You Shall Know a Word's Gender by the Company it Keeps: Comparing the Role of Context in Human Gender Assumptions with MT

Janiça Hackenbuchner, Arda Tezcan, Aaron Maladry and **Joke Daems**
Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

In this paper, we analyse to what extent machine translation (MT) systems and humans base their gender translations and associations on role names and on stereotypicality in the absence of (generic) grammatical gender cues in language. We compare an MT system's choice of gender for a certain word when translating from a notional gender language, English, into a grammatical gender language, German, with the gender associations of humans. We outline a comparative case study of gender translation and annotation of words in isolation, out-of-context, and words in sentence contexts. The analysis reveals patterns of gender (bias) by MT and gender associations by humans for certain (1) out-of-context words and (2) words in-context. Our findings reveal the impact of context on gender choice and translation and show that word-level analyses fall short in such studies.

1 Introduction

Aligned with a growing interest and use of language technologies as well as a demand for gender inclusiveness in society, gender bias in Machine Translation (MT) systems and Large Language Models (LLMs) is an increasingly studied phenomenon with varying research approaches. Due to the nature of how MT systems, and Natural Language Processing (NLP) systems in general, are trained based on large language corpora, these systems exhibit and exacerbate biases present in these corpora (Vanmassenhove, 2024). With biases being an inherently useful characteristic for machine learning systems to generalise on unseen data (Mitchell, 1980), they can lead to unfair and harmful stereotypes, such as when referring to a person using an inaccurate gender (Vanmassenhove, 2024).

Previous research on potential triggers of gender bias in MT is often limited to word-level analyses and does not take context into account. The study presented in this paper is part of a broader research project that aims to fill gaps in current studies by focusing on the influence of sentence context on gender translations (Hackenbuchner et al., forthcoming). MT systems primarily translate into generic masculine (Monti, 2020), however, we hypothesise that context can be a deciding factor for MT systems, as well as for humans, to change the gender inflection in their output. To raise awareness of why this might be happening or of where MT should be adapting gender, the goal of a broader research project, of which this study is a part of, is the creation of a detection system that analyses English source data to detect and mark words and phrases that are considered to influence the gender inflection in target translations. In comparison to what MT systems do, it is important to understand how humans perceive gender of words in isolation, out-of-context, and how those perceptions change for words in context. Humans would be well aided to have additional support when machine translating text to ascertain correct and fair gender translations.

The study presented in this paper compares gender bias in MT systems with inherent gender associations perceived by humans. We comparatively analysed (1) how an MT system translates a person's gender of a word out-of-context (i.e. in iso-

© 2024 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies, pages 31–41
June 27, 2024 ©2024 European Association for Machine Translation

lation) versus in a sentence context, (2) the individual differences between human annotators of gender associations of words out-of-context and in-context, and (3) the comparison of MT with human associations with a focus on gender.

In the following sections, we cover related research (Section 2), how the data was collected (Section 3), the process of participatory data annotation (Section 4), data analysis of both human annotators and MT outputs (Section 5), limitations (Section 7) and a conclusion (Section 6).

2 Related Research

Research shows that humans are strongly influenced by how gender is expressed in languages, by role names and by general stereotypes (Gygax et al., 2008; Lardelli and Gromann, 2023; Misersky et al., 2014). Humans construct their own individual representations of gender, which, if available, they base on grammar in language (e.g., *waitress*) but when lacking grammatical cues, they base on stereotype information (Gygax et al., 2008). In grammatical gender languages, such as German (Stahlberg et al., 2007), people are often referred to in the generic masculine which is intended to be *generic* but is not typically interpreted as such (Gygax et al., 2008). Stereotypicality and bias further come into play when language has no grammatical gender cues, lacks pronouns or other gender referents, and the gender interpretation is up to the reader to define or an MT system to translate. Based on previous research, we analysed to what extent, in the absence of grammatical gender cues, MT systems and humans base their gender translations and associations on role names and on stereotypicality.

