Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT

Maja Popović[1], **Ekaterina Lapshinova-Koltunski**[2]
[1] ADAPT Centre, School of Computing, Dublin City University, Ireland
maja.popovic@adaptcentre.ie
[2] Language and Information Sciences, University of Hildesheim, Germany
lapshinovakoltun@uni-hildesheim.de

Abstract

This paper presents an analysis of first-person gender in five different translation variants of Amazon product reviews: those produced by professional translators, by translation students, with different machine translation (MT) systems and with ChatGPT. The analysis revealed that the majority of the reviews were translated into the masculine first-person gender both by humans and by machines. Further inspection revealed that the choice of the gender in a translation is not related to the actual gender of the translator. Finally, the analysis of different products showed that there are certain bias tendencies, because the distribution of genders notably differ for different products.

1 Introduction

In this paper, we focus on the distribution of gendered words in human and machine translations of product reviews from English into Croatian and Russian. In contrast to English, both Croatian and Russian have gender marking not only on pronouns, but also on nouns, adjectives, verbs, determiners and numbers. The gender implicit in the English source needs to be specified in the target. This may result in translation errors, mismatches and inconsistencies, as well as gender bias in train and test data.

In reviews, the texts are written in the first person form as illustrated in example (1). While translating from English into Croatian or Russian, the gender of the adjectives and verb past and passive participles should be specified: обожал (masculine) vs. обожала (feminine).

(1) a. *I **loved** using this makeup*
b. я обожал(а) пользоваться этой косметикой.

The decision for either feminine or masculine form is required not only in case of machine translation. Human translators need to specify this form, too. If no information on the text author is available and no specific instructions are given for translators, this may result in inconsistencies and individual decisions by human translators.

Therefore, we decide to look into this variation analysing and comparing translations produced by two different groups of translators (professional and student) as well as with two machine translation systems and ChatGPT large language model in the two language pairs at hand.

Our work is similar to the studies of gender bias in machine translation (MT). However, our primary focus is not on reducing the gender bias, but rather on regularities in human and machine translation data that may follow in the emerging gender bias in the data.

Gender bias (preference or toward one gender over the other) exists in training data, pre-trained models such as word embeddings and also algorithms themselves (Zhao et al., 2018a; Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al.2018), so that a machine translation system containing bias can produce gender biased predictions. Although this issue belong to active research topics, detection and evaluation of gender bias in machine translation systems have not been thoroughly investigated yet.

Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies, pages 22–30
June 27, 2024 ©2024 European Association for Machine Translation

In our analysis, we focus on the following research questions:

RQ1: What is the distribution of first person gender in different translations?

RQ2: Is choice of the gender in human translations related to the gender of the translator?

RQ3: Is choice of the gender related to the topic/product?

The remainder of the paper is organised as follows: Section 2 provides and overview of related studies. The data is described in Section 3. The analyses and the results are presented in Sections 4 and 5, and conclusions in Section 6.

2 Related Work

Our work is similar to the studies of gender bias in natural language processing and specifically in machine translation. However, our descriptive aims differ from those existing in most studies. Some studies do describe bias in the data. For instance, Zhao et al. (2018) addressed gender bias in word embeddings and Sun et al. (2019) provides an overview of existing biases.

Some works focus on the creation of challenge or test suites. Stanovsky et al. (2019) presented a challenge set and evaluation protocol for the analysis of gender bias in MT. Their automatic gender bias evaluation method was developed for eight target languages (including Russian) with grammatical gender. They tested six MT systems themselves, including also Google. Vanmassenhove and Monti (2021) presented an English–Italian challenge set focusing on the resolution of natural gender phenomena by providing word-level gender tags on the English source side and multiple gender alternative translations, where needed, on the Italian target side. The data analysed in our study can potentially serve as a test suite as well.

In our work, we also address bias dependence on topic or product. Similarly, bias variation was addressed in (Zhao et al., 2017) who found that on the one hand, data sets for specific tasks (e.g. cooking) contain significant gender bias and, on the other hand, models trained on these datasets further amplify existing bias.

Some works showed that bias can be measured, see e.g. (Cho et al., 2019) who proposed a measure called 'translation gender bias index' (TGBI).

