What the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-centered Study

Beatrice Savoldi[■], **Sara Papi**[■], **Matteo Negri**[■],
Ana Guerberof Arenas[★], **Luisa Bentivogli**[■]
[■]Fondazione Bruno Kessler, Italy
[★]University of Groningen, Netherlands
{bsavoldi,spapi,negri,bentivo}@fbk.eu
a.guerberof.arenas@rug.nl

Abstract

Gender bias in machine translation (MT) is recognized as an issue that can harm people and society. And yet, advancements in the field rarely involve people, the final MT users, or inform how they might be impacted by biased technologies. Current evaluations are often restricted to automatic methods, which offer an opaque estimate of what the downstream impact of gender disparities might be. We conduct an extensive human-centered study to examine if and to what extent bias in MT brings harms with tangible costs, such as quality of service gaps across women and men. To this aim, we collect behavioral data from ∼90 participants, who post-edited MT outputs to ensure correct gender translation. Across multiple datasets, languages, and types of users, our study shows that feminine post-editing demands significantly more technical and temporal effort, also corresponding to higher financial costs. Existing bias measurements, however, fail to reflect the found disparities. Our findings advocate for human-centered approaches that can inform the societal impact of bias.

1 Introduction

Natural language processing (NLP) has evolved from an academic specialty to countless commercial applications that can both benefit and negatively affect people's lives. With the widespread use of these technologies, researching the ethical and social impact of NLP has become increasingly crucial (Hovy and Spruit, 2016; Sheng et al., 2021), with gender fairness being a major concern (Sun et al., 2019; Stanczak and Augenstein, 2021).

In machine translation (MT) gender bias has received significant attention, also in the public domain (Olson, 2018). Numerous studies have shown that MT perpetuates harmful stereotypes (Stanovsky et al., 2019; Triboulet and Bouillon, 2023) and is skewed towards masculine forms that under-represent women (Vanmassenhove et al., 2018; Alhafni et al., 2022b).

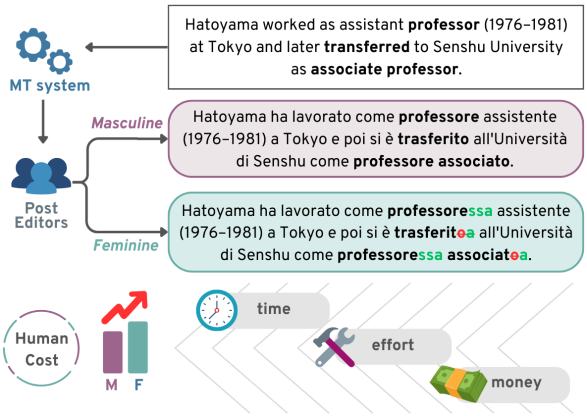


Figure 1: Harms as assessed in our study design. We task participants with the post-editing of an MT output into both feminine and masculine gender. We collect behavioural data (i.e. time and technical effort) and assess higher workload and economic costs associated with feminine translations.

As emphasized by Savoldi et al. (2021) – if we regard MT as a resource in its own right – such representational disparities might directly imply *allocative* harms, i.e. differential access to material benefits that make a social group or individual worse-off (Barocas et al., 2017; Chien and Danks, 2024). For instance, a woman using an MT system to translate her biography (i.e. the first sentence in English in Figure 1) into Italian would need more effort (i.e. represented by insertions – in green, and substitutions – in red and green – in Figure 1) to revise incorrect masculine references, thus experiencing a disparity in the quality of the service.

Despite acknowledging the potential harm to individuals, research on gender bias in MT primarily focuses on in-lab automatic evaluations. Such assessments, however, are only assumed to reflect a real-world downstream effect, without verifying if and to what extent biased models might concretely impact users interacting with a system.