Previous studies on gender in monolingual English data focused on gender inherently manifested in word embeddings by measuring the gender-inflection on a word level (Bolukbasi et al., 2016; Caliskan et al., 2022). This has not yet been applied to sentence level nor in the context of MT. Previous research on gender in MT includes the creation of challenge sets to test gender bias in MT outputs, for instance based on professions and adjectives, to balance out the gender of these pre-determined words in machine translations (Stanovsky et al., 2019; Saunders and Byrne, 2020; Troles and Schmid, 2021). Such challenge sets follow the format of, for example, containing a female and a male sentence for *"The choreographer finished her work. / The choreographer finished his work."* to fine-tune and therefore balance an MT system on both gendered versions (Saunders and Byrne, 2020). Moreover, existing work on gender bias in MT has predominantly focused on translations of the binary gender, namely male and female, not taking into account the non-binary community, with the exception of few approaches taken, as by Savoldi et al. (2024), Lardelli and Gromann (2023) and Saunders et al. (2021).

A recent study on the comparison of human and model evaluations of gender bias concluded that, under constrained settings, "model biases reflect human decision-making" and that humans make (sometimes wrong) predictions based on societal and cognitive presupposition (Lior and Stanovsky, 2023). In the study presented here, we analyse to what extent MT gender translations (model choices) coincide with human associations of gender.

Starting with words taken out-of-context whose word embeddings have an inherent gender-inflection as well as a list of sentences featuring these words in varying contexts, this research focuses on how differently or similarly humans and MT associate gender with certain words on an individual out-of-context level and how this gender-inflection is affected by sentence context. In this way, we expand on previous research but broaden the scope by collecting natural contexts that influence gender-inflections in translations rather than artificially constructed test sentences, and by extending the gender categories to include non-binary.

3 Data Description

The data used for this study is in English, a notional gender language (McConnell-Ginet, 2013), where role names generally do not have a gender assigned, e.g., *poet*, apart from kinship relations (*mother, father*) or a few exceptions (*actor, actress*). English data was filtered from monolingual English corpora (StatMT's news-crawl[1], as well as c4 (Raffel et al., 2019) and wiki (Foundation, nd) as made available on HuggingFace[2]). The MT output is analysed in German, a grammatical gender language (McConnell-Ginet, 2013), where gender is specified.

[1] https://data.statmt.org/news-crawl/
[2] https://huggingface.co/

Compiled List of Individual Words

coordinator	**flight attendant**	*musician*	*opponent*	**socialite**	**therapist**°	**lover**°
mechanic	**dancer**°	visitor	colleague	companion	**author**°	**clerk**°
student	**accountant**	**designer**°	**baker**	**writer**°	**consumer**	**poet**
bookkeeper	**counselor**	friend	*guard*	*officer*°	*user*	*supporter*
judge	*fighter*	*dealer*	*soldier*	*player*	*manager*	*contractor*
captain	*farmer*	*maestro*	*boss*	*driver*	*idiot*	*cook*
filmmaker	*admirer*	*follower*	*salesperson*	*buddy*	**winner**°	*construction worker*

Table 1: Individual list of 49 words where those words with a female word embedding gender-inflection are marked in bold, those words with a male gender-inflection are marked in italics, and all others have a neutral gender-inflection. All words with a superscript° appeared more than once in the sentence-context.

3.1 Compiling Gender-Ambiguous Words and Sentences

Our focus lay on compiling a list of words, role names, referring to people where the gender is not specified in English (e.g., *poet*) but, as previous research outlined above has shown, their word embeddings are indeed often gender-inflected, which influences MT systems' choice of gender when translating from a notional gender language to a grammatical gender language.

To further analyse the impact of context, this study is based on the annotation and translation of selected words both on an individual level and in varying sentence contexts, in which gender is ambiguous. In total, 150 words were compiled, where 50 had a female-inflected word embedding, 50 were male-inflected and 50 were considered neutral (having neither a measurable female nor male gender inflection). The initial word list was compiled based on previous studies, outlined above and on gender-inflections in word embeddings. In addition, this list was further augmented by prompting ChatGPT for lists of female-inflected, male-inflected and neutral-inflected words. The ChatGPT prompted lists were compared with previous research and where words did not overlap, they were added.

These words were then translated from English into German using the DeepL API between February and March 2024. The German MT output was noted and the gender inflection of the MT was documented, i.e. whether *poet* was translated as *Dichter* (male) or *Dichterin* (female).