We analyse both human and machine-translated texts. The latter were analysed in several other works. For instance, Saunders et al. (2020) explored the potential of gender-inflection controlled translation in case the gender is identifiable either from a human reference or when it can be automatically gender-tagged. The authors found out that simple existing approaches could overgeneralize a gender-feature to multiple entities in a sentence, and suggested effective alternatives in the form of tagged co-reference adaptation data. They also proposed an extension to assess translations of gender-neutral entities from English given a corresponding linguistic convention in the target language. In another study, the authors analyse and evaluate gender bias comparing bias measurements across multiple metrics for pre-trained embeddings and the ones learned by their own machine translation model (Ramesh et al., 2021). A summary of various analyses of gender bias in machine translation was presented by Savoldi et al. (2021). The authors also discussed the mitigating strategies proposed in various studies. Měchura (2022) presented a taxonomy of phenomena which caused bias in machine translation. Interestingly, it included not only gender bias on people being male and female, but also number and formality bias (singular *you* vs. plural *you* as well as informal *you* vs. formal *you*).

In our study, we focus not only on the machine translations but also on the human ones and compare them across each other. We also distinguish two groups of translators according to their experience: professionals and students. In this way, we also consider the bias introduced by the human translators, which has not been thoroughly analysed so far. Human bias has been addressed in a few studies only. For instance, Hada et al. (2023) investigated the generation and consequent receptivity of manual annotators to bias of varying degrees. The authors created the first dataset of GPT-generated English text with normative ratings of gender bias. The variation of themes of gender biases in the observed ranking was then systematically analysed. The authors showed that identity-attack was most closely related to gender bias. They also showed the performance of existing automated models trained on related concepts on their dataset.

We believe that our work has an added value to the studies existing in the area of machine translation and natural language processing, as it adds to the awareness (Daems and Hackenbuchner, 2022) of the bias existing in the translation data, both in human and machine translations.

3 Data

For our analysis, we use the publicly available corpus DiHuTra[1] (Lapshinova-Koltunski et al., 2022). The corpus contains 196 English Amazon product reviews (14 reviews in each of 14 different product categories) and their human and machine translations into three languages, Croatian, Russian and Finnish. Since the Finnish language does not have grammatical gender in any word category, not even in personal pronouns, only Croatian and Russian were included in our analysis. The number of running words and vocabulary size for the source text and for each of the translations can be seen in Table 1.

In most of the reviews, the gender of the writer is not known, and not specified by any information in the English source. In two reviews only, the text indicates that the writer was a female. The human translations were produced by two groups of translators: several professional translators and several students. The translators were only instructed to keep the given segmentation and not to use any MT system. They did not receive any guidelines about how to treat the gender in the target language. Therefore, the corpus is appropriate to explore the subjectivity.

The machine translations in the corpus were generated by different MT systems. Croatian MT outputs are the two best ranked outputs by human evaluation from the WMT 2022 shared task[2] (Kocmi et al., 2022). Russian MT outputs were generated using Google Translate[3] and DeepL Translator[4]. ChatGPT [5] translations for all target languages were generated using the publicly available GPT 3.5 version. Since human translators were given only simple instructions, a similar approach was used for ChatGPT as well, namely a simple prompt "translate into Croatian/Russian".

text	running words	vocabulary
en source	15,236	3,155
hr prof	13,981	4,359
hr stud	13,931	4,446
hr mt1	13,467	4,309
hr mt2	13,465	4,247
hr gpt3.5	14,170	4,265
ru prof	14,217	4,414
ru stud	14,247	4,523
ru mt1	14,472	4,348
ru mt2	14,635	4,391
ru gpt3.5	15,015	4,397

Table 1: Corpus statistics.

4 Analysis of first-person gender

As mentioned in Section 3, the gender of the writer is not known, and with the exception of two reviews, not specified by any information in the English source. Therefore, the choice of the first person gender in the translation is totally free. The analysis of first-person gender was carried out manually, finding that the majority of the first-person gendered words are verb past participles, followed by adjectives and verb passive participles. This analysis revealed that some student translations and many ChatGPT translations contain the inclusive gender forms. These words were not properly recognised by the part-of-speech tagger and were tagged as masculine nouns.

For each review, a gender label was assigned according to the gendered words it contained. If all first-person gendered words within a review have the same gender (feminine, masculine or inclusive), the review was assigned this gender label. If there was a mixture of first-person genders, the review got the label "mixed".

