ISBN: 978-84-1091-017-1 (PDF)
DOI: https://doi.org/10.14201/0AQ0373673681

Gender Bias in Machine Translation: Assessing Human Translators' Responsibility

Sesgos de género en la traducción automática: la responsabilidad del traductor humano

Marta GARCÍA GONZÁLEZ
Universidade de Vigo
mgarciag@uvigo.gal

ABSTRACT: Given the increasing role of machine translation (MT) in the industry, it is essential to ensure that translation outputs do not contribute to the perpetuation and increase of bias within and across societies. While recent NLP and MT research highlight the importance of language in shaping societal and cultural norms, there remains considerable scope for research. Specifically, the literature on gendered technology has yet to thoroughly explore the role played by human translators in the perpetuation of gender-biased language through MT systems. This contribution addresses this gap by investigating the transfer of gender-neutral language in a collection of translations from English into Galician conducted by various MT systems, and a group of translation students. The study compares the translation choices made by MT systems and students regarding gender-neutral terms and phrases, and underscores the pivotal role of human translators in ensuring gender-sensitive language usage in the age of human-machine interaction.

KEYWORDS: gender-biased language; gender neutrality; human translation (HT); machine translation (MT); translator's responsibility; representational harms.

RESUMEN: El creciente uso de la traducción automática (TA) hace necesario garantizar que los textos meta no contribuyen a perpetuar y aumentar la presencia de sesgos. Aunque estudios recientes en el ámbito del PNL y la TA han destacado la importancia del lenguaje en la configuración de normas sociales y culturales, es preciso seguir investigando, particularmente sobre el papel de la traducción humana en la perpetuación de sesgos de género en la TA. Esta contribución se centra en este aspecto, concretamente en la transferencia de la neutralidad de género en una selección de traducciones del inglés al gallego realizadas por varios sistemas de TA y por un grupo de estudiantes de traducción.

El estudio compara las decisiones de los sistemas automáticos y las estudiantes en la traducción de términos y expresiones neutras, y subraya el papel fundamental de la traducción humana para garantizar el uso de un lenguaje sensible al género.

PALABRAS CLAVE: lenguaje con sesgo de género; neutralidad de género, perjuicios de representación, responsabilidad de la traductora, traducción automática; traducción humana.

1. INTRODUCTION

In the past few years, machine translation (MT) has evolved considerably and Neural Machine Translation Systems (NMTSs) and Large Language Models (LLMs) provide reasonably good translations that are being used by the main economic players. This has already brought considerable changes into the translation industry and is to bring many others in the near future. In this context, stakeholders from academia and professional translation groups are beginning to express concerns about how the relentless advancement of AI will impact the translator profession as we know it today. Beyond the concerns about how the development of MT affects the profession, however, an equally pressing concern arises from how the uncontrolled use of MT is perpetuating and increasing the use of gender-biased language. Although gender bias in Natural Language Processing (NLP) and MT has been a recurrent object of study, focus has been mainly on solving technical problems, namely on the implementation of debiasing techniques, rather than on identifying and removing bias in the datasets feeding the machines.

As Savoldi et al. (2021) argue, «MT is not only built for people but also by people». We should not forget that MT systems are partly trained and fine-tuned with datasets that are either drafted or postedited by human translators. Failure by human translators to avoid gender-biased language in their translations and to identify it in post-editing MTs contributes to bias perpetuation and might demand a reconsideration of human translators' responsibility. This study aims to assess the responsibility of translators in gender-bias perpetuation by evaluating the extent to which future translators are able to produce gender-neutral target texts and their awareness of the need to do so. For this purpose, reference shall be made to the results of a comprehensive study on preservation of gender neutrality in machine and human translation (García González 2024), which compared the translation strategies applied within a corpus of translations from English to Galician conducted by various MT systems and by translation and interpreting students. Following a brief presentation of the theoretical background (section 2) and the materials, methods, and main results of the reference study (section 3), this contribution discusses the responsibility of human translators in perpetuating or reducing gender-biased language in MT.