These 150 words were used to automatically filter the monolingual English corpora (newscrawl, c4 and wiki) for sentences containing these words. This resulting data was then manually filtered for gender-ambiguous sentences excluding those sentences that contain a gender cue, a pronoun or name referring to the person in question. In total, 892 gender-ambiguous sentences have been collected.

Similarly, all these gender-ambiguous sentences were translated from English into German using the DeepL API between February and March 2024. The gender of the word in the output sentence was noted, i.e. whether the sentence *Who's the worst poet in Miami?* was translated as *Wer ist der schlechteste Dichter in Miami?* (male) or as *Wer ist die schlechteste Dichterin in Miami?* (female). Of these 892 gender-ambiguous sentences, 75% were translated by DeepL into (the generic) male, only 6.6% were translated into female and the rest were mistranslated or translated as neutral (e.g., *the pilot* as *das Pilotprojekt*).

From all sentences, a sample of 60 sentences was selected for this study. These 60 sentences were selected based on the fact that their German machine translation gender-inflections showed a broader distribution, i.e. some sentences were translated as male, some as female. As a result, 18% of the sentences were translated as female and 82% as generic masculine. The focus lay on the gender-ambiguous role names (e.g., *poet*) in the sentences. There were 49 different role names, of which 19 words had a female word embedding gender-inflection, 25 a male gender-inflection and 4 a neutral gender-inflection. This is depicted in Table 1.

There were only 49 individual role names in the 60 sentences because some occurred in different sentences. The difference in gender perceptions for the same word (role name) in different sentence contexts is an interesting factor, further outlined in section 5 and will be further analysed in

Gender-Inflections by MT

coordinator	*flight attendant*	**musician**	**opponent**	***socialite***	**therapist**[1]	**lover**[1]
mechanic	dancer[1]	**visitor**	colleague	***companion***	author[1]	clerk[1]
student	accountant	designer[1]	baker	writer[1]	consumer	poet
bookkeeper	counselor	**friend**	**guard**	officer[1]	user	supporter
judge	fighter	dealer	soldier	player	manager	contractor
captain	farmer	maestro	boss	driver	idiot	cook
filmmaker	admirer	follower	salesperson	buddy	winner[1]	construction worker
officer[2]	therapist[2]	lover[2]	author[2]	writer[2]	designer[2]	clerk[2]
dancer[2]	winner[2]	winner[3]	author[3]			

Table 2: Gender-inflections by the MT system of words in and out-of-context. All words in italic were female-inflected out-of-context. All words in bold were female-inflected in-context. All other words were male. Superscript 1, 2 and 3 are used to refer to in-context sentence 1, 2 or 3 when there are multiple sentences for a word.

the broader research project, of which this study is a part of (Hackenbuchner et al., forthcoming). An example would be the analysis of the gender association and translation of the word *therapist* in the following two contexts:

- Kensington massage **therapist** jailed for sexually assaulting clients.
- There are 52 weeks in a year, my **therapist** continued matter-of-factly, "I know you can't go on a date every single week, but how many do you think you should be going on?"

We wanted to analyse whether, for the same word, the two different contexts affected the choice of gender. For this example, as depicted in Appendix B, the MT system translated the therapist as female in the first sentence and as male in the second sentence. We want to analyse such differences and whether the choice of gender by human annotators coincides with the gender selected by MT (which in this case it does not as humans annotated the therapist as male in the first context and as female in the second).

3.2 Translation Comparison of Words and Sentences

After the data was compiled, a comparison was drawn between the translation of the individual word out-of-context with the translation of the word in a sentence context. This comparison is depicted in Table 2. We can clearly see that the words were predominantly translated as (generic) masculine both in- and out-of-context. Out-of-context, the MT predominantly translated words as male and a mere three words (*flight attendant, socialite, companion*) were translated as female. In sentence context, the MT translated fewer words as male, with a slightly lesser majority of 82%, and in 18% of the cases, as female. We can therefore see that out-of-context, the MT predominantly translates into the male gender inflection. In a sentence context, the MT still predominantly translates into the male gender inflection but to a lesser extent. This shows that the MT, for certain sentences and role names, is influenced by context. Words that the MT had individually translated as male but in a sentence context as female are: *coordinator, mechanic, musician, visitor, friend, opponent, guard, therapist, lover*. The sentences are depicted in Appendix B.

	out-of-context	in-context
male	.95	.82
female	.05	.18

Table 3: Label distribution gender-associations of MT translations in-context and out-of-context.