An example of gender labels for Croatian and Russian translations[6] is shown in Table 2. The English source text contains two words referring to the first person (one verb past participle *received* and one adjective *upset*) which should be gendered in the translations. The first translation is labelled as feminine since both relevant words are in the feminine form. Analogously, the second translation is labelled as masculine, and the third one as inclusive. The fourth and fifth translation are labelled as mixed, because the two relevant words have different genders.

[1] http://hdl.handle.net/21.11119/0000-000A-1BA9-A

[2] https://www.statmt.org/wmt22/translation-task.html

[3] https://translate.google.com/, accessed in February 2023

[4] https://www.deepl.com/en/translator, accessed in August 2023

[5] https://chat.openai.com/, accessed in November 2023

[6] Sentences are shown instead of entire reviews for the sake of space and clarity.

	en	this is fake MAC, i just **received** mine and super **upset** to find out it isnt real MAc.
fem.	hr	Ovo je fejk MAC, upravo sam **dobila** svoj i jako sam **ljuta** što nije pravi MAC.
	ru	Это подделка МАС, я только что получила свою косметику и ужасно расстроена, потому что это не настоящая косметика МАС!
masc.	hr	Ovo je fejk MAC, upravo sam **dobio** svoj i jako sam **ljut** što nije pravi MAC.
	ru	Это подделка МАС, я только что получил свою косметику и ужасно расстроен, потому что это не настоящая косметика МАС!
incl.	hr	Ovo je fejk MAC, upravo sam **dobio/la** svoj i jako sam **ljut/a** što nije pravi MAC.
	ru	Это подделка МАС, я только что получил(а) свою косметику и ужасно расстроен(а), потому что это не настоящая косметика МАС!
mixed	hr	Ovo je fejk MAC, upravo sam **dobila** svoj i jako sam **ljut** što nije pravi MAC.
	ru	Это подделка МАС, я только что получил свою косметику и ужасно расстроена, потому что это не настоящая косметика МАС!
mixed	hr	Ovo je fejk MAC, upravo sam **dobio/la** svoj i jako sam **ljut** što nije pravi MAC.
	ru	Это подделка МАС, я только что получила свою косметику и ужасно расстроен(а), потому что это не настоящая косметика МАС!

Table 2: Example of gender labels according to first-person gendered words.

It should be noted that there are still no non-binary forms in the analysed target languages. Neuter gender is never used for people, only for objects, and would sound awkward, and even possibly offensive. Also, while in some texts it is possible to avoid the gender and generate a "neutral" translation, it is very difficult to avoid all adjectives and past participles. The only way for a proper inclusion is to use the "inclusive" form, comprising both gender variants in a word.

5 Results

5.1 Distribution of first-person gender

First of all, it was found out that about two thirds of the translated reviews (slightly more in Croatian than in Russian) are found to contain indicators of the writer's gender. The rest does not contain any indicator of the writer's gender and was not taken into account in the analysis.

The gender distribution of the gendered reviews is shown in Figure 1: feminine reviews are presented in red, masculine in blue, inclusive in orange, and mixed in grey. For each gender category, lighter nuance represents Croatian and darker nuance Russian.

It can be seen that masculine first-person gender is dominant for both languages and all translation variants, both human and machine-generated. The difference between the percentage of masculine and feminine reviews is smaller in human translations, but still notable. For both target languages, there are slightly less feminine reviews in student translations than in professional ones.

As for machine-generated translations, distributions are slightly different for different systems and target languages, but the overall tendency is the same: the vast majority of the reviews are written in masculine. The most extreme are Russian ChatGPT translations with only 0.5% of all gendered reviews being written in the feminine gender.

The inclusive reviews are mainly found in Croatian ChatGPT translations, although there are a few Russian ones, too. One Croatian student also opted to use the inclusive form. The rest of translations (MT outputs, professional translations, Russian student translations) do not contain any inclusive reviews.

Mixed reviews were found in all machine-generated translations, more in Croatian than in Russian. The smallest amount of mixed reviews was found in the Russian ChatGPT output (0.5%, the same as feminine reviews). It should be noted that in ChatGPT translations there was no mixing of masculine and feminine forms as in MT outputs, but of inclusive and masculine or feminine forms.