To address this gap, we examine the effect of gender bias in MT with a human-centered perspec-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18048–18076
November 12-16, 2024

tive. Specifically, we ask: ***Does gender bias in MT imply tangible service disparities across men and women?*** And if so, can we meaningfully quantify them via more human-centered measures? To take stock of the current research landscape, we review the involvement of human subjects in prior literature on gender and MT. Motivated by the outcome, we conduct extensive experiments across multiple datasets, languages, and users. In a controlled setup, 88 participants post-edited MT outputs to ensure either feminine or masculine gender translation.[1] In the process, we track behavioral data – i.e. time to edit and number of edits – to compare effort across genders. Based on this, we estimate the associated cost for post-editing into each gender if the work were assigned to a third-party translator. Our main findings are:

1. Most of current assessments of gender bias in MT either overlook human involvement, or treat individuals as models' evaluators rather than potentially affected users (§2).
2. We find substantial gender disparities in the time and technical effort required to post-edit MT, with feminine translation taking on average twice longer and four times the editing operations (§4).
3. The cost of the found disparities is also economic, and can unfairly fall onto various stakeholders in the translation process (§5).
4. The automatic evaluation of gender bias does not accurately reflect the found human-centered effort disparities (§5).

To sum up, our study marks a step towards understanding the implications of gender bias in MT. While harms have so far been implied, or inferred from automatic scores as a proxy for downstream impact, here we empirically show that gender bias can bring unfair service disparities. What's more, we quantify bias with measures that are more meaningful for potentially impacted individuals: workload and economic costs.

Behvaioural data and post-edits are made available at `https://huggingface.co/datasets/FBK-MT/gender-bias-PE`.

2 Where are the people? `Evaluator != User`

Language technologies have reached a level of quality that enabled laypeople to integrate them into their day-to-day activities (Nurminen and Papula, 2018). With this shift, understanding users' needs, and how they might be impacted becomes of utmost importance. Indeed, NLP is witnessing increasing emphasis towards more human-centered approaches[2] (Robertson et al., 2021; Goyal et al., 2023), but still little is known about the experience of people interacting with such technology – even for wide-reaching, user-facing applications such as MT (Guerberof-Arenas and Moorkens, 2023).[3]

Similarly, the study of bias is emphasized as an intrinsically human-centered endeavour (Bender, 2019) that requires understanding which behaviour might be harmful, how and to whom (Blodgett et al., 2020). Nonetheless, there is a paucity of work that foregrounds human engagement (Cercas Curry et al., 2020; Mengesha et al., 2021; Wang et al., 2024). Arguably, truly informative measurements on the downstream effects of bias and its potential for harm should assume people as target. But in what capacity, if any, have people been involved so far in the study of gender bias in MT?

ACL Anthology search For a systematic review of prior work, we query the ACL anthology.[4] As keywords, we specify our application of interest – e.g. "MT" and "translation" – combined with "gender" or "bias". For a more channelled query focusing on people involved in bias assessment, we also add other human-engagement dedicated keywords (e.g. "user", "survey"). As of April 2024, our search returned 251 articles, 96 of which also matched the human-engagement keywords. Upon manual inspection, we retained 105 *in-scope* manuscripts,[5] and discarded unrelated papers focusing on other definitions of the keywords (e.g. "inductive bias"). The in-scope papers were finally reviewed by focusing on the presence, or lack thereof, of human involvement. For more details on our search and selection, see Appendix A.

Review We report the results of our review in Figure 2. As the image shows, we attest to a steady increase in publications related to gender in MT, in particular from 2020 onwards. In line with our expectations and the general trend in NLP, however,

[1] We discuss the implications this binary setup in §8.

[2] A case in point being the introduction of the "human-centered NLP" track in ACL* conferences.

[3] Briva-Iglesias et al. (2023) claim that also for professional translators existing studies mostly focus on industry-oriented productivity gains rather than on user experience.

[4] `https://aclanthology.org/`

[5] Works focusing on (human) gender translation, bias or fairness in the context of MT.