The MT's translation behaviour of gender is later compared to human gender associations of the same words both out-of-context and in a sentence context.

4 Annotation & Guidelines

4.1 Annotators

Unlike regular annotation tasks where correct word categories are requested to be annotated, the annotations for this research are highly subjective and individual as there was, e.g., no pre-defined part of speech (POS) that had to be annotated. There were no *right* or *wrong* annotations. To

cover a variety of viewpoints, we tried to recruit a diverse set of annotators. A total of 22 annotators were recruited who are highly proficient in English and vary in native language, origin and gender, as detailed in Appendix A. This allowed for a balanced gender representation, minimising the possibility for one certain gender to highly influence the annotations.

All annotators were duly informed of the study and their role as annotators, and signed the informed consent form, allowing their annotations to be analysed within the context of this study.

4.2 The Annotation Task

The annotation task consisted of two parts. In the first annotation step, the annotators were asked to annotate the associated gender for words in isolation, for each of the 49 individual words (i.e. role names) in an Excel table. They could choose a gender from a pre-defined list (female/male/non-binary) and had the option to select N/A if they really did not associate any gender with the word. For example, annotators had to indicate their gender association for the role name *poet* without any context.

In the second annotation step, given that the aim is to understand how and to what extent context influences the human perception of gender, they annotated the same words presented in a sentence on the annotation platform *Label Studio*[3]. The annotators had to equally indicate from the pre-defined list (female/male/non-binary) which gender they most strongly associated with the word (role name, e.g., *poet*), but this time in a (gender ambiguous) sentence context, e.g., *Who's the worst poet in Miami?*

5 Analysis

In this paper, we focus on a quantitative analysis of selected aspects of the annotations we obtained. We comparatively analysed (1) how an MT system translates a person's gender out-of-context versus in a sentence context, (2) the individual differences between human annotators of gender associations of words out-of-context and in-context, and (3) the comparison of MT with human associations with a focus on gender.

[3]Label Studio `https://labelstud.io/`

	Human		MT	
	OOC	IC	OOC	IC
male	.58	.58	.96	.82
female	.19	.28	.04	.18
Non-binary	.03	.01	/	/
N/A	.19	.13	/	/

Table 4: Label distributions for gender associated with words out-of-context (OOC) and in-context (IC). The label distribution is shown in percentages and was averaged for the human annotators.

5.1 Words Out-of-Context vs. In-Context

As shown in Table 4, human annotators associated 58% of the words both out-of- and in-context with the male gender. Furthermore, in the out-of-context scenario, the annotators indicated the words as female for 19% of the cases and did not assign a specific gender (i.e., annotated N/A) also in 19% of cases. When moving to the in-context scenario, the percentage of male-associated words remains the same, but the number of female words increased by 9% and the number of non-binary associations drops minimally (from 3% to 1%). However, the 58% male annotations did not refer to the same words out-of- and in-context. And the N/A labels from out-of-context did not simply change to become female. There was an overall change for which gender was associated with which word, as further explained in the analysis. Interesting to note here is that 19% of the words would not evoke a gender association for human annotators without context, however, annotators are less likely to use the N/A label in-context.

Compared to the human annotations, the MT system shows a clear bias for the male gender, where out-of-context, 96% of words were translated as male. As the MT system does not translate words into gender-inclusive non-binary or 'N/A' genders, the remaining 4% was labeled as female. In-context, the MT system shows less bias, with only 82% of words being translated as male and 18% as female. Overall, the annotation and translation distributions indicate that both MT and human annotators had a tendency towards the male gender, but this bias is much more predominant in the MT system and seems to drop in context.

