https://doi.org/10.33919/JCAL.23.1.5

GENDER-NEUTRAL LANGUAGE USE IN THE CONTEXT OF GENDER BIAS IN MACHINE TRANSLATION
(A REVIEW OF LITERATURE)

Aida Kostikova

New Bulgarian University, Ghent University

aida.kostikova@ugent.be

Abstract

Gender bias has become one of the central issues analysed within natural language processing (NLP) research. A main concerns in this field relates to the fact that many NLP tools and automatic machine learning systems not only reflect, but also reinforce social disparities, including those related to gender, and language technology is one of the areas in which this issue is pronounced. This paper analyses the problem of gender-neutral language use from the standpoint of gender bias in machine translation (MT). We determine which types of harms can be caused by the failure to reflect gender-neutral language in translation, provide the general definition of gender bias in MT, describe its sources and provide an overview of existing mitigating strategies. One of the main contributions of this work is that it focuses not only on females, but also non-binary people, whose linguistic visibility has been receiving only limited attention from academia. This literature review provides a firm foundation for further research in this area aimed at

addressing the problem of gender bias in machine translation, especially bias linked to representational harms.

Keywords: *gender bias, machine translation, NLP tools, gender-neutral language use, non-binary gender*

1. Introduction

As the adoption of gender-neutral language (GNL) becomes more widespread, it is increasingly important to consider how these trends can be reflected in natural language processing (NLP) applications, especially given the fact that the purpose of GNL is to "reduce gender stereotyping, promote social change and contribute to achieving gender equality" (Papadimoulis, 2018, 3). Failure to adopt more equitable and balanced linguistic practices can lead to bias associated with representational and, ultimately, allocational harms (Crawford, 2017). The major concerns raised by the researches in this field are related to the fact that any type of bias in technology can be detrimental for ensuring social justice, as by hindering the visibility of speech patterns of certain groups and allocating certain stereotypes to them, such systems can perpetuate inequality (Levesque, 2011; Régner et al., 2019).

While much of prior work in the field of gender bias studies gender identity, most is built on techniques which assume that gender is binary. At the same time, there is growing recognition of non-binary gender identities, with numerous ways to refer to non-binary people or to simply not indicate a binary gender (Sun et al., 2021). That is why in it is necessary to take into account strategies aimed at increasing the linguistic visibility of non-binary people in NLP, and, in particular, in machine translation (MT). In this paper, we attempt to analyze the problem of gender bias from the standpoint of GNL use. The goal is to define and classify types of gender bias generated by a biased MT, and identify harms which might occur due to the failure to reflect gender-neutral language in translation; in addition, we provide an overview of gender-neutral strategies and discuss a rationale for their use. Special attention is paid to non-binary language and its application in machine translation.

2. The issue of gender bias in languages/translation/MT

Although natural language processing (NLP) research does not directly involve human subjects (Hovy and Spruit, 2016; Bender et al., 2021), its engagement with language – the main mediator of the human experience –, which shapes communication as well as such cognitive processes as categorization and perception – raises the question of the social impact of language technologies. The major concern raised by researchers in this field is that bias in technologies can undermine any efforts to establish social justice and equality, as they have a direct impact on the allocation of resources integration and the inclusion of certain social groups (Hovy and Spruit, 2016). Among the narrower, but no less significant, issues related to bias in NLP and languages, are exclusion, stereotyping, bias reinforcement and denigration (Bender et al., 2021).

Overall, there is a close link between bias in technology and prejudice (Ferrer et al., 2021), which has certain psychological and sociological implications (Bourguignon et al., 2015). Machine translation (MT) systems are no exception, as they are known to reflect asymmetries, including those related to gender (Prates et al., 2020), and this phenomenon can be manifested in many ways, with issues ranging from gender stereotyping (Olson, 2018) to over-reliance on the so-called "masculine default" (Schiebinger, 2014). Particular attention must be paid to adverse effects that MT systems may have, as it is one of most widely used artificial Intelligence (AI) applications on the Internet, which is also employed indirectly, e.g., through social media (Monti, 2020).