Overall, even human translations "prefer" to write in masculine gender, and the "preference" is even stronger in MT systems and ChatGPT, especially Russian ChatGPT.

As for the two reviews with indicators of a female author, all human translators used the feminine gender, while most MT translations had mixed gender. As for ChatGPT, both Russian translations were feminine, while one Croatian

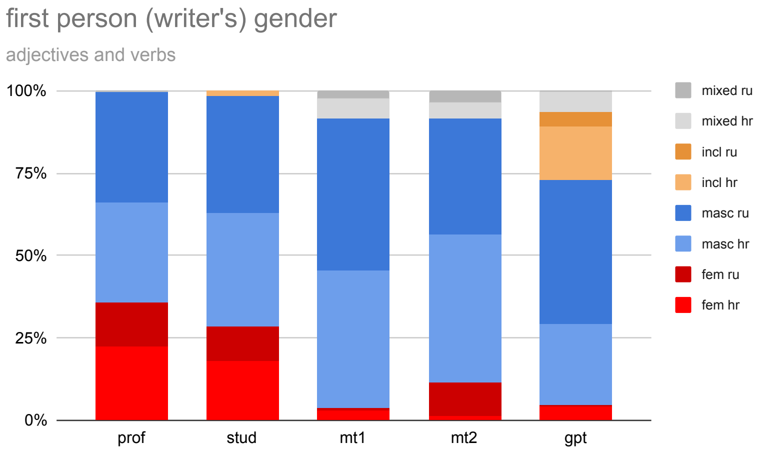

Figure 1: Distribution of first-person genders in different translations: red = feminine, blue = masculine, orange = inclusive, grey = mixed; darker shade = Russian, lighter shade = Croatian.

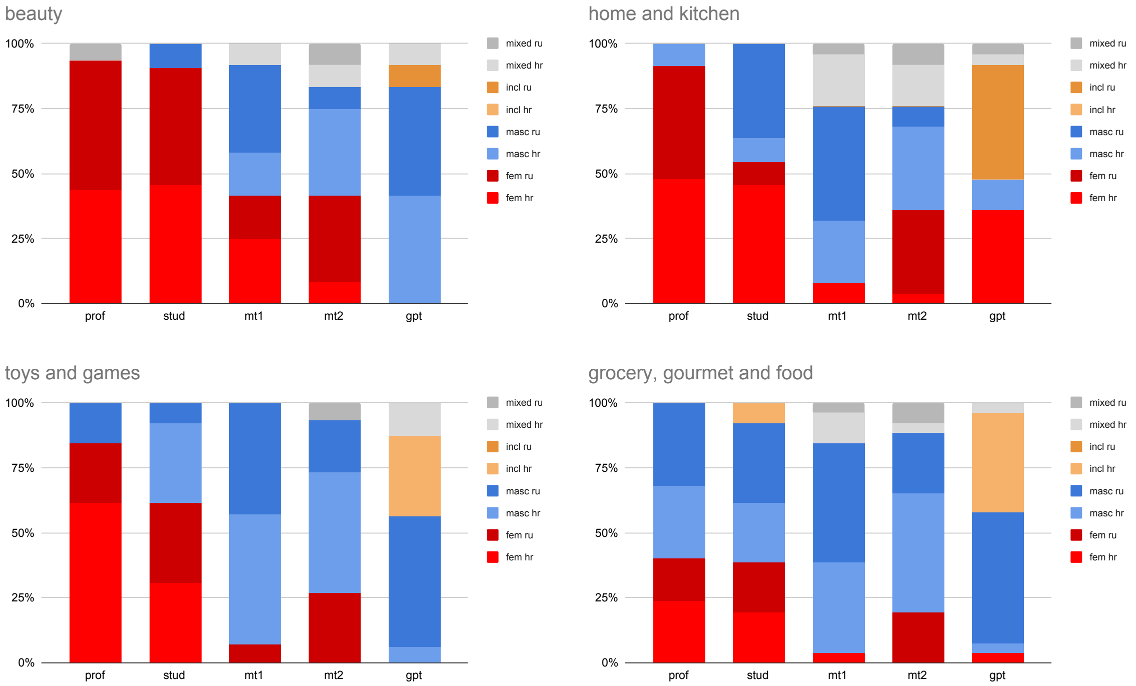

Figure 2: Distribution of first person gender for different products, part 1.

translation was masculine and one mixed, containing feminine and inclusive forms.

