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

Manuel Lardelli[♣*], **Giuseppe Attanasio**[♠*], **Anne Lauscher**[♢]

[♣] University of Graz, Austria
[♠] Instituto de Telecomunicações, Lisbon, Portugal
[♢] University of Hamburg, Germany
manuel.lardelli01@gmail.com

Abstract

The translation of gender-neutral person-referring terms (e.g., *the students*) is often non-trivial. Translating from English into German poses an interesting case—in German, person-referring nouns are usually gender-specific, and if the gender of the referent(s) is unknown or diverse, the generic masculine (*die Studenten (m.)*) is commonly used. This solution, 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 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 natural 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. We release code and data at https://github.com/g8a9/building-bridges-gender-fair-german-mt.

1 Introduction

Gender equality is one of the United Nation's sustainable development goals.[1] 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)*.[2]

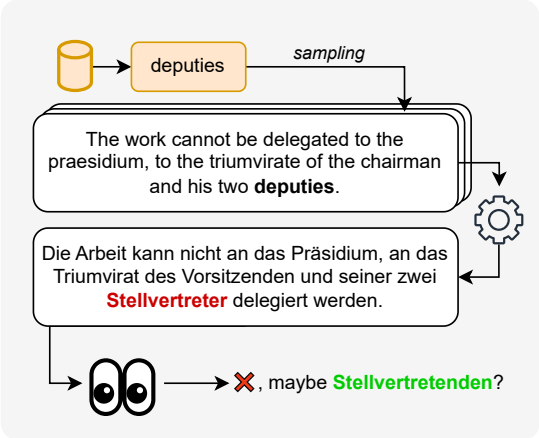

Figure 1: **Study overview.** We collect English person nouns (yellow, top box) and sample passages representing their mentions in context. We translate those passages with MT systems (white, central boxes) and conduct a human as well as an automatic evaluation on gender forms (bottom) used in German 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 being exclusive to other gender identities (Dev et al., 2021), and reinforcing stereotypical biases (Stanovsky et al., 2019).

[*] Equal contribution.
[1] https://sdgs.un.org/goals/goal5
[2] See, for instance, this recommendation by the European Parliament: http://www.europarl.europa.eu/RegData/publications/2009/0001/P6_PUB(2009)0001_EN.pdf

Findings of the Association for Computational Linguistics: ACL 2024, pages 7542–7550

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—none of which focuses on German specifically. In this short paper, we address this gap by presenting the first study on GFL in English-to-German MT. See Figure 1 for an overview.

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 compile this 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 into 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 overt 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 natural contexts?* Across two domains (encyclopedic and parliament speeches) we show that additional context does not yield 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 beyond the masculine and feminine forms.

2 Background

2.1 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, to make all genders visible.

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."*

2.2 GFL Strategies for German

In German, there are four main approaches to gender-fair language (Lardelli and Gromann, 2023c). *Gender-neutral rewording* uses passive constructions, indefinite pronouns, gender-neutral terms, or participles instead of gendered nouns. *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. Eng: the reader). *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"*. *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-referring nouns. Second, we sample passages from Wikipedia and Europarl (§3.2) to study our terms in natural contexts. Words in isolation allow for testing priors in translation systems, i.e., the most likely gender form for a noun when no context is provided. Natural passages enable studying the effect of contextual clues.

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 to masculine generics (e.g., Piergentili et al., 2023a).

3.1 Gender-Fair Dictionary

Acknowledging the importance of hearing the voices of affected individuals in GFL research (Gro-

Preceding Context	First release On July 15, 2019, EndeavourOS released their first ISO. The team did not expect that much of the Antergos community would follow them, but the response and the numbers of community members that joined exceeded their expectations.
Matching Sentence	Not only did the community receive the first release very well, but several **bloggers** and vloggers gave it very positive reviews, even shortly after launch.
Trailing Context	Immediately after the launch of the distribution, the EndeavourOS team began to develop a net-installer to install with different Desktop Environments directly over the internet.

Table 2: **Multi-Sentence Passage from Wikipedia.** Seed noun *"bloggers"* in bold. The gender of the seed is ambiguous and cannot be resolved from either context.

mann et al., 2023), we start from the *"Gender-wörterbuch."*[3] This website hosts a community-created German 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—experienced with GFL and translation—enriched 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 manually translated each term into English. Our final dictionary counts 115 nouns in their singular and plural forms (see Table 1 for an example). 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 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 in their plural form. We focus on plural occurrences because they yield to gender-ambiguous cases more frequently, providing a more challenging scenario for translation systems.