2.1 Bias statement and implications of gender bias

Overall, a model can be regarded as biased in cases when, while being created by and for people (Schnoebelen, 2017), it "systematically and unfairly discriminates against certain individuals or groups in favor of others" (Friedman and Nissenbaum, 1996) and entails risks associated with social exclusion and stigmatisation (Bender et al., 2021). Bias can be represented in multiple parts of a system, including the training data, resources, pretrained models, and algorithms themselves (Zhao et al., 2018; Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018), which can lead to the production of biased predictions and the further reinforcement of biases present in the training sets (Zhao et al., 2017).

Such systems can, therefore, cause representational harms (i.e., diminishing the role and exclusion of social groups and their identity) or allocational harms (i.e., cases were a system limits the access to resources for certain groups or allocates them in an unfair way) (Crawford, 2017). By drawing on the classification used by Savoldi et al., we also consider such harmful dynamics within representational harms, as stereotyping and under-representation (Savoldi et al., 2021). Stereotyping involves the propagation of generalized beliefs about a social group, for example, by assigning less prestigious occupations or negative physical characteristics to women. Under-representation refers to the cases where the visibility of certain social groups is reduced, which in most cases affects women and non-binary individuals. More emphasis will be placed on the second category of harms (under-representation), as it involves cases of misgendering and ignoring gender-neutral forms, which is precisely the object of our study.

Within the classification framework developed by Dinan et al., who defines harms based on gender dimensions (bias when speaking about someone or gender of the topic, bias when speaking to someone or gender of the addressee, and bias from speaking as someone or gender of the speaker), failure to convey gender-fair language can be described as, on the one hand, misrepresentation when talking "about" certain groups, and on the other hand as reduced visibility of the language used "by" speakers of such groups, which can be detrimental for reflection of their identity and communicative repertoires. In other words, an MT system which does not recognize or reflect certain linguistic expressions of gender might present a barrier for communication and produce an output that "indexes unwanted gender identities and social meanings" (Dinan et al., 2020).

In a broader context, such trends also have an impact on indirect stakeholders, because a biased MT system does not only contribute to the reinforcement of stereotypical assumptions and prejudices (Levesque, 2011; Régner et al., 2019), but promotes language features used by the dominant group, and consequently their establishment as appropriate or prestigious variants (Tallon, 2019). The issue is compounded by prioritization of the overall quality of an MT output, which in most cases is viewed as acceptable by an MT user and perceived as the linguistic norm in a given language (Martindale and Carpuat, 2018). Therefore, there is a close link between representational and allocational harms, which manifests itself in performance disparities across users in the quality of service *(*Savoldi et al., 2021*)*.

2.2 Sources of gender bias in MT

Considering the complexity of implications of gender bias in MT described above, it can be assumed that this problem goes beyond the scope of machine translation. MT and NLP models are considered to exemplify unwanted gender biases present in society (Bolukbasi et al., 2016; Hovy and Spruit, 2016; Caliskan et al., 2017; Rudinger et al., 2018; Garg et al., 2018; Gonen and Goldberg, 2019; Dinan et al., 2020). Some researchers have also emphasized multidimensionality of gender bias sources, among which, for example, there are such broad categories as pre-existing, technical and emergent bias (Friedman and Nissenbaum, 1996).

Pre-existing bias refers precisely to any asymmetries which are rooted in society at large or which reflect personal biases of individuals responsible for the system development. In the context of NLP, this could also include subtle connotational characteristics that permeate language structure and use, as well as gender imbalances. These are manifested most notably through the generic masculine, in which referents in discourse are considered to be men by default – unless explicitly stated (Silveira, 1980; Hamilton, 1991). This affects affects not only women, but also non-binary people (Barker and Richards, 2015).

Technical bias emerges during data collection, system design, training and testing procedures. If present in the data used by these processes, asymmetries in the semantics of language use and gender distribution are respectively inherited by the output of the MT (Caliskan et al., 2017). Methods of mitigating bias at this stage include careful data curation (Barocas et al., 2019; Paullada et al., 2020; Koch et al., 2021; Bender et al., 2021), paired with analyses of what is acceptable from the social and pragmatic points of view (Sap et al., 2020; Devinney et al., 2020, Hovy and Yang, 2021), as well as credible annotation practices (Waseem, 2016, Gaido et al., 2020).

Emergent bias typically occurs after design completion and includes cases of mismatch between users and system design, loss of relevance due to shifts in context of use. An example of emergent bias in MT might be the inability of a system to preserve the linguistic style of a social group or to assign correct gender to its potential users (Hovy et al., 2020).