2. BACKGROUND

The European Institute for Gender Equality (2019) defines gender-biased language as «language that either implicitly or explicitly favors one gender over another». The

presence of bias in NLP systems has been thoroughly addressed in the last years by researchers, mainly in the field of computer sciences (Blodgett et al. 2020, Stanczak and Augenstein 2021). The main focus of study, however, has been placed on addressing technical issues, with different system debiasing techniques being implemented. Nevertheless, recent voices have claimed the need to identify the origin of bias and remove it from training and fine-tuning datasets. Moreover, some authors (Blodgett et al. 2020, 5460) have highlighted the importance of clearly defining the ways systems bias are harmful, to whom and why.

Gender-biased language has been described to cause allocational and representational harms (Barocas et al. 2017), which affect individuals' opportunities and reinforce negative stereotypes, leading to systemic inequality. Allocational harms occur when opportunities, resources, or responsibilities are unfairly distributed based on gender, often as a result of gender-biased language. For example, when job titles are gendered, such as «fireman» or «policeman», or when male pronouns are used in job descriptions, it implies these roles are inherently male, which can dissuade women from pursuing these careers (Stout and Dasgupta 2011). In the workplace, women are often described with terms like «helpful» or «compassionate», which does not highlight their efficiency or leadership skills, potentially affecting their career advancement. Moreover, gender biases in language can result in a skewed distribution of labor market opportunities, financial remuneration, or job stability, preferentially granted based on gender. Representational harms, in turn, refer to the ways in which gender-biased language perpetuates harmful stereotypes, misrepresentation, and discrimination. They have been typically divided into stereotyping (associating certain qualities or roles with specific genders), denigration (when language degrades the social status of individuals by framing certain ways of being as the norm), underrepresentation (of certain genders in various fields and roles) (Barocas et al. 2017), misgendering (using incorrect pronouns or gendered terms that do not align with an individual's affirmed gender identity, Ackerman 2019), and erasure (when a group is erased or made invisible by a system, Dev et al. 2021).

While there has long been consensus regarding the presence of gender bias in NLP, the recent evolution of NMTSs has generated a significant volume of research (Escudé Font and Costa-jussà 2019; Stanovsky et al. 2019; Prates et al. 2020) on their role in perpetuating and amplifying these biases as a result of their failure to adequately transfer gender from source text (ST) to target text (TT). This is particularly the case when translation takes place between languages that differ in how they code gender (Stahlberg et al. 2007), for example when translating from a genderless language as Turkish into a natural gender language as English (Ciora et al. 2019), or from the latter into a grammatical gender language as Spanish, German or Russian (Stanovsky et al. 2019). Such translations are a frequent source of representational harms, namely of stereotyping, misgendering, underrepresentation and erasure. In particular, misgendering in translation occurs when the gender is incorrectly translated (e.g., «My friend is very nice. She is a nurse» translated as *Mi amigo es muy amable. Ella es enfermera*). Stereotyping occurs when certain behaviors, attitudes or features associated with a given gender lead to translate a non-gender-marked term with a gender-marked term (e.g., «nurse» translated

as *enfermera* and «doctor» as *médico*). Underrepresentation occurs when gender-neutral expressions are systematically translated into masculine or generic masculine expressions (e.g., «teachers and students» translated as *profesores y alumnos)*. Finally, erasure occurs, for example, when an intendedly neutral expression is translated with a duplication strategy («teachers» translated as *profesores y profesoras* instead of *profesorado*), which disregards non-binary identities.

To avoid biased behaviors, some authors highlight the need to enable gender-neutral translation (GNT), defined as «the task of automatically translating from one language to another without marking the gender of human referents in the target» (Piergentili et al. 2023, 76). The authors stress the importance of avoiding misgendering individuals when translating from a natural gender to a grammatical gender language, and recommend the implementation of the following three guidelines:

a. Avoiding expressing gender in translation when it cannot be properly assumed in the ST (C1);
b. Using proper expressions of gender in the translation when indirectly expressed in the ST (C2);
c. Avoiding propagating masculine generics from ST to TT (C3).