[3] https://geschicktgendern.de

Data Sources. We sample sentences from Europarl (Koehn, 2005) and Wikipedia.[4] Europarl is a widely recognized benchmark dataset for MT displaying institutional language from parliamentary speeches—perhaps amongst the first contexts GFL was devised 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.[5] The seed's gender assignment might require cross-sentence resolution. Thus, limited to Wikipedia, we extract the matching sentence along with two preceding sentences and one following (see Table 2).

Concretely, we sample passages in two steps. First, we randomly selected five passages per seed noun, yielding an initial batch of 358 single-sentence passages from Europarl and 400 multi-sentence passages from Wikipedia. Respectively, 36 and 35 seeds did not match any sentence in Europarl and Wikipedia, and we matched only one or two sentences for some seeds. Then, we manually filtered passages via quality checks on the matching sentence. Specifically, we ensure that (i) the overt gender of the seed words is ambiguous or it refers to a mixed-gender group, (ii) the passage meaning is self-contained, and (iii) the passages do not exceed a length of 100 words.[6]

4 Experiments

4.1 Translation System Selection

Acknowledging that, today, people are exposed to MT in multiple ways, we include in our study a va-

[4] We use Europarl's release v7 (parallel corpus English-German) and the Wikipedia snapshot at 01–03–2022 at https://huggingface.co/datasets/wikipedia.

[5] 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.

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

	# SN	# RP	# AP
Europarl	79	358	215
Wikipedia	80	400	218

Table 3: **GENDER-FAIR GERMAN DICTIONARY statistics.** Number of seed nouns (SN), and retrieved (RP) and annotated passages (AP) from Europarl and Wikipedia.

riety 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. We also include open-weight models: two supervised MT models, NLLB (Costa-jussà et al., 2022) and OPUS MT (Tiedemann and Thottingal, 2020), Flan-T5 (Chung et al., 2024), a multi-task instruction fine-tuned model, and Llama 2 (Touvron et al., 2023). See Appendix A for full details.

4.2 Translation and Evaluation

We machine-translated all seed words in isolation (singular and plural) and the passages retrieved from Europarl and Wikipedia (§3.2). The same author from §3.1 manually annotated whether the term is translated with a masculine, feminine, gender-inclusive, or gender-neutral form. Since words in isolation were sometimes mistranslated, we noted the type of errors. Mistakes were due to semantics (i.e., the German term has a different meaning than the English source), grammar (e.g., wrong number or no agreement between article and nouns), and hallucinations. We finally annotate three out of the five passages retrieved from Europarl and Wikipedia with the same criteria.[7] The final number of annotated passages is 215 and 218 for Europarl and Wikipedia, respectively. In order to validate our analysis, a German native speaker student research assistant with previous experience in MT output annotation repeated the analysis on a the OPUS MT's outputs in the singular and plural, and the GPT3.5's translations of the Wikipedia passages.

Additionally, to answer **RQ3**, we prompt GPT 3.5 in zero-shot to detect GFL in the translations (see Appendix B for details). We compare these results with the manual annotations.

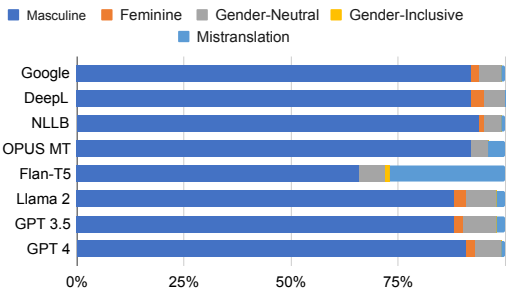

Figure 2: **Results for plural words in isolation.** Gender form distribution and mistranslations for each translation system.

4.3 Results

Words-in-Isolation. As shown in Table 5 (see Appendix C), **all models are heavily biased towards masculine forms** (93–96% of all translations, (**RQ1**)). MT systems use feminine forms seldom (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 even rarer (0–2%). One example is the term *"newcomer,"* translated by nearly all models with the gender-neutral *"Neuling"* that is very common in German. While its grammatical gender is masculine, it is used for all genders. Interestingly, Flan-T5 produced many mistranslations. For instance, the seed noun *"traveller"* was translated to *"Reisenden"* with a grammatical mistake in the noun declension, i.e., the suffix "n." The model also created non-existing words, e.g., *"Antwortent"* for *"respondent,"* instead of *"Befragten"* ("Antwort" is "answer" in English).

The analysis of plural translations yielded similar results (see Figure 2). Gender-neutral forms occur slightly more frequently (4–8% of all translations), probably because of two reasons. First, while some nouns, e.g., *"practitioner"* are gender-specific in the singular (*"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"*). We report full results in Table 6, Appendix C.[8]

[7] The two passages excluded contain complex phrasing, formatting problems—e.g., Wikipedia section titles, which are not proper preceding context—or severe translation mistakes.