2.3 Challenges and bias mitigation strategies

The majority of mitigating strategies address technical bias: some studies considered, for example, model debiasing with the help of both internal components – like gender tags (Vanmassenhove et al., 2018) and debiased word embeddings (Bolukbasi, 2016; Escudé Font and Costa-jussà, 2019) – and external components integrated with the MT model, such as lattice re-scoring modules (Saunders and Byrne, 2020) and black-box injections (Moryossef et al., 2019). Research is also being carried out within the context of training data (Reddy and Knight 2016; Zhao et al., 2017; Webster et al., 2018) and evaluation methods (Rudinger et al., 2018; Zhao et al., 2018) improvement. However, as some experts have pointed out, these efforts follow a more focused approach within NLP, and lack a human-computer interaction component which is crucial for the development of gender-inclusive systems (Savoldi et al., 2021; Monti, 2020).

What is more, within these proposed strategies, with a few notable exceptions (Cao and Daumé III, 2020; Saunders et al., 2020; Sun et al., 2021), the discussion around gender bias has been reduced to the binary dichotomy. Current language models can perpetrate harms such as the cyclical erasure of non-binary gender identities (Uppunda et al., 2021) rooted in model and dataset biases "due to tainted examples, limited features, and sample size disparities" (Dev et al., 2021), which, in turn, result from the exclusion and an underrepresentation of non-binary genders in society (Rajunov and Duane, 2019). Therefore, an additional challenge in addressing gender bias in MT concerns the need in reshaping the understanding of gender in language technologies in a more inclusive manner – a problem which is well documented in the field (Dev et al., 2021; Savoldi et al., 2021; Misiek, 2020).

3. Gender-neutral language

Being centered around such a complex social phenomenon as gender, gender-neutral language has not yet achieved universal understanding. Moreover, there is no consensus concerning the definition of gender-fairness in language, also referred to as gender-inclusive, gender-fair or genderless, while the exact approach really depends on the conceptual model of a language and social group it is aimed at. In this section, we provide an overview of gender-neutral language and strategies in this field.

3.1 Definition and general information

Gender-fair language (GFL) was introduced as a response to linguistic gender asymmetry and as part of a broader attempt to reduce stereotyping and discrimination in language (Fairclough, 2003; Maass et al., 2013). By avoiding unfounded, unfair and discriminatory reference to certain social groups, it helps to reduce unfavorable cognitive and behavioral biases and promotes gender equality (Stahlberg et al., 2007). Past research has revealed that gender-fair forms evoke fewer male representations than masculine generics (e.g. Irmen, 2007) and influence individuals' attitudes and perceptions: for example, they lead to more favorable hiring decisions for women and positively influence women's motivation and self-assessment in job interviews (Horvath and Sczesny, 2016; Stout and Dasgupta, 2011). Ultimately, an overall purpose of gender-fair language is to include everybody, regardless of gender and/or sexuality (Douglas and Sutton, 2014; Sczesny et al., 2016). Given that language not only reflects stereotypical beliefs but also affects recipients' cognition and behavior (Menegatti, 2017), the use of expressions consistent with social groups' gender and self-perception can help prevent reinforcement of a biased belief system and prevent discrimination.

However, while a lot of effort has been put into representing female populations in language, non-binary language use has not received enough attention in academia. New developments aimed at ensuring gender equality in languages are often perceived as *excess*ive, and this especially concerns the cases when people "do not conform to cis-normative standards of femininity or masculinity" (Airton, 2018). Additionally, there is a lack of non-binary studies within the machine translation field, as has been pointed out by a number of researchers (Dev et al., 2021, Savoldi et al., 2021, Misiek, 2020). All these factors might result in the adverse effects described in the previous section, especially given the fact that language has been central to the emergence of non-binary gender identities, as challenging cis-normativity – the idea that linguistic categories such as man and woman are "normal" or "natural" – is at the heart of non-binary thinking (Cordoba, 2020).

Moreover, a number of GFL guidelines developed by major international organizations (such as the UN and the European Parliament) still make no mention of strategies to address non-binary people in language, and focus on discrimination and exclusion of women (Trainer, 2021); existing strategies in ensuring gender-fair language are not always aimed at other social groups apart from males and females (Lindqvist et al., 2019) or are not sufficiently disseminated (Harris et al., 2017; McGlashan and Fitzpatrick, 2018; Zimmer and Carson, 2012).

