Building Bridges: A Dataset for Evaluating Gender-Fair Machine Translation into German

Anonymous ACL submission

Abstract

The translation of gender-neutral person terms (e.g., *the students*) is often non-trivial. An interesting case poses the translation from English to German – in German, every noun is gendered, and if the gender of the referent(s) is unknown or diverse, the generic masculine (*die Studenten (m.)*) is commonly used. This, however, reduces the visibility of other genders, such as women and non-binary people. To counteract gender discrimination, a societal movement towards using gender-fair language exists (e.g., by adopting neosystems). However, gender-fair German is currently barely supported in Machine Translation (MT), requiring costly post-editing or manual translations. We address this research gap by studying gender-fair language in English to German MT. Concretely, we enrich a community-created gender-fair language dictionary, and sample multi-sentence test instances from encyclopedic text and parliamentary speeches. Using these novel resources, we conduct the first benchmark study involving two commercial systems and six neural MT models for translating words in isolation and words in larger contexts across two domains. Our findings show that most systems produce mainly masculine forms, and rarely gender-neutral variants, highlighting the need for future research. [1]

1 Introduction

Gender equality is one of the United Nation's sustainable development goals.[2] As psychological research shows that linguistic forms influence the mental representation of gender identities (Sczesny et al., 2016), many organizations are officially adopting *gender-fair language (GFL)*. [3]

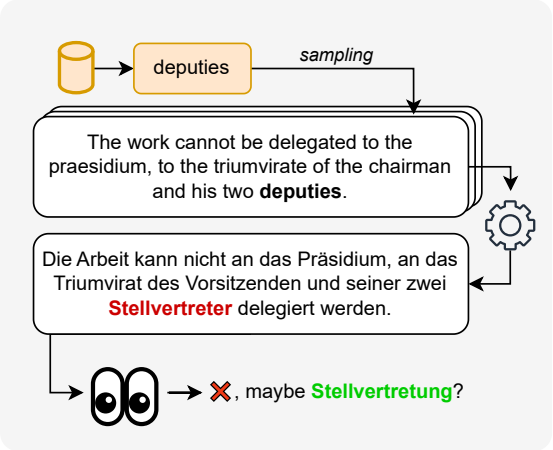

Figure 1: Study overview. We collect English person nouns (yellow, top box) and sample passages representing their mentions in context. We translate those passages with MT systems (white, central boxes) and conduct a human as well as an automatic evaluation on gender forms (bottom) used in German translations.

Towards reaching equality and inclusion, language technology should account for GFL. In this context, recent research in Natural Language Processing (NLP) explores issues around machine translation (MT; e.g., Piergentili et al., 2023a; Savoldi et al., 2023). For instance, when translating gender-neutral person words (e.g., *the students* in English) to a language with grammatical gender, the output may default to a specific gender (e.g., *die Studenten (m.)* in German), thus be exclusive to other gender identities (Dev et al., 2021), and reinforce stereotypical biases (Stanovsky et al., 2019).

However, the existing landscape of research on gender-fair MT is still scarce (Lardelli and Gromann, 2023a): previous studies are limited to covering only a few languages, scenarios, and domains – resources for testing MT systems are barely available. In this short paper, we address this gap by presenting the first large-scale study on GFL in English to German MT to date (see Figure 1).

[1] Code and data at: https://anonymous.4open.science/r/german-fair-mt

[2] https://sdgs.un.org/goals/goal5

[3] See, for instance, this recommendation by the European Parliament: http://www.europarl.europa.eu/RegData/publications/2009/0001/P6_PUB(2009)0001_EN.pdf

Contributions. (**1**) We present GENDER-FAIR GERMAN DICTIONARY, a novel resource that lists gender-neutral and gender-inclusive variants in German and their English translation. We create the resource by enriching a community-created dictionary for German GFL. (**2**) We collect multi-domain data for testing the translation of gender-neutral terms from English to German in context, aligned with our dictionary. (**3**) We benchmark GFL in English to German translations involving two dedicated MT systems and six instruction-tuned models. We answer the following questions: (**RQ1**) *Which grammatical genders are prevalent in English to German MT outputs?* We demonstrate that modern MT systems are systematically biased towards the masculine gender. GFL is extremely rare (0–2% of all translations). (**RQ2**) *Do we observe significant differences when translating isolated words in comparison to their mentions in actual contexts?* Across two domains (Wikipedia, Europarl) we show that additional context does not trigger a significantly higher portion of GFL translations. (**RQ3**) *To what extent can the benchmarking of gender-fair German MT be automatized?* Our results show that GPT models struggle to recognize the overt gender of referents in English to German translations.