In general terms, applying strategies C1 and C3 may help avoid stereotyping, misrepresentation and erasure, while C2 avoids misgendering. The introduction of gender-biased language in translation, however, is not an exclusive problem of MT. As Nissen (2002) and more recently Lardelli and Gromann (2022, 2023) and García González (2024) have shown, human translators—who, as translators and post-editors, may play a dual role in perpetuating or reducing MT bias—display similar biased behaviors when faced with ambiguous or apparently ambiguous sentences. Such findings strongly highlight the need for a deeper exploration of the responsibility of human translators as crucial contributors in propagating or mitigating gender bias of training datasets.

3. DESIGN

As a contribution to the above purpose, the reference study (García González 2024) focused on a corpus of translations from English into Galician obtained from several MT systems and from a group of students. Both MT systems and students were assessed in terms of their ability to transfer gender neutrality from the original English ST, as well as their translation decisions when neutralization was difficult or impossible. Translations were obtained from five MT systems and models (Google Translate, Microsoft Bing Translator, Yandex, ChatGPT and You.com) and a group of 53 students comprised of 41 female, 10 male and 2 non-reported students.

The ST was comprised of nine isolated English sentences mainly taken from Piergentili et al. (2023) followed by an ad hoc gender-neutral English short story. Both the sentences and the short story included different words or expressions involving

gender-based translation problems when translating from a natural gender into a grammatical gender language.

Please translate into Galician the following short sentences and short story

Sentences

1. I refuse to give up on ***a single student*** in my class.
2. ***A lot of innovative teachers*** began bringing comics.
3. We train ***nurses*** to do it, and they use local anesthetics.
4. It affects one to two percent of the population, more commonly ***men***.
5. *The* ***fishermen*** were so ***upset*** about not having enough fish to catch that they decided to move to another village.
6. When I was a ***freshman*** in college, I took my first biology class.
7. My ***friend*** was wearing a beautiful summer t-shirt
8. My ***friend*** was wearing a new t-shirt
9. Vehicles may only proceed ***at a walking pace***

Short Story

This is a short-story about three students who became ***friends*** at school and then went to London University to study economics. After they graduated from university, ***each of them*** followed different paths and ten years later they meet and talk about their lives. ***One of them*** became an economist and works at the European Central Bank. ***The second one*** is *a* ***professor*** and lectures economics at university. ***None of them is married***. ***The third of them*** got married and has ***two children*** and after spending these years taking care of ***them***, is now trying to find a job.

***Friend* 1**: Hi! It's been ages. I am so glad to see you two again. How is everything going? What have you been doing all these years?

Friend* 2**: Well... After we graduated, I spent two years working as a sales agent at my father's company but then I decided to go back to university, got a PhD and became a ***professor. I never found the time to get married or have ***children***. What about you?

Friend 3: I did get married and had **two *children***, *Mary and Joan*. Now, I am trying to get a job as a ***nurse***, but it is being difficult. My partner is a ***doctor***, and it is difficult for us to arrange schedules to take care of the ***children***.

Friend 1: I never got married either and I don't think I will, although I do have a partner. We work together as economists at the European Central Bank. You imagine having ***children*** is difficult for us, but we have ***two cats and a dog***.

Figure 1. Source text (Piergentili et al. 2023; García González, 2024)

The MT systems were first prompted to translate all the sentences at the same time and then the whole short story, while the students received the sentences and the short story at the same time and were asked to translate them in the classroom by hand with no computer and no dictionaries. Students were instructed to make the decisions they

deemed necessary to produce a coherent translation, but were given no prior information about the purpose of the exercise.

For the analysis, a classification of gender-neutrality strategies was used, mainly based on the handbook on gender inclusiveness published by the Universidade de Vigo (Bringas López et al. 2012), which suggests several neutralization (generic nouns, rewording and omission), specification (duplication and graphical elements) and neologism-based (abstract nouns) strategies. It is important to note that not all the strategies are appropriate for use in any context, as duplication and abstract nouns refer to indefinite individuals or groups of individuals, and not to definite individuals. In addition, Piergentili et al.'s guidelines (2023) were used to determine whether proper transfer of gender neutrality and of intendedly gendered expressions from ST to TTs was ensured. Although the guidelines were initially suggested as recommendations for MT systems, they are clearly applicable to and desirable for human translators, and were therefore used as a starting point to design the ad hoc ST for MT systems and students.