3.2 Gender-neutral language frameworks

When defining a gender-neutral language strategy, a broader as well as narrower approach can be taken. Firstly, linguistic structures used to refer to the extra-linguistic reality of gender vary across languages (Savoldi, 2021), and their type in terms of grammatical gender system defines the means by which gender-fairness is achieved.

In general, different strategies can be used to make language gender-fair and avoid the detrimental effects of masculine generics. The choice of an appropriate strategy depends on the type of language concerned: there are genderless languages (Finnish, Turkish), where gender-specific repertoire is at its minimum; notional gender languages (Danish, English), which display characteristics of lexical gender (*mom/dad*), as well as a system of pronominal gender (*she/he*, *her/him*); and grammatical gender languages (e.g., German, French, Arabic), where each noun pertains to a class such as masculine, feminine, and, if present, neuter. Grammatical gender languages are also characterized by the semantic assignment of gender markings to human referents and a system of morphosyntactic agreement (Stahlberg, 2007; Savoldi et al., 2021).

A gender-fair strategy that has been especially recommended for notional gender languages (Hellinger and Bußmann, 2003) and genderless languages is neutralization. In the framework of neutralization gender-marked terms are replaced by gender-indefinite nouns (English *policeman* by *police officer*). In grammatical gender languages, gender-differentiated forms are replaced, for instance, by epicenes (e.g., *Staatsoberhaupt,* or *Fachkraft* in German). In contrast, feminization which is based on the replacement of masculine generics by feminine-masculine word pairs (e.g., *Elektrikerinnen und Elektriker)* has been recommended for grammatical gender languages.

Even though feminization increases women's visibility, and hence creates more diverse mental images to whom individuals referred (Stahlberg et al., 2001), previous research is inconclusive regarding whether paired forms can eliminate the male bias (Lindqvist et al., 2019). What is more, while neutralization helps avoid male bias and therefore indirectly takes into account all genders, feminization does not solve the problem with the exclusion of non-binary people. Therefore, recent research has been proposing such approaches as gender-neutrality (which is closer to the idea of neutralization) and gender-inclusivity (del Rio-Gonzalez, 2021). These approaches can be considered as the same concept (Papadimoulis, 2018; Lindqvist et al., 2019; Bonnin, 2021), as different aspects or degrees of the single phenomenon (Sczesny et al., 2016), (EIGE, 2019) or two separate strategies, where the term gender-neutral language (GNL) is used to describe a language which avoids any classification of sex or gender, whereas gender-inclusive language (GIL) explicitly challenges binary notions of gender and recognizes the plurality of identities beyond feminine–masculine dimensions (del Río González, 2021).

Some researchers also distinguish between direct and indirect non-binary language (López, 2019a, 2019b). Indirect non-binary language, or INL, aims to refer to all genders without using gender markers – by employing certain linguistic strategies such as using participles instead of adjectives (*Studierende* instead of *Studenten und Studentinnen*) or the use of epicenes *(el pueblo argentino* or *las personas argentinas* instead of *los argentinos*), which makes it similar to the gender-neutral strategies described above. Direct non-binary language, or DNL, is much more obvious because it uses neomorphemes and neopronouns such as *ze* and *zir*, and this strategy can therefore be considered within the framework of gender-inclusive approach. Both categories are considered to be equally important and deserve the attention of practitioners because, although their main objective is to break the generic conception of the masculine, the two categories convey radically different messages: DNL communicates unequivocally that the author respects and supports non-binary people, while the use of INL is perfect for mixed-gender contexts (López, 2020).

Although the use of new grammatical gender systems and direct non-binary language in general (López, 2020) seems to be a rather controversial decision in translation, one should not lose sight of the fact that language is a marker of social belonging (Cordoba, 2020), and the refusal to recognize any social groups in language can contribute to discrimination and social exclusion (Sczesny, 2016). Increasing the linguistic visibility of non-binary people and women takes on special significance in the case of grammatical gender languages, as countries with this language type were found to reach lower levels of social gender equality than countries with notional gender languages or genderless languages. This suggests that there is a close link between the level of gender asymmetries present in language and societal gender inequalities (Hausmann et al., 2009, Wasserman and Weseley, 2009). Additionally, despite the difficulties in implementation and promotion of gender-fair language, there are general positive trends in the language communities in supporting strategies aimed at linguistic inclusion of different social groups (Hekanaho, 2020). Hostile and negative reactions towards new language trends challenging the binary gender system seem to normalize rather quickly (Sendén et al., 2015), especially with active efforts to raise awareness about the advantages, benefits and importance of gender-fair languages (Sczesny and Koeser, 2014).