2 Background

Gender-Fair Language (GFL). Drawing on Sczesny et al. (2016), we use *"gender-fair"* as an umbrella term subsuming two distinct approaches: *gender-neutral* and *gender-inclusive* language. Gender-neutral describes strategies to avoid gender reference, e.g., by using passive constructions, and gender-neutral nouns. In contrast, gender-inclusive refers to the use of different typographical characters, e.g., the interpoint (·) in French; and symbols, e.g., schwa (ə) in Italian; etc., to make all genders visible.

GFL Strategies for German. In German, there are four main approaches to gender-fair language (Lardelli and Gromann, 2023c):

(1) Gender-neutral rewording is the use of passive constructions, indefinite pronouns, gender-neutral terms, or participles instead of gendered nouns;

(2) Gender-inclusive characters such as gender star (*), colon (:), or underscore (_) are used to combine masculine and feminine forms as in *"der*die Leser*in"* (*m.*f.* article *m.*f.* noun, the reader);

(3) Gender-neutral characters and/or endings are similar to the previous approach and include the use of *"x"* or *"*"* to, however, replace gender suffixes as in *"dix Lesx"*;

(4) Gender-fair neosystems introduce a fourth gender in German alongside masculine, feminine and neuter. New pronouns, articles, and suffixes are proposed, e.g., *"ens"* in *"dens Lesens"*.

In this paper, we focus on strategies (1) and (2) because these are currently the most common approaches in general language use and the most likely to be adopted by professional translators (Lardelli and Gromann, 2023b).

3 Data for Gender-Fair MT

We release two resources for studying GFL in English to German MT: first, we assemble a dictionary (§3.1) of person nouns; second, we sample passages from Wikipedia and Europarl (§3.2) representing the use of terms in the wild.

3.1 Gender-Fair Dictionary

Acknowledging the importance of hearing the voices of affected individuals in GFL research (Gromann et al., 2023), we start from the *"Gender-wörterbuch"*[4]. This website hosts a community-created vocabulary: users add gender-fair, usually neutral, alternatives to commonly gendered terms. Next, we sample and select suitable terms for our research, and further enrich the dictionary.

Term Selection. We start from 128 randomly selected terms. We filter out those that were already neutral, e.g., *"Star"*, which is an anglicism and does not have variants for other genders in German. To facilitate back-translation into English, we remove polysemous terms, e.g., *"aid"*.

Dictionary Enrichment. One of the authors—expert in GFL and translation—annotates every noun with its masculine, feminine, gender-inclusive, and gender-neutral form in singular and plural. We use gender star (*) for gender-inclusive forms, as it is common in German-speaking countries (Körner et al., 2022). Finally, we automatically add the English translation for each term. Our final dictionary counts **115 nouns** in their singular and plural forms (see Table 1). Notably, the final list contains both professions (e.g., *"deputy"*) as well as common nouns (e.g., *"donor"*). While, to date, most research on gender bias in MT focused

[4] https://geschicktgendern.de

on the translation of profession terms only (Prates et al., 2020), we expand the focus and include common nouns referring to people in a broader sense.

3.2 Multi-Sentence Multi-Domain Mentions

We collect an additional set of English passages that mention our dictionary entries.

Data Selection. We sample sentences from Europarl (Koehn, 2005) and Wikipedia.[5] Europarl is a widely recognized benchmark dataset for MT displaying institutional language from parliamentary speeches—perhaps amongst the first contexts GFL was designed for (Piergentili et al., 2023b). Wikipedia presents encyclopedic text, opening to new contexts where our seed nouns appear.

Passage Retrieval. For each of our 115 terms, we retrieve all sentences in a given corpus with at least one occurrence of the noun.[6] Since gender assignment might require cross-sentence resolutions, we extract the matching sentence along with two preceding sentences and one following. We sample five passages per seed noun. Finally, we manually inspect the outputs to ensure the quality of the retrieved passages. We ensure that (i) the gender of the seed words is ambiguous or mixed (e.g., masculine and feminine), (ii) the passage meaning is self-contained, and (iii) the passages do not exceed a length of 100 words.[7] Respectively, 36 and 35 of the words did not match any sentence in Europarl and Wikipedia, whereas in some other cases we matched one or two sentences only.