Clearly depicted in Figure 1, all human annotators often changed the gender annotation for each word from out-of-context to in-context. On average, annotators chose a different associated gender when annotating in-context for 44% (27/60) of

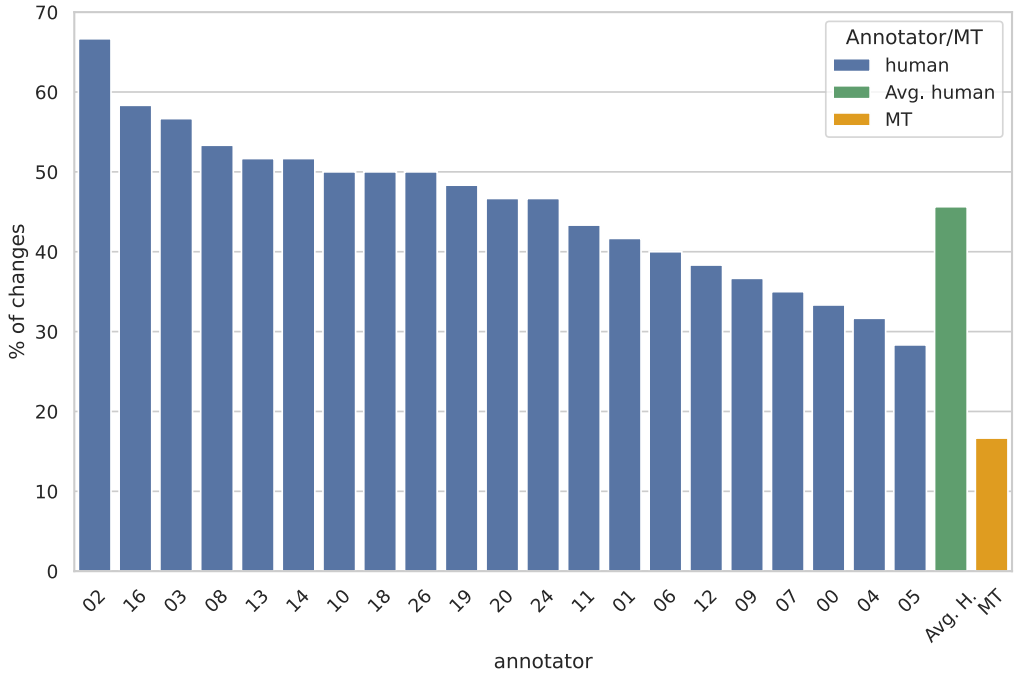

Figure 1: Gender changes for words from out-of-context to in-context for (individual and average) annotators and MT

words. In comparison, the MT system only translated a word in context with a different gender for 17% (10/60) of the words. This shows that the MT system predominantly translated each word whether in- or out-of-context in its male generic form, whereas the human association of gender was highly subjective to sentence context. The annotator's association with gender was less consistent for words out-of-context but much more decisive, and also in higher agreement as discussed below, when words were presented in context.

5.2 Agreement

For all human annotators, we calculated inter-annotator agreement scores with Fleiss' Kappa (Fleiss, 1971) and Krippendorff's alpha scores (Krippendorff, 2011), which resulted in fair agreement score of 34% and 35% respectively. However, we focus our analysis on an average of the pairwise Cohen's Kappa (Cohen, 1960), to enable averaged pairwise comparison of agreement between the annotators on the one hand, and the annotators versus MT on the other. Since MT-human agreement has to be calculated between each human annotator versus MT and is then averaged across annotators, it made sense to compare agreements this way.

In Table 5, we present the scores for inter-human agreement (human), calculated pairwise, and MT-human agreement on the gender of words in- and out-of-context. These agreement scores, the pairwise Cohen's kappa, indicate fair agreement for in-context labeling and slight agreement for out-of-context labeling both for inter-human and MT-human agreement. Although there are no right or wrong labels, there is a noteworthy increase in agreement for in-context, 18% for inter-human and 15% for MT-human agreement. Notably, inter-human is consistently higher than the agreement between MT and human annotations, despite highly varying annotator profiles. For in-context labeling, inter-human agreement results in an 8% higher agreement than MT-human annotations, and for out-of-context labeling, this results in a 5% higher agreement.

Figure 2 and Figure 3 show the difference between human annotations in and out-of-context, by looking at the percentage of annotators that marked words with the same gender. The x-axes depict the percentage of annotators that agreed on the gender of a word, with a higher percentage, meaning a larger majority. The y-axes depict the number of words that have been agreed on.