3.3 Gender-neutral language in machine translation

The problem of GNL is receiving increasing attention from academia. Studies related to gender bias concern not only trends which could potentially harm women, but also non-binary people – for example, Dev et al., analyze the complexity of gender and its linguistic representation, and provide the results of a survey on gender-related harms associated with language technologies conducted among non-binary persons. Among three common NLP tasks (Named Entity Recognition, Coreference Resolution, and Machine Translation) included in the survey, misgendering was one of the most frequently mentioned issues, and in terms severity of harms machine translation was the cause of major concern (Dev et al., 2021).

Some efforts in the NLP community were mainly aimed at solving a problem of underrepresentation of non-binary individuals in task-specific data sets: for example, Cao and Daumé III (2020 and 2021) introduce a gender-inclusive dataset GICoref for coreference resolution; in MT, Saunders et al. have presented a method of tagging words with target language gender inflection (Saunders et al., 2020). Apart from approaches that incorporate additional meta-data during training and testing, allowing for a controlled generation of gender alternatives (Bau et al., 2019; Habash et al., 2019; Alhafni et al., 2020), research in this area also concerns generation of gender variants or gender rewriting. For example, Sun et al. (2021) and Vanmassenhove et. al (2021) present a rule-based and neural rewriter for the generation of gender-neutral singular *they* sentences; however, research in this area is monolingual and is limited to English-specific gender-neutral writing, and, more specifically, only the *they* pronoun.

Although the underlying goal of works in this field is to provide more possibilities for the users to make their preferred linguistic choices, thereby empowering people and whole social groups "to interact with technology in a way that is consistent with their social identity" (Sun et al., 2021), there are still challenges at the intersection of gender-fair language and machine translation: firstly, there is insufficient real-world data for all the GNL strategies (and, more specifically, neopronouns); secondly, solutions in this field consider non-binary genders as a static third category which exists next to male and female genders (Dev et al., 2021), when in reality it is of a fluid and diverse nature.

4. Conclusion

This literature review lays the groundwork for further research, the purpose of which will be to assess the efficiency of machine translation in relation to gender-neutral language use. To this end, we categorized the gender-neutral language problem in terms of gender bias in machine translation, presented existing approaches to gender-neutral language and provided an overview of different strategies in machine translation aimed at mitigating representational harms caused by a biased system.

References

Airton, L. (2018). The De/politicization of Pronouns: Implications of the No Big Deal Campaign for Gender-expansive Educational Policy and Practice. *Gender and Education,* 30(6), 790–810.

Alhafni, B., Habash, N. and Bouamor, H. (2020). Gender-aware Reinflection Using Linguistically Enhanced Neural Models. In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 139–150.

Bau, A., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F. and Glass, J. (2019). Identifying and Controlling Important Neurons in Neural Machine Translation. In: *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*.

Barocas, S., Hardt, M. and Narayanan, A. (2017). Fairness in Machine Learning. *Nips Tutorial,* 1, 2.

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V. and Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016).* NY: Curran Associates Inc., 4356–4364.

Bonnin, J. E. and Coronel, A. A. (2021). Attitudes Toward Gender-Neutral Spanish: Acceptability and Adoptability. *Frontiers in sociology,* 6, 35.

Bourguignon, D., Yzerbyt, V. Y., Teixeira, C. P. and Herman, G. (2015). When Does it Hurt? Intergroup Permeability Moderates the Link Between Discrimination and Self-esteem. *European Journal of Social Psychology,* 45(1), 3–9.

Caliskan, A., Bryson, J. J. and Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science,* 356(6334), 183–186. Available at: http://opus.bath.ac.uk/55288/.

Cao, Y. T. and Daumé III, H. (2021). An Analysis of Gender and Bias Throughout the Machine Learning Lifecyle. *Computational Linguistics*, 47(3), 615–661. Available at: https://doi.org/10.1162/coli_a_00413.