In total, we collected 358 passages from Europarl and 400 from Wikipedia, e.g., "The work cannot be delegated to the praesidium, to the triumvirate of the chairman and his two *deputies*."[8]

4 Experiments

4.1 Translation System Selection

Acknowledging that, today, people are exposed to MT in multiple ways, we include in our study a variety of systems. As commercial representatives of dedicated MT systems, we include Google Translate and DeepL. Additionally, we study GPT 3.5 and GPT 4 (OpenAI, 2023), accessible through online APIs. Complementary to these, we also include open models: two supervised MT models, NLLB (Costa-jussà et al., 2022) and OPUS MT (Tiedemann and Thottingal, 2020), Flan-T5 (Chung et al., 2022), a multi-task instruction fine-tuned model, and Llama 2 (Touvron et al., 2023). See Appendix A for full details.

Gender Form	Singular	Plural
Masculine	*Berater*	*Berater*
Feminine	*Beraterin*	*Beraterinnen*
Gender-neutral	*Beratende*	*Beratenden*
Gender-inclusive	*Berater*in*	*Berater*innen*

Table 1: Dictionary entry for *"counsellor(s)."*

4.2 Translation and Evaluation

We translate the seed words in our dictionary, both in the singular and then in the plural form, as well as the passages retrieved from Europarl and Wikipedia (§3.2). We then manually evaluate (same author as in §3.1) the translations, annotating whether the person term is translated with a masculine, feminine, gender-inclusive, or gender-neutral form. Since words in isolation were sometimes mistranslated, such mistakes are annotated accordingly. We annotate three out of the five passages retrieved from Europarl and Wikipedia. This allows us to exclude passages that are too complex, present formatting problems (e.g., Wikipedia section titles), or are not translated by the models. Additionally, to answer **RQ3**, we prompt GPT-3.5 to zero-shot label GFL in the translations (see Appendix B for details). We compare these results with the manual annotations.

4.3 Results

Words-in-Isolation. As shown in Table 3, all models are heavily biased towards masculine forms (93–96% of all translations, (**RQ1**)). Feminine forms are found quite rarely (2–4%), usually when nouns relate to professions that are stereotypically associated with femaleness like *"children's day carer"*, *"kindergarten teacher"*, and *"secretary"*. Gender-neutral and inclusive forms are rare (0–2%).

The analysis of plural translations yields similar results (Figure 2). Gender-neutral forms occur slightly more frequently (4–8% of all translations) for two reasons. First, while some nouns, e.g., *"practitioner"* are gender-specific in the singular

[5] We use the snapshot at 01–03–2022 at `https://huggingface.co/datasets/wikipedia`.

[6] We use `nltk` to split paragraphs into sentences. Since several words can be used as adjectives, we extract POS tags with `spacy`'s morphological utility and match only NOUNs.

[7] The average passage length is 34 and 92 words for Europarl and Wikipedia, respectively.

[8] Note that while this work focuses on evaluating German translations, our resource can be enriched with any grammatical gender language where gender-fair language approaches ought to be preferred on masculine generics (e.g., Piergentili et al., 2023a).

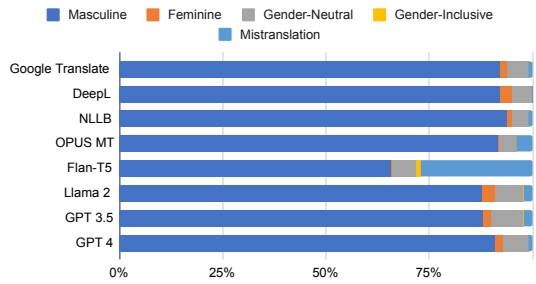

Figure 2: Results of our analysis on plural words in isolation.

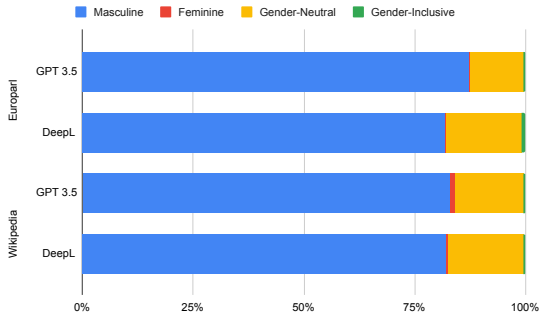

Figure 3: Results of our analysis on words in context.

(*"Praktiker"/"Praktikerin"*), gender-neutral alternatives are common for plural (*"Praktizierende"*). Second, some nouns have the same form for masculine and feminine but the article is gender-specific in the singular only, e.g., *"the relative"* (*"die Angehörige"/ "die Angehörigen"*).