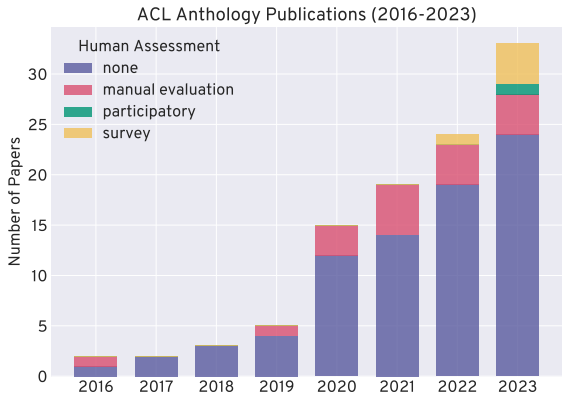


Figure 2: Human involvement in the assessment and framing of gender (bias) in MT, based on an ACL Anthology search. For studies with human participants, we distinguish qualitative, but yet model-centric MANUAL EVALUATION, and more human-centric designs – i.e. SURVEY studies and PARTICIPATORY approaches.

we attest a severe lack of human engagement. In fact, only 24 works rely on humans to measure bias, though in a different capacity, which we distinguish into three conceptual categories. In 18 papers, we find that people – often expert linguists (e.g. Vanmassenhove et al. (2021a); Soler Uguet et al. (2023)) – are involved in MANUAL EVALUATION. This serves to either ensure correlation with bias metrics (e.g. Kocmi et al. (2020)) or to gain qualitative insights that defy automatic approaches (Popović, 2021). While indeed valuable, such analyses are a support for structured, often annotation-based *model-centric* evaluations – i.e. that inform and quantify models' behaviour. Differently, the 5 papers in the SURVEY category focus on the feedback and experiences of potentially impacted groups of users (e.g. Piergentili et al. (2023b); Daems and Hackenbuchner (2022)). For instance, to grasp user preference in how models should handle the translation of novel, non-binary pronouns from English – e.g. *ze, xe* (Lauscher et al., 2023)[6], or to understand the potential trade-off between overall quality and inclusivity goals (Amrhein et al., 2023). Finally, the study by Gromann et al. (2023) recounts a PARTICIPATORY Action Research, where a community-led approach with different stakeholders informs the state and potential direction for gender fair MT.

Overall, despite this recent trend towards surveys or participatory methods, humans are rarely involved to estimate gender bias in MT. Moreover, **if involved, people mostly serve in the capacity of *evaluators***, supporting model-centric assessments **rather than being considered as potentially impacted *users***. Our finding stands in contrast with a qualitative survey by Dev et al. (2021), which found MT as an application with a high risk for downstream harms in the context of gender bias.

Further motivated by such evidence, we carry out a quantitative, empirical study – to the best of our knowledge, the very first of its kind – focusing on human-centered assessments. In particular, we examine whether gender bias in machine translation leads to disparities in the quality of service offered to women and men, by considering different *datasets, languages*, and *users* (§3.1).

3 Experimental setup

We simulate the conditions in which an end user needs the translation of a text referring to them – as described in §1 and exemplified in Figure 1. To strike a balance between controlled conditions for reliable findings while keeping a realistic scenario, the study is realized as a *post-editing task* (PE), where participants are asked to also ensure that human references are rendered as either feminine or masculine. The same output sentences are edited twice (once per gender), thus allowing to isolate any difference in effort as a gender-related factor.

Note that our experiments are based on sentences that always require to translate gender and enable focused analyses. As we further discuss in 7, we thus mimic scenarios that often require to manage gender mentions to human referents, as in the case of biographies, CVs, and administrative texts.

3.1 Settings

Languages We include three language pairs – English→Italian/Spanish/German – which are representative of the challenges of translating into languages with extensive gendered morphology – e.g. *the friend*→ es: ***el/la amig*o/*a***. Overall, these pairs feature sufficiently diverse gender phenomena (Gygax et al., 2019). The selection was also bound to their representation within available datasets.

Datasets We rely on MT datasets that represent naturally occurring gender translation phenomena. Namely, MT-GenEval (Currey et al., 2022) – which is built upon Wikipedia biographies – and the TED-derived Must-SHE corpus

[6]Interestingly, all SURVEY works focus on non-binary linguistic strategies beside feminine/masculine ones. See §8.