5.2 Translators' gender

In order to analyse the preference for masculine gender in human translations, we looked into the meta-data which provide the actual gender of the translator for each review. Overall, there were more female than male translators, and consequently more reviews translated by female translators, which already indicated that the translators do not necessarily use their own gender in translations.

Table 3 presents the percentage of translated reviews written in particular gender for each group of the translators. For example, the first row should be interpreted in the following way: of all Croatian professional translations, 50 reviews were translated by a male translator. Of these reviews, 44% were written in masculine gender (meaning that the translator kept his own gender) and 34% in feminine gender (meaning that the translator changed his own gender).

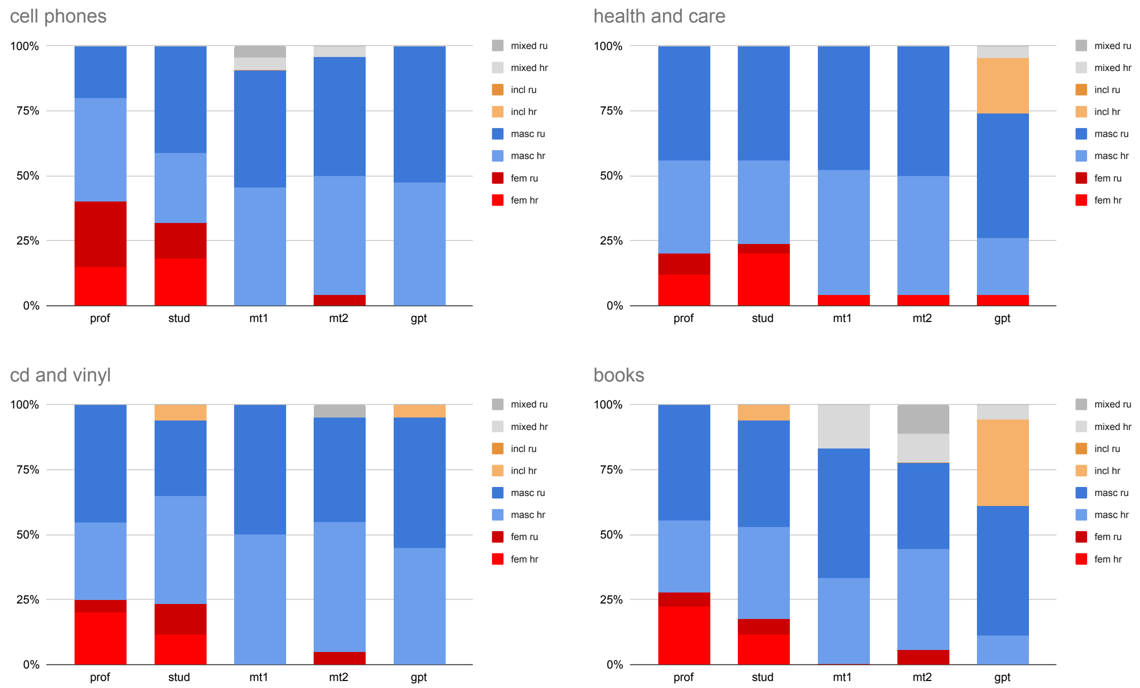

Figure 3: Distribution of first person gender for different products, part 2.

translator				translations		
			number of	gender		
group	lang.	gender	reviews	masc.	fem.	incl.
prof.	hr	m	50	44.0	34.0	0
		f	146	39.7	28.8	0
	ru	m	20	40.0	20.0	0
		f	176	45.4	18.2	0
stud.	hr	m	51	54.9	17.6	0
		f	145	40.7	25.5	2.8
	ru	m	0	0	0	0
		f	196	46.4	39.8	0

Table 3: Translators' reported gender and percentage of gender chosen for the translations.

In total, the numbers in Table 3 shows that translators choose masculine gender more often, regardless of their actual gender.

5.3 Tendencies for different products

Since the previous analysis showed that both female and male translators "prefer" the masculine writer's gender, we decided to look into the gender distributions for different products.