Words-in-Context. For answering **RQ2**, we test GPT 3.5 and DeepL on passages from Europarl and Wikipedia because they are the models that produced the highest number of non-masculine translations among MT systems and language models. The results are shown in Figure 3. Both models are strongly biased towards masculine forms (85%). While feminine and gender-inclusive forms are rare, in about 1% of cases, gender-neutral forms are more common (~15%) – usually for nouns that are already gender-neutral (e.g., *"travellers"*, *"respondents"*, and *"relatives"*) or for which a gender-neutral alternative is common in the plural (e.g., *"practitioners"*, *"chairpeople"*, *"newcomers"*).

Zero-shot GFL Detection. As described in (§4.2), we test whether GPT-3.5 and GPT-4 can serve as viable tools for automatizing GFL evaluations. To this end, we prompt the models to label the translations of words in context produced with GPT-3.5 and compare the results with our manual annotations. Both GPT-3.5 and GPT-4 perform poorly (Table 2), demonstrating the continuous need for experts when studying GFL in MT.

Gender	P	R	S
Masculine	92.9	69.7	188
Gender-Inclusive	4.8	100	1
Gender-Neutral	30.0	11.5	26
Masculine	96.3	96.3	188
Gender-Inclusive	6.2	100	1
Gender-Neutral	75.0	11.5	26

Table 2: (P)recision, (R)ecall, and (S)upport of GPT-3.5 (top) and GPT-4 (bottom) zero-shot labeling when compared to human analysis. We found no feminine forms. Europarl EN-DE (n=215).

5 Related Work

Due to stereotypical and exclusive biases present in the training data, the output of MT may discriminate against certain genders (e.g., Stanovsky et al., 2019; Attanasio et al., 2023). In this context, recent research has focused on the issue of gender exclusivity (Piergentili et al., 2023a). Towards a better understanding, much attention has been paid to studying existing strategies chosen by human subjects, like translation team leaders (Daems, 2023), and MT post editors (Lardelli and Gromann, 2023b; Paolucci et al., 2023). Related to this, Gromann et al. (2023) pointed to participatory research as a promising avenue. Another research thread focuses on assessing the capabilities of existing MT systems: Lauscher et al. (2023) investigated the translation of pronouns in commercial MT, Saunders and Olsen (2023) the translation of named entities, and Piergentili et al. (2023b) benchmarked gender-neutral MT from English to Italian. Savoldi et al. (2023) report the results of a shared task, designed to assess the GFL ability of MT systems from German to English. The only existing work, which also focuses, like ours, on English to German GFL is Kostikova et al. (2023). However, the authors study 15 sentences only. In contrast, we focus on 115 words in multiple translation scenarios.

6 Conclusion

In this work, we have presented the first large-scale study on gender fairness in EN to DE MT involving eight MT systems tested across several scenarios (e.g., on parliamentary data). To this end, we presented two novel resources, grounded in community contributions. Our findings clearly call for more research on GFL in modern MT, towards fairer and more inclusive language technology.

Limitations

To date, there is no common standard for gender-fair language (Ackerman, 2019). Hence, investigating its use in MT requires manual analysis from experts in the field. This limits the number of language pairs that can be analyzed. For this reason, the present contribution focuses on EN to DE MT only.

Ethical Considerations

By investigating gender bias in MT, our work focuses on the exclusionary potential of language technologies which might impact the visibility and/or mental health of minoritized groups such as women and non-binary people (Sczesny et al., 2016; McLemore, 2018). Here, we also enrich a community-created GFL dictionary. Since there is no acknowledged standard for GFL, the alternatives we present in our work are not prescriptive.

References

Lauren M Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1):1–17.

Giuseppe Attanasio. 2023. Simple Generation. https://github.com/MilaNLProc/simple-generation.

Giuseppe Attanasio, Flor Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Joke Daems. 2023. Gender-inclusive translation for a gender-inclusive sport: strategies and translator perceptions at the international quadball association. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 37–47, Tampere, Finland. European Association for Machine Translation.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff 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*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dagmar Gromann, Manuel Lardelli, Katta Spiel, Sabrina Burtscher, Lukas Daniel Klausner, Arthur Mettinger, Igor Miladinovic, Sigrid Schefer-Wenzl, Daniela Duh, and Katharina Bühn. 2023. Participatory research as a path to community-informed, gender-fair machine translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 49–59, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Anita Körner, Bleen Abraham, Ralf Rummer, and Fritz Strack. 2022. Gender representations elicited by the gender star form. *Journal of Language and Social Psychology*, 41(5):553–571.