[8] A second annotator (see 4.2) annotated OPUS MT's outputs in the singular and plural. Agreement on gender forms was perfect, with a Cohen's kappa of 1.00.

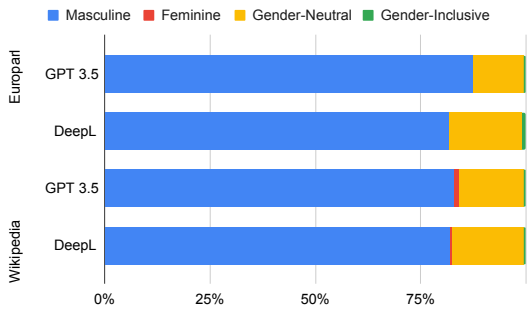

Figure 3: **Results for words in context (plural).** Gender form distribution in GPT 3.5 and DeepL translations for each data source.

Words-in-Context. For answering **RQ2**, we conducted a focused analysis on GPT 3.5 and DeepL translations of the passages retrieved from Europarl and Wikipedia. These models produced the highest number of non-masculine translations among MT systems and language models. The results are shown in Figure 3 and absolute frequencies are reported in Table 7 in Appendix C. **Both models are strongly biased towards masculine forms** (85% of all translations). While feminine and gender-inclusive forms are rare (about 1% of cases), gender-neutral forms are more common (~15%). Systems use them 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"*).[9]

Zero-shot GFL Detection. We test whether GPT 3.5 and GPT 4 can serve as viable tools for automatic detection of GFL. To this end, we prompted the models to label the translations of words in context produced with GPT 3.5 and compare the results with our manual annotations. Note that we found no feminine forms in this set of translation.

Table 4 reports agreement results with human evaluation. Both GPT 3.5 and GPT 4 achieve an extremely low recall (11.5%) for gender-neutral cases. However, GPT 4's precision is relatively high (75%) compared to GPT 3.5 (30%), showing an improvement model generations. These findings highlight that **zero-shot automatic detection of GFL in German with recent GPT models is hard**, and underscore the importance of expert human oversight when studying GFL in MT.

[9]In this case, the second rater repeated the analysis on a portion of data, i.e., GPT 3.5's translations of the Wikipedia passages. Cohen's kappa was 0.954. The few disagreements were mistakes that were corrected.

Gender	P	R	S
Masculine	92.9	69.7	188
Feminine	-	-	0
Gender-Inclusive	4.8	100	1
Gender-Neutral	30.0	11.5	26
Masculine	96.3	96.3	188
Feminine	-	-	0
Gender-Inclusive	6.2	100	1
Gender-Neutral	75.0	11.5	26

Table 4: **Automatic detection of GFL.** (P)recision, (R)ecall, and (S)upport of GPT 3.5 (top) and GPT 4 (bottom) zero-shot labeling when compared to human analysis. 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

We have presented the first study on gender-fair EN-DE MT. We introduced two novel resources grounded in community contributions. We experimented with eight translation systems and several setups, including words in isolation, natural passages from the encyclopedic domain and parliamentary speeches. Our findings call for more research on GFL in modern MT towards fairer and more inclusive translation technology.

Acknowledgments

Giuseppe Attanasio was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020. He conducted part of the work as a member of the MilaNLP group at Bocconi University, Milan.

Limitations

This work comes with several limitations.

We focus on a single language pair and direction, English->German. Our choice is dictated by the lack of consensus for gender-fair language (Ackerman, 2019). German is a notable exception where seminal work has been recently conducted. Hence, we opted to limit the scope and study whether MT systems keep up with the growing trends.

Our study is limited to a relatively small number of seed nouns and sampled sentences. We acknowledge this aspect but highlight that our procedure generalizes easily to new seeds and data sources. Moreover, when sampling natural passages, we did not control for specific factors related to gender and gender inflection. First, we focus on sentences where the entity's gender is ambiguous or mixed. Therefore, we discard all cases where the entity's gender is disambiguated, for example, by lexical clue within the matching sentence. Second, we do not control for the presence of other human entities that might act as a confounding factor.

Ethical Considerations

By investigating gender 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, though they represent common strategies.

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, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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`[10] for Google Translate and DeepL, and accessed

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

`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.

B Automatic Evaluation

We prompted GPT 3.5 and GPT 4 with default decoding parameters to evaluate whether machine translated passages used any gender-fair form.

The prompt we used is:

```
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}
```

We substituted `seed_noun` and `translation` accordingly.

C Detailed Results

Tables 5–7 report the total occurrences of different gender forms and mistranslations in our setups.

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 5: **Results of the words in isolation analysis (singular).** For each seed word, we count masculine (M), feminine (F), gender-neutral (GN), and gender-inclusive forms (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 6: **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) forms, and mistranslations (Mi).

Source	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 7: **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) forms.