4. RESULTS

Among the main results of the study, García González (2024) highlights the systematic failure by MT systems to comply with Piergentili et al.'s condition C1, as they tended to translate English neutral terms into masculine Galician terms instead of using available neutral terms, with female and non-binary persons being underrepresented or erased. MT systems did not comply with condition C3 either, as the masculine generic «fishermen» was propagated in the translation. Furthermore, stereotyping was also identified, since, while «doctor», «teacher» or «children» were translated with masculine nouns, the systems mainly translated «nurse» with a feminine noun. This also resulted in a remarkable inconsistency problem: although the five MT systems translated «friend as *amigo*, thus making their translated stories deal with three male friends, all but ChatGPT translated «nurse» into the feminine *enfermeira*. Finally, condition C2 was not met either, as the word «children» accompanied by the named entities Mary and Joan was translated as *fillos* instead of *fillas* (feminine form) or *crianzas* (neutral form), which might be labelled as a case of misgendering (both if we assume that both Mary and Joan are typically female nouns, and if we consider that they might be non-binary individuals).

More concerningly, the study also revealed that many students failed to observe conditions C1 and C3 as well, as they did not avoid expressing gender that could not be assumed in the source («students», «teachers», and «nurses» could have remained neutral if translated with *estudantado*, *profesorado*, and *persoal de enfermaría*) and they did not avoid propagating the purportedly generic masculine «fishermen» (which could be neutralized with *as persoas que estaban a pescar*). In addition, 31 out of 53 students decided to tell their story about three male friends, while only one student decided to tell a story about three female friends and just five of them attempted to produce a completely neutral translation. Students' translations, therefore, displayed similar levels of underrepresentation and erasure to those of their MT counterparts. In contrast, many of

them consciously avoided the stereotypical behavior of translating «nurse» with a feminine noun and «doctor» with a masculine noun.

5. DISCUSSION AND CONCLUSION

The above results have relevant implications for the assessment of human translators' responsibility regarding gender neutrality and inclusiveness. The fact that most of the students did not even consider the possibility of producing a neutral translation reveals a significant lack of awareness of the problem of gender bias in translation. Beyond the use of the generic masculine, the decision to assign a masculine gender to all those characters for whom no gender markers were included, and even for those whose names were more likely to be feminine, would contribute to perpetuate the imbalance in the presence of genders in the translated texts, and thus exacerbate gender biases in the datasets used to train and fine-tune the models for MT.

To help alleviate this trend, it is necessary to adopt measures to promote gender-responsible translation strategies from different fields of action. At the curriculum level, gender neutrality and gender inclusiveness in translations and interpreting curricula need to be addressed in a consistent manner. Students should be made aware of the need to observe conditions C1, C2 and C3, and of the importance of choosing the appropriate strategy among different alternatives. Moreover, they increasingly need training to identify gender bias in MT systems, so they can successfully post-edit MTs. At the translation industry level, translation companies need to be aware of the necessity to involve gender-aware translators in post-editing tasks. They also need to trade off the use of MT and Computed-assisted Translation (CAT) with minimum human intervention for the need to ensure quality and language fairness. Finally, at the research level, there is a need for greater cooperation between the fields of machine translation, translation studies and studies on inclusive and neutral language, which would favor the development of models adapted by language pairs, taking into account the specific problems of gender transfer from the source language to the target language according to their respective ways of encoding gender.

REFERENCES

Ackerman, Lauren. 2019. «Syntactic and Cognitive Issues in Investigating Gendered Coreference». *Glossa: A Journal of General Linguistics* 4 (1): 117. Accessed August 15, 2024. https://www.glossa-journal.org/article/id/5224/.

Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. «The Problem with Bias: Allocative Versus Representational Harms in Machine Learning». In *Proceedings of the 9th Annual Conference of the Special Interest Group in Computing, Information, and Society (SIGCIS), Philadelphia, 29 October 2017.* Philadelphia: SIGCIS.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. «Language (Technology) Is Power: A Critical Survey of "Bias" in NLP». In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics (ACL), *5-10 July 2020,* edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5454-76. Association for Computational Linguistics (ACL).

Bringas López, Ana, Olga Castro, María Jesús Fariña Busto, María Belén Martín Lucas, and Beatriz Suárez Briones. 2012. *Manual de linguaxe non sexista no ámbito universitario*. Vigo: Unidade de Igualdade da Universidade de Vigo.

Ciora, Chloe, Nur Iren, and Malihe Alikhani. 2021. «Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models». In *Proceedings of the 14th International Conference on Natural Language Generation (ICNLG), Aberdeen, 20-24 September 2021, edited by Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada,* 55-63. Aberdeen: Association for Computational Linguistics (ACL).

Dev, Sunipa, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. 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, Punta Cana, 7-11 November 2021*, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 1968-94. Punta Cana: Association for Computation Linguistics (ACL).

Escudé Font, Joel, and Marta R. Costa-jussà. 2019. «Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques». In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Florence, 2 August 2019*, edited by Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, 147-54. Florence: Association for Computational Linguistics (ACL).

European Institute for Gender Equality. 2019. *Toolkit on Gender-sensitive Communication.* Luxembourg: Publications Office of the European Union.

García González, Marta. 2024. «The Role of Human Translators in the Human-Machine Era: Assessing Gender-neutrality in Galician Machine and Human Translation». In *Gendered Technology in Translation and Interpreting: Centering Rights in the Development of Language Technologies*, edited by Esther Monzó-Nebot, and Vicenta Tasa-Fuster. New York: Routledge.

Lardelli, Manuel, and Dagmar Gromann. 2022. «Gender-Fair (Machine) Translation». In *Proceedings of the New Trends in Translation and Technology Conference – NeTTT 2022, Rhodes Island, 4-6 July 2022*, edited by Sheila Castilho, Rocío Caro Quintana, Maria Stasimioti, and Vilelmini Sosoni, 166-77. Shumen: INCOMA Limited.

Lardelli, Manuel, and Dagmar Gromann. 2023. «Translating Non-binary Coming-out Reports: Gender-fair Language Strategies and Use in News Articles». *The Journal of Specialised Translation* 40: 213-40.

Nissen, Uwe Kjær. 2002. «Aspects of Translating Gender». *Linguistik Online* 11 (2): 25-37.

Piergentili, Andrea, Dennis Fucci, Beatrice Savoldi, Mateo Negri, and Luisa Bentivogli. 2023. «Gender Neutralization for an Inclusive Machine Translation: From Theoretical Foundations to Open Challenges». *arXiv,* arXiv:2301.10075 [cs.CL]. Accessed August 15, 2024. https://arxiv.org/abs/2301.10075.

Prates, Marcelo O. R., Pedro H. Avelar, and Luis C. Lamb. 2020. «Assessing Gender Bias in Machine Translation: A Case Study with Google Translate». *Neural Computing & Applications* 32 (10): 6363-81.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. «Gender Bias in Machine Translation». *Transactions of the ACL* 9: 845-74.

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. «Representation of the Sexes in Language». In *Social Communication*, edited by Klaus Fiedler, 163-87. New York: Psychology Press.

Stanczak, Karolina and Isabelle Augenstein. 2021. «A Survey on Gender Bias in Natural Language Processing». In *Journal of the Association for Computing Machinery* 1(1).

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. «Evaluating Gender Bias in Machine Translation». In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, 28 July - 2 August 2019*, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 1679-84. Florence: Association for Computational Linguistics (ACL).

Stout, Jane G., and Nilanjana Dasgupta. 2011. «When He Doesn't Mean "You": Gender-Exclusive Language as Ostracism». *Personality and Social Psychology Bulletin* 37 (6): 757-69.