		# src-W	# out-W	# tgt-GW
en-it	MTGEN-UN	24	25	4.57
en-it	MUST-SHE	25	24	1.58
en-it	MTGEN-A	17	17	2.38
en-es	MTGEN-A	18	19	2.34
en-de	MTGEN-A	17	17	2.61

Table 1: Data statistics. For each dataset and language, we provide the average number of words for source **(src-W)** and output sentences **(out-W)**, as well as the average number of target gendered words **(tgt-GW)** in the reference translations.

(Bentivogli et al., 2020). Our data samples are organized as follows. ***(i)*** MTGEN-A, a subset of MT-GenEval sentences where gender in the source is ambiguous.[7] ***(ii)*** MTGEN-UN, which contains feminine/masculine versions of gender-unambiguous English sentences,[8] thus offering favourable conditions for correct translation based on available gender cues in the source. Finally, ***(iii)*** a subset of MUST-SHE featuring ambiguous first-person references in the English source sentence.[9] This sample is included for the sake of phenomena variability – given that MUST-SHE entails gendered translation for many parts-of-speech –whereas both Wikipedia-derived samples mainly represent gendered translations for occupational nouns.

As a key feature of these datasets, for each English source sentence, two contrastive feminine/masculine pairs of reference translations are provided. These are designed to isolate gender as a factor from overall quality evaluation.[10]

As described in §4, we conduct multi-*dataset* (§4.1) experiments for en-it, whereas the multi-*language* (§4.2) study with en-es/de is based on MTGEN-A. For each dataset (statistics in Table 1), we retrieve a random sample of 250 sentences, while maximizing the number of common sentences across language pairs.[11]

User types The study aims to reflect an average user, who fixes an MT output by themselves. While including lay users with different levels of language expertise or MT familiarity would represent a comprehensive case study, such a setup adds a notable level of complexity and potential noise to deal with (e.g. gendered expressions to be fixed might be overlooked). To guarantee higher control of our variables, we thus rely on professional translators as a proxy. Still, to also mimic MT interactions with less experienced users, for en-it we carry out multiuser experiments (§4.3) involving high-school students, native speakers of Italian with a B2 level of English (further details in the upcoming §3.2). To avoid fitting our results to the potentially subjective post-editing activity of one person, we allocate 16 post-editors for MUST-SHE and 16 for MTGEN-UN. Since it consists of shorter sentences (see Table 1), we task 14 subjects for each of the four MTGEN-A conditions – for a total of 88 participants overall.

Model Reliable behavioural assessments require a sizable data sample and number of participants, which we prioritize during budget allocation. For this reason, we do not consider MT models as a variable and only use Google Translate (GT). Besides being state-of-the-art, GT is chosen as it represents one of the most widely used consumer-facing commercial MT systems (Pitman, 2021).

3.2 Study design

Task instructions and platform Given a source sentence and its MT output, participants were instructed[12] to carry out a *light* PE – i.e. targeting only essential fixes to adjust the overall quality of the translation (O'Brien, 2022) – with a focus on ensuring either feminine or masculine translation for human referents, based on provided gender information. We choose a light PE *i)* given the high quality of the MT output,[13] and crucially *ii)* to limit the number of preferential edits that might introduce noise. The task was performed with Matecat,[14] a mature, computer-assisted translation (CAT) tool supporting PE that is freely available online.[15]

Within-group design For each data sample of 250 <English source, GT output> pairs, we design a within-subjects study with counterbalancing (Charness et al., 2012), which ensures variation of the order of conditions in the study. Namely, each participant performs *i)* both feminine (F) and masculine (M) post-edits, *ii)* in equal amounts (blocks

[7] e.g. "*Hatoyama worked as assistant professor*[...]"

[8] e.g. "*She was appointed Archdeacon of Lismore* [...]" vs. "*He was appointed Archdeacon of Lismore*[...]."

[9] "*I immediately began to doubt myself* [...]"

[10] These references allow us to compare human-centered results with those of automatic metrics in §5. We adjusted a few inconsistencies in MTGEN-A references – see B.1.2.

[11] See Appendix B.1.1 for details on sample extraction.