Figure 2 shows a relatively equal balance between words that have a small majority (on the left-hand side of the figure) and a strong majority (on

		OOC	IC
Human	avg	.18	.36
	max	.50	.96
	min	-.13	.08
	med	.16	.37
MT-Human	avg	.13	.28
	max	.37	.51
	min	-.08	.04
	med	.11	.32

Table 5: Out-of-context (OOC) and in-context (IC) Pairwise Cohen's kappa scores for inter-human and MT-human agreement (including average, minimum, maximum and median for each pair).

the right-hand side of the figure). When comparing these results to Figure 3, which displays the same for in-context labels, this shows us that there are a lot more words with strong agreement (on the right-hand side of the figure).

Table 6 displays the top 10 most agreed-upon words both out-of- and in-context. This shows us that words like *construction worker*, *judge* and *opponent* were annotated with high agreement both in and out-of-context, meaning that annotators had a clearer associated gender for these role names both when seeing the words on their own or when reading the word in a sentence context.

In Table 7, on the other hand, we display the top 10 words with the least agreement in- and out-of-context. These results suggest that words like *baker*, *colleague* and *visitor* were highly ambiguous. Notably, although words like *accountant* and *fighter* have a clear out-of-context associated gender, their in-context annotations have low agreement. The word *fighter* is an interesting example to look at more closely as out-of-context, a decisive 91% of annotators marked the word as male, whereas in-context only 38% of annotators marked the word as male, and the others as female, N/A or non-binary. The sentence this word occurred in was: *It's not the end of the world just yet - I like to think of myself as a fighter and I will keep fighting right until my last run.* This sentence strongly appeals to the individuality of the reader.

We can clearly see here that the human gender association for words is highly dependent on the context that these words are seen in. This phenomenon is much more present in humans than can be seen MT outputs, which predominantly defaults to male.

out-of-context	in-context
construction worker	construction worker
judge	judge
opponent	opponent
dealer	dealer
farmer	buddy
guard	filmmaker
fighter	maestro
captain	boss
accountant	manager
mechanic	student

Table 6: Top 10 most agreed-upon words (over 90% of the votes). With the exception of opponent in-context, all words had *male* as their majority label both in and out-of-context.

out-of-context	in-context
baker	baker
colleague	colleague
visitor	visitor
consumer	salesperson
clerk	cook
follower	fighter
friend	user
coordinator	lover
musician	accountant
designer	dancer

Table 7: Top 10 least agreed-upon words (with majority votes between 33 and 50% of all annotators).

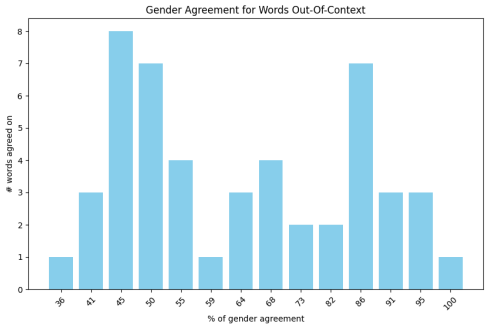

Figure 2: Comparison of human annotators' choice of gender for words out-of-context

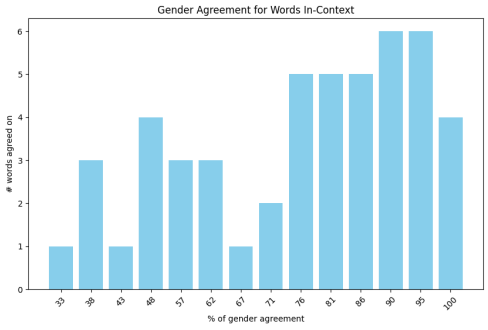

Figure 3: Comparison of human annotators' choice of gender for words in-context

6 Conclusion

In this paper, we analysed to what extent an MT system translates and humans associate gen-

der with role names both out-of-context and in-context, with no grammatical gender cues in (the source) language.

We note that the MT system is very rigid in its gender translations of role names, primarily translating into generic masculine, particularly for words out-of-context, and seldom changing a role name's gender in certain sentence contexts. Human associations of gender are much more varied, both for words out-of-context and in-context. We particularly see that all annotators have been greatly influenced by the sentence contexts, annotating role names with a gender they associated with that specific context.

The results from this study show that, in comparison to MT, humans have a much more varied understanding of gender and are highly influenced by context. This underlines the diversity and complexity of human associations and gender roles in society and therefore critically highlights the problematic generic masculine translation outputs by MT systems.