Aida Kostikova, Joke Daems, and Todor Lazarov. 2023. How adaptive is adaptive machine translation, really? a gender-neutral language use case. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 95–97, Tampere, Finland. European Association for Machine Translation.

Manuel Lardelli and Dagmar Gromann. 2023a. Gender-fair (machine) translation. In *Proceedings of the New Trends in Translation and Technology Conference - NeTTT 2022*, pages 166–177.

Manuel Lardelli and Dagmar Gromann. 2023b. Gender-fair post-editing: A case study beyond the binary. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260, Tampere, Finland. European Association for Machine Translation.

Manuel Lardelli and Dagmar Gromann. 2023c. Translating non-binary coming-out reports: Gender-fair language strategies and use in news articles. *The Journal of Specialised Translation*, (40):213–240.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Kevin A McLemore. 2018. A minority stress perspective on transgender individuals' experiences with misgendering. *Stigma and Health*, 3(1):53–64.

OpenAI. 2023. Gpt-4 technical report. *preprint*.

Angela Balducci Paolucci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland. European Association for Machine Translation.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore. Association for Computational Linguistics.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Danielle Saunders and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93, Tampere, Finland. European Association for Machine Translation.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7.

Gabriel Stanovsky, 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. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Details on Translation Systems

We used paid APIs and `deep-translator`[9] for Google Translate and DeepL, and accessed `gpt-3.5-turbo-0613` (GPT 3.5) and `gpt-4-0613` (GPT 4). For all open-weight models, we used code and implementation from `transformers` (Wolf et al., 2020) and `simple-generation` (Attanasio, 2023) as the inference engine. In particular, we used Helsinki-NLP/opus-mt-en-de (OPUS MT), facebook/nllb-200-3.3B (NLLB), google/flan-t5-xxl (Flan-T5), and meta-llama/Llama-2-70b-chat-hf (Llama 2).

To run the experiments, we used an in-house computing center and run all the experiments on one A100 GPU.

Prompt and Decoding. We used no prompts from supervised MT models, whereas for Llama 2 and GPTs we used:

```
Translate the following sentence into German. Reply only with the translation. Sentence: {sentence}
```

Finally, we followed FLAN's (Longpre et al., 2023) translation templates for Flan-T5:

```
{sentence}\n\nTranslate this into German?
```

We used the default generation configuration for GPTs, beam search decoding (n=5) for OPUS MT, NLLB, and Flan-T5, and nucleus sampling (top p=1, top k=50, temperature=0) for Llama 2.

[9] https://github.com/nidhaloff/deep-translator

B Automatic Evaluation

We prompted GPT-3.5 (`gpt-3.5-turbo-0613`) with default decoding parameters to evaluate whether machine translated passages used any gender-fair form.

The prompt we used it:

```
If the following sentence contains the German translation for the English word "seed_noun", tell me which overt gender it displays among Masculine, Feminine, Gender-Neutral, or Gender-Inclusive. If no translation is found, reply with None. Sentence: "translation"
```

Detailed Results

Model	Gender				Mi
	M	F	GN	GI	
DeepL	108	5	1	1	0
GT	107	3	1	0	4
GPT 3.5	107	2	1	0	5
GPT 4	108	2	1	0	4
NLLB	111	2	1	0	1
OPUS MT	110	3	0	0	0
Flan-T5	51	9	3	0	39
Llama 2	107	3	2	0	2

Table 3: Results of the words in isolation analysis (singular). For each seed word, we count masculine (M), feminine (F), gender-neutral (GN), and gender-inclusive (GI) and mistranslations (Mi)

Model	Gender				Mi
	M	F	GN	GI	
DeepL	106	3	6	0	0
GT	105	2	6	0	1
GPT 3.5	101	2	10	0	2
GPT 4	105	2	7	0	1
NLLB	108	1	5	0	1
OPUS MT	105	0	5	0	5
Flan-T5	76	0	7	1	31
Llama 2	101	3	8	0	2

Table 4: Results of the words in isolation analysis (plural). For each seed word, we count masculine (M), feminine (F), gender-neutral (GN), and gender-inclusive (GI) and mistranslations (Mi)

Domain	Model	Gender			
		M	F	GN	GI
Europarl	GPT 3.5	188	0	26	1
	DeepL	177	0	37	1
Wikipedia	GPT 3.5	181	2	34	1
	DeepL	178	1	38	1

Table 5: Results of the words in context analysis (plural). For each seed word, we count masculine (M), feminine (F), gender-neutral (GN), and gender-inclusive (GI)