Cordoba, S. (2020). Exploring non-binary genders: language and identity. [PhD diss.]. De Montfort University.

Crawford, K. (2017). The Trouble with Bias – NIPS 2017 Keynote – Kate Crawford #NIPS2017. *YouTube.* [Video]. Available at: https://www.youtube.com/watch?v=fMym_BKWQzk&t=10s.

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M. and Chang, K. W. (2021). Harms of Gender Exclusivity and Challenges in Non-binary Representation in Language Technologies. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1968–1994.

Devinney, H. Björklund, J. and Björklund, H. (2020). Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish. In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 79–92. Available at: https://aclanthology.org/2020.gebnlp-1.8/.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D. and Williams, A. (2020). Multidimensional Gender Bias Classification. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 314–331.

Douglas, K. M. and Sutton, R. M. (2014). "A Giant Leap for Mankind" But What About Women? The Role of System-justifying Ideologies in Predicting Attitudes Toward Sexist Language. *Journal of Language and Social Psychology*, 33(6), 667–680.

EIGE – European Institute for Gener Equality. (2019). *Toolkit on Gender-sensitive Communication. A resource for policymakers, legislators, media and anyone else with an interest in making their communication more inclusive*. Publications Office of the European Union. Available at: https://eige.europa.eu/sites/default/files/20193925_mh0119609enn_pdf.pdf.

Fairclough, N. (2001) *Language and Power.* 2nd ed. Harlow: Pearson Education.

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M. and Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective. *IEEE Technology and Society Magazine*, 40(2), 72–80.

Font, J. E. and Costa-Jussa, M. R. (2019). Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 147–154.

Friedman, B., and Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems (TOIS)*, 14(3), 330–347.

Gaido, M., Savoldi, B., Bentivogli, L., Negri, M. and Turchi, M. (2020). Breeding Gender-aware Direct Speech Translation Systems. *arXiv.org e-Print arXiv:2012.04955.* Available at: https://arxiv.org/abs/2012.04955.

Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J., 2018. Word Embeddings Quantify 100 years of Gender and Ethnic Stereotypes. In: *Proceedings of the National Academy of Sciences*, 115(16), E3635-E3644.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings but do not Remove Them. *arXiv.org e-Print arXiv:1903.03862.* Available at: https://arxiv.org/abs/1903.03862.

Gustafsson Sendén, M., Bäck, E.A. and Lindqvist, A. (2015). Introducing a Gender-neutral Pronoun in a Natural Gender Language: The Influence of Time on Attitudes and Behavior. *Frontiers in psychology*, 6, 893.

Hamilton, M.C. (1991) Masculine Bias in the Attribution of Personhood: People = male, male = people. *Psychology of Women Quarterly*, 15(3), 393–402.

Harris, C. A., Biencowe, N. and Telem, D. A. (2017) What's in a Pronoun? Why gender-fair Language Matters. *Annals of Surgery,* 266(6), 932.

Hausmann, R., Tyson, L. D., and Zahidi, S. (2009). *The Global Gender Gap Report 2009*. Geneva: World Economic Forum.

Hekanaho, Laura. (2020). Generic and nonbinary pronouns: usage, acceptability and attitudes. [PhD diss.]. Helsingfors University, Helsinki.

Hellinger, M., and Bußmann, H. (2001, 2002, 2003). *Gender Across Languages: The Linguistic Representation of Women and Men, Vol. 1, 2, 3.* John Benjamins Publishing Company.

Horvath, L. K., and Sczesny, S. (2016).Reducing Women's Lack of Fit with Leadership? Effects of the Wording of Job Advertisements. *European Journal of Work and Organizational Psychology,* 25(2), 316–328.

Hovy, D. and Yang, D. (2021). The Importance of Modeling Social Factors of Language: Theory and Practice. In: *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, 588–602.

Hovy, D. and Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2 (Short Papers)*, 591–598.

Hovy, D., Bianchi, F. and Fornaciari, T. (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*. Online: Association for Computational Linguistics, 1686–1690. Available at: https://aclanthology.org/2020.acl-main.154/.

Irmen, L. (2007). What's in a (Role) Name? Formal and Conceptual Aspects of Comprehending Personal Nouns. *Journal of Psycholinguistic Research,* 36(6), 431–456.

