the social role theory and its impact on gender stereotypes in society. We also explain how societal gender bias spilled over onto language while being fed back into our perceptions as stated in the Sapir-Whorf hypothesis.

We hope that our comprehensive taxonomy of gender bias enables the technical community working on gender bias in NLP to focus on the identification and mitigation of relevant biases in text for improved fairness in NLP systems. We also hope that by addressing and connecting the sources and implications of gender bias in text from a linguistic, sociological, and real-life perspective, we would encourage the community to further push the research in this direction and raise more awareness on bias and discrimination in NLP systems. In future work, we will work on expanding the taxonomy to include other languages and address other forms of bias such as racial and ethnic biases.

References

Andrea E Arthur, Rebecca S Bigler, Lynn S Liben, Susan A Gelman, and Diane N Ruble. 2008. Gender stereotyping and prejudice in young children: A developmental intergroup perspective.

Manuela Barreto and Naomi Ellemers. 2005. The burden of benevolent sexism: How it contributes to the maintenance of gender inequalities. *European journal of social psychology*, 35(5):633–642.

Julia C Becker and Stephen C Wright. 2011. Yet another dark side of chivalry: Benevolent sexism undermines and hostile sexism motivates collective action for social change. *Journal of personality and social psychology*, 101(1):62.

Rebecca S Bigler and Campbell Leaper. 2015. Gendered language: Psychological principles, evolving practices, and inclusive policies. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):187–194.

Rebecca S Bigler and Lynn S Liben. 2006. A developmental intergroup theory of social stereotypes and prejudice. *Advances in child development and behavior*, 34:39–89.

Maija S Blaubergs. 1980. An analysis of classic arguments against changing sexist language. *Women's studies international quarterly*, 3(2-3):135–147.

Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020a. Language (Technology) is Power: A Critical Survey of "Bias" in NLP.

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

Ann Bodine. 1975. Androcentrism in prescriptive grammar: singular 'they', sex-indefinite 'he', and 'he or she'. *Language in Society*, 4:129 – 146.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lera Boroditsky. 2011. How language shapes thought. *Scientific American*, 304(2):62–65.

Jennifer K Bosson, Elizabeth C Pinel, and Joseph A Vandello. 2010. The emotional impact of ambivalent sexism: Forecasts versus real experiences. *Sex Roles*, 62(7):520–531.

John Briere and Cheryl Lanktree. 1983. Sex-role related effects of sex bias in language. *Sex roles*, 9(5):625–632.

Kay Bussey and Albert Bandura. 1999. Social cognitive theory of gender development and differentiation. *Psychological review*, 106(4):676.

Andrei Cimpian, Yan Mu, and Lucy C Erickson. 2012. Who is good at this game? linking an activity to a social category undermines children's achievement. *Psychological science*, 23(5):533–541.

Rachel A Connor, Peter Glick, and Susan T Fiske. 2017. Ambivalent sexism in the twenty-first century.

Emily J Cross and Nickola C Overall. 2018. Women's attraction to benevolent sexism: Needing relationship security predicts greater attraction to men who endorse benevolent sexism. *European Journal of Social Psychology*, 48(3):336–347.

David Danks and Alex John London. 2017. Algorithmic Bias in Autonomous Systems A Taxonomy of Algorithmic Bias. *26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, (Ijcai):4691–4697.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.0(Mlm):4171–4186.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Alice H Eagly, Wendy Wood, and Amanda B Diekman. 2000. *Social role theory of sex differences and similarities: A current appraisal*, pages 123–174. Erlbaum.

Beverly I Fagot, Mary D Leinbach, and Richard Hagan. 1986. Gender labeling and the adoption of sex-typed behaviors. *Developmental Psychology*, 22(4):440.

Océane Foubert and Maarten Lemmens. 2018. Gender-biased neologisms: the case of man-x. *Lexis. Journal in English Lexicology*, (12).

Nancy Fraser. 2000. Redistribution, recognition and participation: towards an integrated concept of justice. *World Culture Report*, pages 48–57.

Victor Gay, Estefania Santacreu-Vasut, and Amir Shoham. 2013. The grammatical origins of gender roles. *Berkeley Economic History Laboratory Working Paper*, 3.

Peter Glick and Susan T Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3):491.