The study conducted shows the necessity for continued research to further understand what these diverse human associations for gender in context mean for MT translations. On the one hand, we criticise the generic masculine output of MT systems but on the other hand we see that, for some sentences, the MT does change the gender of role names based on context. This pattern of when and why MT changes gender based on context, and to what extent this relates to human gender associations, will be further studied in the broader research project.

7 Limitations and Future Work

A limitation of this exploratory study is that it is only done in a single language direction. Four annotators explicitly noted their mother tongue's influence on their choice of gender annotation for certain words, particularly out-of-context. Many of the annotator's native languages are grammatical gender languages, where role names have a gender assigned. The vast majority of words in grammatical gender languages are traditionally highly influenced by culture and predominantly referred to in the generic masculine.

Two aspects that have been excluded from the analysis of this study but that annotators were asked to annotate were (1) how strongly they associated a word with a specific gender (on a scale of 1-3) and (2) which specific words in the sentence context influenced their choice of gender for the role name in that specific context. Our future research will analyse these aspects and particularly focus on the specific context that influences gender, and relate it to MT. In comparison to analysing influences for human gender associations for words in context, our overarching question that we will focus on is: *What are the triggers that make MT systems change a role name's gender when translating in a specific sentence context?*

8 Acknowledgements

This study is part of a broader project, a strategic basic PhD research (1SH5V24N) fully funded by The Research Foundation – Flanders (FWO) for the timespan of four years, from 01.11.2023 until 31.10.2027, and hosted within the Language and Translation Technology Team (LT3) at Ghent University. This research, including the information letter, study guidelines and informed consent form, has been ethically approved by the ethics committee at the Faculty of Arts and Philosophy at Ghent University. The authors would like to thank all annotators for their voluntary and patient annotations, without whom this study could not have been done, and Colin Swaelens and Jasper Degraeuwe for early feedback on the (analysis) of this work.

9 Bias Statement

In this paper, we study machine translations of and human associations with gender for role names out-of-context and in-context. The human annotators base their gender associations on language or (stereotypical) cultural and societal knowledge. MT systems predominantly and by default translate role names into the generic masculine, establishing a skewed image of gender in society, thus creating representational harm. Our assumptions are that humans may be stereotyping their assumptions but are nevertheless much more diverse in their overall gender associations for role names, representing a more colourful society, whereas MT systems default to generic masculine but break this pattern for specific and highly stereotypical sentence contexts.

References

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

Caliskan, Aylin, Pimparkar P. Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *AAAI/ACM Conference on AI, Ethics, and Society*, page 156–170.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Foundation, Wikimedia. n.d. Wikimedia downloads.

Gygax, Pascal, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. In *LANGUAGE AND COGNITIVE PROCESSES*, volume 23:3, pages 464–485.

Hackenbuchner, Janiça, Arda Tezcan, and Joke Daems. forthcoming. Automatic detection of (potential) factors in the source text leading to gender bias in machine translation.

Krippendorff, Klaus. 2011. Computing krippendorff's alpha-reliability. *Computing*, 1:25–2011.

Lardelli, Manuel and Dagmar Gromann. 2023. Gender-fair (machine) translation. In *New Trends in Translation Technology (NeTTT)*, page 166–177, Rhodes Island, Greece.

Lior, Gili and Gabriel Stanovsky. 2023. Comparing humans and models on a similar scale: Towards cognitive gender bias evaluation in coreference resolution. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, page 755–762.

McConnell-Ginet, Sally. 2013. Gender and its relation to sex: The myth of 'natural' gender. *In G. G. Corbett (Ed.), The expression of gender. DE GRUYTER.*, page 3–38.

Misersky, Julia, Pascal M. Gygax, Paolo Canal, Ute Gabriel, Alan Garnham, Friederike Braun, Tania Chiarini, Kjellrun Englund, Adriana Hanulikova, Anton Öttl, Jana Valdrova, Lisa Von Stockhausen, and Sabine Sczesny. 2014. Norms on the gender perception of role nouns in czech, english, french, german, italian, norwegian, and slovak. *Behav Res*, 46:841–871.

Mitchell, Tom M. 1980. The need for biases in learning generalizations.