[12] For each condition, we prepared dedicated guidelines, which are available at https://github.com/bsavoldi/post-edit_guidelines

[13] E.g., COMET scores are between 82.3-85.3 across all languages and data. See Appendix F.1 for full results.

[14] https://www.matecat.com/

[15] For more details on the Matecat setup see Appendix B.2.

User	Lang	Dataset	TE (↓) FEM	TE (↓) MAS	TE (↓) Δ*abs*	TE (↓) Δ*rel*	HTER (↓) FEM	HTER (↓) MAS	HTER (↓) Δ*abs*	HTER (↓) Δ*rel*	# EDITED SENT (↓) FEM	# EDITED SENT (↓) MAS	# EDITED SENT (↓) Δ*abs*	# EDITED SENT (↓) Δ*rel*
P	*en-it*	MTGEN-UN	2:58	2:11	0:47	36.3	8.17	5.16	3.01	58.3	142	83	59	71
P	*en-it*	MUST-SHE	2:33	1:27	1:06	76.1	8.07	3.16	4.91	155.4	226	58	168	290
P	*en-it*	MTGEN-A	2:38	0:57	1:41	177.6	16.51	5.47	13.08	201.8	243	70	173	247
P	*en-es*	MTGEN-A	2:13	1:13	0:59	81.1	14.88	5.76	9.12	158.3	242	93	149	160
P	*en-de*	MTGEN-A	2:12	0:30	1:42	334.0	15.62	5.47	11.04	515.0	228	40	188	470
S	*en-it*	MTGEN-A	2:08	0:29	1:38	329.8	13.18	1.79	11.39	636.3	242	38	204	573
		AVG.	2:27	1:08	1:19	116.2	12.74	3.98	8.76	220.1	221	64	157	245

Table 2: Multidataset (top), multilanguage (center) and multiuser (bottom) results. Results are shown for all users – both (P)rofessional and (S)tudents – languages, and datasets. We provide time to edit (TE, i.e. hour:minutes), HTER, and the number of post-edited sentences (out of 250 per each gender).

of around 15 sentences each), *iii)* balancing at the sample level which block – F or M – they will post-edit first. A within-subject approach is ideal to distribute potential extraneous effects (e.g. participants' tendency to edit more or take longer) across F and M post-edits. Also, counterbalancing handles carryover effects such as *order* and *fatigue*[16] (Price et al., 2017). Crucially, to control for *familiarity* effects, we ensure that a participant never post-edits the same output twice across genders.

The design remains the same for all samples, but always involving different subjects, so as to ensure the generalization and replicability of our results.

Participants recruitment and task organization Experiments for en-it include data from both *i)* professional translators based on voluntary participation, and *ii)* paid professionals. We attested no significant difference between these conditions (for further details see Appendix C.2). For en-de/es, we exclusively relied on paid professionals. Experiments were allocated 50m (i.e. ∼10m instructions and ∼40m PE), which vastly ensured the sufficient time to complete the task.[17] The experiment with students was carried out as part of their school activities: we allocated double the time and included a warm-up phase to get acquainted with the PE task. No participant was informed of the scope of our study beforehand, and all recorded data are anonymous. For further information on the recruited participants and compensation see Appendix C.

Data collection and effort measures At the end of the process, for each sample of 250 source sentences we collect 500 post-edits (250 F and 250 M). We then measure the corresponding "feminine"/"masculine" effort for the *temporal* and *technical* dimension (Krings, 2001). Respectively, *i)* time to edit (TE) is recorded within Matecat for each output sentence,[18] whereas *ii)* the amount of edits is computed with HTER (Snover et al., 2006).

We frame the difference between feminine and masculine efforts (Δ) as our human-centered measure for gender-related quality of service disparities. We also compute statistical significance tests between F and M effort values. We use paired bootstrap resampling (Koehn, 2004) for HTER, and Wilcoxon (Rey and Neuhäuser, 2011)[19] for both HTER and TE, with p-value < 0.05. Tests were calculated for all results presented in the paper, and are all statistically significant.

4 Results

4.1 Multidataset Results

In Table 2 (top), we report the cumulative results for TE and the number of edits across genders for three en-it datasets. Consistently, though with variation across datasets, **our results confirm a significant effort difference across genders**.