Koch, B., Denton, E., Hanna, A. and Foster, J. G. (2021). Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). arXiv.org e-Print arXiv:2112.01716.* Available at: https://arxiv.org/abs/2112.01716.

Koeser, S. and Sczesny, S. (2014). Promoting Gender-fair Language: The Impact of Arguments on Language Use, Attitudes, and Cognitions. *Journal of Language and Social Psychology,* 33(5), 548–560.

Levesque, R. J. (2011). Sex Roles and Gender Roles. In: *Encyclopedia of Adolescence*. Springer International Publishing, 2622–2623.

Lindqvist, A., Renström, E. A. and Gustafsson Sendén, M. (2019). Reducing a Male Bias in Language? Establishing the Efficiency of Three Different Gender-fair Language Strategies. *Sex Roles,* 81(1), 109–117.

López, Á (2020). Cuando el lenguaje excluye: consideraciones sobre el lenguaje no binario indirecto, *Cuarenta naipes*, (3), 295–312.

López, Á (2021). Direct and Indirect Non-binary Language in English to Spanish Translation. In: *27th Annual Lavender Languages and Linguistics Conference*, Online, 21–23.

Maass, A., Suitner, C. and Merkel, E. M. (2013). Does Political Correctness Make (social) Sense? In: Forgas, J. P., Vincze, O. and László J., (eds.). *Social Cognition and Communication*. Psychology Press, 345–360.

María del Río-González, A. (2021) To Latinx or not to Latinx: A Question of Gender Inclusivity Versus Gender Neutrality. *American Journal of Public Health*, 111(6), 1018–1021.

Martindale, M.J. and Carpuat, M. (2018). Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Association for Machine Translation in the Americas, 13–25. Available at: https://aclanthology.org/W18-1803.pdf.

McGlashan, H. and Fitzpatrick, K. (2018). "I Use Any Pronouns, and I'm Questioning Everything Else": Transgender Youth and the Issue of Gender Pronouns. *Sex Education*, 18(3), 239–252.

Menegatti, M. and Rubini, M. (2017). Geneder Bias and Sexism in Language. In: *Oxford Research Encyclopedia of Communication*. Oxford University Press, 451–468.

Misiek, S. (2020) Misgendered in Translation? Genderqueerness Polish Translations of English-language Television Series. *Anglica. An International Journal of English Studies,* 29(2), 165–185.

Monti, J. (2020). Gender Issues in Machine Translation: An Unsolved Problem? In: von Flotow, L. and Hālah, K., (eds.). *The Routledge Handbook of Translation, Feminism and Gender*. Abingdon Oxon: Routledge, 457–468.

Moryossef, A., Aharoni, R. and Goldberg, Y. (2019). Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing,* 49–56. *arXiv.org e-Print arXiv:1903.03467.* Available at: https://arxiv.org/abs/1903.03467.

Habash, N., Bouamor, H. and Chung, C. (2019). Automatic Gender Identification and Reinflection in Arabic. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 155–165.

Olson, P. (2018). The Algorithm that Helped Google Translate Become Sexist. *Forbes.* [Online]. Available at: https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/?sh=d675b9c7daa2.

Papadimoulis, D. (2018). *Gender-neutral Language in the European Parliament.* Brussels: European Parliament.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E. and Hanna, A. (2021). Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns*, 2(11), 100336.

Prates, M. O. R., Avelar, P. H. and Lamb, L. C. (2020). Assessing Gender Bias in Machine Translation: A Case Study with Google Translate. *Neural Comput & Applic*, 32, 6363– 6381. Available at: https://doi.org/10.1007/s00521-019-04144-6.

Reddy, S. and Knight, K. (2016). Obfuscating Gender in Social Media Writing. In: *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics, 17–26.

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T. and Huguet, P. (2019). Committees with Implicit Biases Promote Fewer Women When they do not Believe Gender Bias Exists. *Nature Human Behavior,* 3(11), 1171–1179.

Richards, C. and Barker, M. J. (2015). *The Palgrave Handbook of the Psychology of Sexuality and Gender.* Palgrave Macmillan.

Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B. (2018). Gender Bias in Coreference Resolution. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2 (Short Papers)*. Association for Computational Linguistics, 8–14.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A. and Choi, Y. (2019). Social Bias Frames: Reasoning about Social and Power Implications of Language. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 5477–5490. *arXiv.org e-Print arXiv:1911.03891.* Available at: https://arxiv.org/abs/1911.03891.