Monti, Johanna. 2020. *Gender issues in machine translation: An unsolved problem?* The Routledge Handbook of Translation, Feminism and Gender.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7724–7736. Association for Computational Linguistics.

Saunders, Danielle, Rosie Sallis, and Bill Byrne. 2021. 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*, page 35–43. Association for Computational Linguistics.

Savoldi, Beatrice, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. A prompt response to the demand for automatic gender-neutral translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, page 256–267. Association for Computational Linguistics.

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of sexes in the language. In *Social Communication, Frontiers of Social Psychology*, page 163–187, Psychology Press, New York, NY.

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*, page 1679–1684.

Troles, Jonas-Dario and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation. impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, page 531–541.

Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. *arXiv e-prints*, page 1–24.

A Appendix: Annotators

Annotators			
Gender	**Annotators**	**Country of origin**	**Mother Tongue**
Female	10 annotators (1 explicitly identifying as trans)	Germany, Belgium, UK, France, The Netherlands, Brazil, Russia	German, Flemish, English, French, Dutch, Portuguese, Russian
Male	10 annotators	Belgium, Turkey, India, UK	Flemish, French, Turkish, Hindi, English
Non-binary	2 annotators	Bulgaria, Belgium	Bulgarian, Flemish

Table 8: List of number of annotators per gender, country of origin and mother tongue.

B Appendix: Examples

MT Translation: Gender Change	
word	**sentence**
friend	After a friend suggested she try it, Ann said, “Sure!”
visitor	A health visitor also contacted RBH to raise the issue in July 2020 and an inspection that month found mould in the kitchen, bathroom and a bedroom cupboard needed treatment.
therapist	Kensington massage therapist jailed for sexually assaulting clients.
musician	In an Instagram video posted last month, the “All Too Well” musician can be seen collaborating with producer Jack Antonoff on the piano.
coordinator	One day, she visited a friend who worked as an assistant production coordinator on a set, and she was intrigued by the location department.
mechanic	It’s important if we want to see a future in which a boy could become a midwife or a girl could become a mechanic.
opponent	On Thursday evening, finally, she stepped out onto the court against a top 10 opponent for just the second time of her life.
guard	The reserve guard stepped up in the absence of fellow rookie guard Jordan Nixon, who injured her hamstring during warmups.
lover	SINGER Matt Goss smooches with his new lover after a dinner date.

Table 9: The MT system translated all words out-of-context (individually) as male, but then as female in the respective sentence context.

Example Comparison

MT translations vs. human annotations

word out-of-context	**sentence in-context**
cook	I always call myself a cook.
MT: male	MT: male
Ann: male	Ann: N/A
poet	Who's the worst poet in Miami?
MT: male	MT: male
Ann: female	Ann: male
colleague	Like me, Imogen gets her "dream job" and thinks her life is finally starting - but her confidence and happiness is constantly threatened and undermined by a toxic colleague.
MT: male	MT: male
Ann: N/A	Ann: female
officer	I am also the the chief executive officer of Global Women Network, a United Kingdom-based Non-governmental Organisation with roots in Nigeria.
MT: male	MT: male
Ann: N/A	Ann: female
follower	I cant even deal with this, one follower wrote alongside two fire emojis, while another wrote: "Love the hair x."
MT: male	MT: male
Ann: N/A	Ann: female

Table 10: A comparison of a sample of words where the MT system translated the gender of the words differently than the gender association as marked by the human annotators.

Comparison: MT gender translations for words in different contexts

word out-of-context	**sentence in-context**	**MT**
therapist	Kensington massage therapist jailed for sexually assaulting clients.	female
	There are 52 weeks in a year, my therapist continued matter-of-factly, "I know you can't go on a date every single week, but how many do you think you should be going on?"	male
clerk	A hotel clerk was caught on video calling a black customer a monkey.	male
	The Newark, New Jersey, native was born in 1954 and adopted at age six months out of an orphanage by a township clerk and an auto parts owner.	male
lover	SINGER Matt Goss smooches with his new lover after a dinner date.	female
	Casual sends a check-in to your friend or lover to see how they're doing or what they're up to.	male

Table 11: Comparison of MT translations of individual words all translated as male out-of-context but then depending on the sentence context, translated the word as either male or female.