Matthew D Hammond, Chris G Sibley, and Nickola C Overall. 2014. The allure of sexism: Psychological entitlement fosters women's endorsement of benevolent sexism over time. *Social Psychological and Personality Science*, 5(4):422–429.

Maya Hickmann. 2000. Linguistic relativity and linguistic determinism: Some new directions.

Lacey J Hilliard and Lynn S Liben. 2010. Differing levels of gender salience in preschool classrooms: Effects on children's gender attitudes and intergroup bias. *Child development*, 81(6):1787–1798.

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

Matthew D Hoffman, Francis R Bach, David M Blei, and Francis R Bach. 2010. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Sarthak Jain, Ramin Mohammadi, and Byron C. Wallace. 2019. An analysis of attention over clinical notes for predictive tasks. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 15–21, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Cheris. Kramarae. 1981. Women and men speaking : frameworks for analysis / cheris kramarae. pages xviii, 194 p. ;.

Tian Lan and Liu Jingxia. 2019. On the gender discrimination in english. *Advances in Language and Literary Studies*, 10(3):155–159.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Campbell Leaper and Rebecca S Bigler. 2004. Gendered language and sexist thought.

Carmen Fernández Martín. 2011. Comparing sexist expressions in english and spanish:(de)-constructing sexism though language. *ES: Revista de filología inglesa*, (32):67–90.

Laora Mastari, Bram Spruyt, and Jessy Siongers. 2019. Benevolent and hostile sexism in social spheres: The impact of parents, school and romance on belgian adolescents' sexist attitudes. *Frontiers in Sociology*, 4:47.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning.

Michela Menegatti, Elisabetta Crocetti, and Monica Rubini. 2017. Do gender and ethnicity make the difference? linguistic evaluation bias in primary school. *Journal of Language and Social Psychology*, 36(4):415–437.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and Phrases and their compositionality. *NIPS*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pages 1–12.

Douglass C North et al. 1990. *Institutions, institutional change and economic performance*. Cambridge university press.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Ssrn*, pages 1–44.

Beata Ozieblowska. 1994. *Generic Pronouns in Current Academic Writing*. Ph.D. thesis, ProQuest Dissertations & Theses,.

Irene Lopez Rodriguez. 2009. Of women, bitches, chickens and vixens: Animal metaphors for women in english and spanish. *Cultura, lenguaje y representación: revista de estudios culturales de la Universitat Jaume I*, pages 77–100.

Kimberly T Schneider, Joe Tomaka, and Rebecca Palacios. 2001. Women's cognitive, affective, and physiological reactions to a male coworker's sexist behavior 1. *Journal of Applied Social Psychology*, 31(10):1995–2018.

Janet A Sniezek and Christine H Jazwinski. 1986. Gender bias in english: In search of fair language. *Journal of Applied Social Psychology*, 16(7):642–662.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social communication*, pages 163–187.

Melanie C Steffens, Maria Angels Viladot, and Maria Àngels Viladot. 2015. *Gender at work: A social psychological perspective*. Peter Lang New York, NY.

Jane G Stout and Nilanjana Dasgupta. 2011. When he doesn't mean you: Gender-exclusive language as ostracism. *Personality and Social Psychology Bulletin*, 37(6):757–769.

Jeffrey L. Stringer and Robert Hopper. 1998. Generic he in conversation? *Quarterly Journal of Speech*, 84(2):209–221.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. (2017):1630–1640.

Songül Tolan. 2018. Fair and Unbiased Algorithmic Decision Making : Current State and Future Challenges. *Digital Economy Working Paper 2018-10; JRC Technical Reports.*, (December).

Nneka Umera-Okeke. 2012. Linguistic sexism: an overview of the english language in everyday discourse. *AFRREV LALIGENS: An international journal of language, literature and gender studies*, 1(1):1–17.

Dries Vervecken, Bettina Hannover, and Ilka Wolter. 2013. Changing (s) expectations: How gender fair job descriptions impact children's perceptions and interest regarding traditionally male occupations. *Journal of Vocational Behavior*, 82(3):208–220.

Sandra R Waxman. 2013. Building a better bridge. *Navigating the social world: What infants, children, and other species can teach us*, pages 292–296.

Benjamin Lee Whorf. 1956. Language, thought, and reality: selected writings of. . . .(edited by john b. carroll.).