Saunders, D., Sallis, R., and Byrne, B. (2020). Neural Machine Translation doesn't Translate Gender Coreference Right Unless You Make It. In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 35–43.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M. (2021). Gender Bias in Machine Translation. In: *Transactions of the Association for Computational Linguistics*, 9, 845–874.

Schiebinger, L. (2014). Scientific Research Must Take Gender into Account. *Nature,* 507, 9.

Schnoebelen, T. (2017). Goal-oriented Design for Ethical Machine Learning and NLP. In: P*roceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 88–93.

Sczesny, S., Formanowicz, M. and Moser, F. (2016). Can Gender-fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in psychology,* 7, 25.

Silveira, J. (1980). Generic Masculine Words and Thinking. *Women's Studies International Quarterly*, 3(2-3), 165–178.

Stahlberg, D., Braun, F., Irmen, L. and Sczesny, S. (2007). Representation of the Sexes in Language. In: Fiedler, K. (ed.). *Social communication*. Psychology Press, 163–187.

Stahlberg, D., Sczesny, S. and Braun, F. (2001). Name Your Favorite Musician: Effects of Masculine Generics and of Their Alternatives in German. *Journal of Language and Social Psychology*, 20(4), 464–469.

Stout, J. G. and Dasgupta, N. (2011). When He doesn't Mean You: Gender-exclusive Language as Ostracism, *Personality and Social Psychology Bulletin*, 37(6), 757–769.

Sun, T., Webster, K., Shah, A., Wang, W. Y. and Johnson, M. (2021). They, Them, Theirs: Rewriting with Gender-neutral English. *arXiv.org e-Print arXiv:2102.06788.* Available at: https://arxiv.org/abs/2102.06788.

Switzer, J. Y. (1990). The Impact of Generic Word Choices: An Empirical Investigation of Age- and Sex-related Differences. *Sex Roles*, 22(172), 69–81.

Tallon, T. (2019). A Century of "shrill": How Bias in Technology has Hurt Women's Voices. *The New Yorker*. Available at: https://www.newyorker.com/culture/cultural-comment/a-century-of-shrill-how-bias-in-technology-has-hurt-womens-voices.

Trainer, T. (2021). The (non) Binary of Success and Failure: A Corpus-based Evaluation of the European Parliament's Commitment to Using Gender-neutral Language in Legislation Published in English and Portuguese. [Master's thesis]. University of Porto.

Turchi, M., Negri, M., Farajian, M. and Federico, M. (2017). Continuous Learning from Human Post-edits for Neural Machine Translation. *The Prague Bulletin of Mathematical Linguistics,* 108, 233–244.

Uppunda, A., Cochran, S.D., Foster, J. G., Arseniev-Koehler, A., Mays, V. M. and Chang, K. (2021). Adapting Coreference Resolution for Processing Violent Death Narratives. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4553–4559. *arXiv.org e-Print arXiv:2104.14703*. Available at: https://arxiv.org/abs/2104.14703.

Vanmassenhove, E., Emmery, C. and Shterionov, D. (2021). NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 8940–8948. *arXiv.org e-Print arXiv:2109.06105*. Available at: https://arxiv.org/abs/2109.06105.

Vanmassenhove, E., Hardmeier, C. and Way, A. (2018). Getting Gender Right in Neural Machine Translation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3003–3008. *arXiv.org preprint arXiv:1909.05088.* Available at: https://arxiv.org/abs/1909.05088.

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In: P*roceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, 138–142.

Wasserman, B. D., and Weseley, A. J. (2009). ¿Qué? Quoi? Do Languages with Grammatical Gender Promote Sexist Attitudes? *Sex Roles,* 61, 634–643.

Webster, K., Recasens, M., Axelrod, V. and Baldridge, J. (2018). Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6, 605–617.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K. W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-level Constraints. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2979–2989. *arXiv.org e-Print arXiv:1707.09457.* Available at: https://arxiv.org/abs/1707.09457.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. and Chang, K. W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2 (Short Papers)*. Association for Computational Linguistics, 15–20. *arXiv.org e-Print arXiv:1804.06876*. Available at: https://arxiv.org/abs/1804.06876.

Zimmer, B. and Carson, C. E. (2012). Among the New Words. *American speech*, 87(4), 491–510.