We have to point out that there are only 14 reviews for each of the 14 products, and not all of them are gendered, so that it is not possible to draw any hard conclusions from this analysis, but certain tendencies can definitively be observed. Figures 2, 3 and 4 show the distributions for each of the products, ordered by the proportion of feminine reviews in human translations.

The main observation is that there are clear differences in gender distributions for certain products (namely bias), and that the product-related differences are even more notable in human translations.

Regarding **human translations**, almost all "beauty" reviews are feminine, followed by "home and kitchen" and "toys and games" (Figure 2, while there are only a few feminine translations of "sports and outdoors", "movies and TV" as well as "patio, lawn and garden", and there is no single feminine review for "musical instruments" (Figure 4).

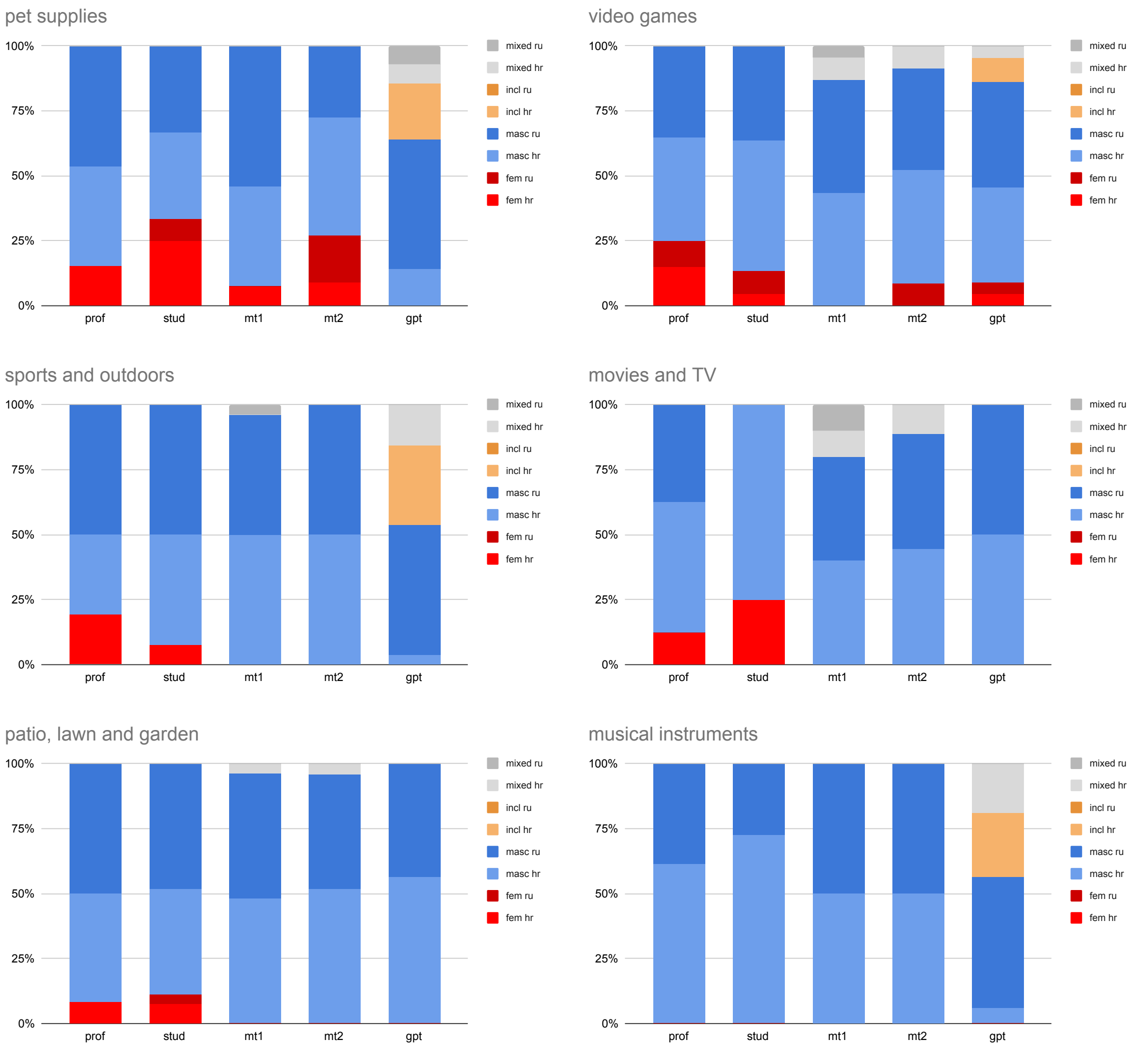

Figure 4: Distribution of first person gender for different products, part 3.

As for **machine-generated translations**, there are less feminine reviews than in human translations for each of the products. For example, for the category "beauty", gender in machine translations is balanced, while the predominant gender in human translations is feminine. For the 'middle-range' products such as "cell phones" or "books", there are about 25-35% of feminine reviews in human translations, but very few or none in machine-generated ones. Finally, for "patio, lawn and garden" there are some feminine reviews in human translations but none in machine-generated ones, and for "musical instruments" there is no single feminine review at all. It should be noted, however, that there are inclusive Croatian ChatGPT outputs.

Another interesting observation is that Russian ChatGPT inclusive reviews are only found in the predominantly "feminine" products, namely "beauty" and "home and kitchen", while there no clear product-related tendencies could be observed for the Croatian ChatGPT inclusive translations.

6 Conclusions

This work presents results of analysis of first-person gender in Russian and Croatian translations of English user reviews. We addressed three research questions concerning the distribution of the first person gender, the relation between the choice of the gender for translation and the real gender of the translator, as well as a tendency towards a product or product group bias. We group the findings according to the three research questions addressed:

RQ1: What is the distribution of first person gender in different translations?

We could observe that in all translations, the predominant gender is masculine. Interestingly, the difference is much stronger in machine-translated texts. This indicates the intensification of the gender bias existing in human translations.

RQ2: Is choice of the gender in human translations related to the gender of the translator?

Our data shows that it is not the case. All translators in our dataset at hand, regardless of their gender, translated more reviews into the masculine form. It is interesting to note that we also observed the cases of a male translator using feminine forms.

RQ3: Is choice of the gender related to the topic/product?

Although the data set is too small to draw hard conclusions, we noticed a clear tendency, especially in human translations. Similar tendencies are observed in machine-generated output, although the overall trend is notably less feminine translations in each of the product categories.

The reported findings also open several directions for future work. Apart from including more target languages from different families, as well as more domains and topics, more language models should be included, also the outputs using different prompts such as giving particular instructions regarding gender specification.

Furthermore, a test suite specifically designed for first-gender analysis should be used in future experiments.

Limitations

First of all, our analysis includes only two target languages belonging to the same language family. Furthermore, only one domain was analysed on a relatively small corpus. Therefore, the analysis of different products/topics, although showing some clear tendencies, is not fully reliable. Furthermore, the corpus is not designed for gender evaluation, so that only two thirds of the corpus were actually convenient for the experiment. Due to the nature of the two languages, only two genders were included. However, the possibilities for inclusive language were discussed.

As for ChatGPT translations, we used the version based on GPT-3.5 instead of the newest one based on GPT-4. However, the free version is still based on GPT-3.5, so that a large number of users are still using this one.

Acknowledgements

ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme under Grant Agreement No. 13/RC/2106_P2.

References

Cho, Won Ik, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy, August. Association for Computational Linguistics.

Daems, Joke and Janiça Hackenbuchner. 2022. DeBiasByUs: Raising awareness and creating a database of MT bias. In Moniz, Helena, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 289–290, Ghent, Belgium, June. European Association for Machine Translation.

Hada, Rishav, Agrima Seth, Harshita Diddee, and Kalika Bali. 2023. “Fifty shades of bias”: Normative ratings of gender bias in GPT generated English text. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1862–1876, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of WMT-2022*, Abu Dhabi, United Arab Emirates (Hybrid), December.

Lapshinova-Koltunski, Ekaterina, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations. In *Proceedings of LREC-2022*, pages 1751–1760, Marseille, France, 20-25 June. ELDA.

Měchura, Michal. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings* *of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington, July. Association for Computational Linguistics.

Ramesh, Krithika, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online, August. Association for Computational Linguistics.

Saunders, Danielle, Rosie Sallis, and Bill Byrne. 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*, pages 35–43, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

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*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

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

Vanmassenhove, Eva and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online, August. Association for Computational Linguistics.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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*, pages 2979–2989, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhao, Jieyu, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November. Association for Computational Linguistics.