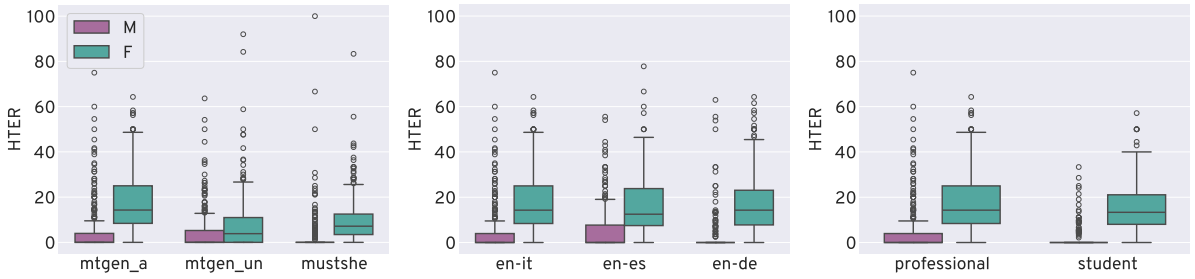
The unambiguous MTGEN-UN exhibits the lowest gap, attesting that, when source text provides gender cues, GT can better handle feminine and masculine gender in the target language. Still, even in this context, F post-editing amounts to a +36.3% and +58.3% increase (Δ*rel*), respectively for TE and HTER. For the ambiguous datasets, the gap clearly widens. This is particularly notable for MTGEN-A, which – compared to MUST-SHE – presents a higher distribution of gendered words

[16] For fatigue, we also only assign 30 sentences per subject.

[17] Based on industry standards, we estimated a PE speed of ∼25 words per minute.

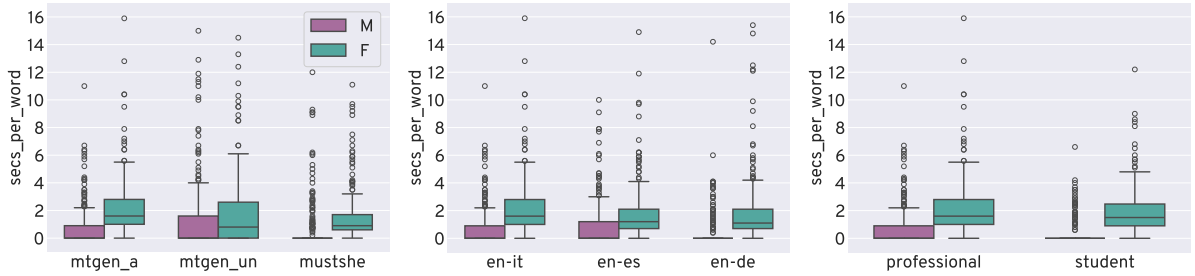
[18] Sentences that do not require any post-editing count as 0.

[19] The Wilcoxon result is computed using `scipy 1.13.1`: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html.



(a) Multidataset (b) Multilanguage (c) Multiuser

Figure 3: HTER distribution across post-edited sentences.



(a) Multidataset (b) Multilanguage (c) Multiuser

Figure 4: Seconds per source word distribution across post-edited sentences.

(see Table 1), which are also more prone to bias, i.e. professions. Compared to its M counterpart, F post-editing for this dataset requires around four times the effort both in time and number of edits.

Overall, effort distribution across post-edited sentences – Figure 3a for technical and 4a for temporal effort – attest that for the vast majority of M sentences, no post-editing at all was required. This mirrors the known GT tendency to masculine default (Piazzolla et al., 2023).

Henceforth, we focus on the particularly biased MTGEN-A sample[20] for multilanguage and multiuser comparisons.

4.2 Multilanguage Results

Moving onto multilanguage assessments with MTGEN-A, we attest that human-centered disparities are present also for en-de and en-es. Although cumulative results in Table 2 (center) show some variation for TE – especially for the masculine set – sentence-level distributions for both types of effort are highly comparable. In figure 3b, median HTERs are the same for en-de/it in the feminine set (14.3), and slightly lower for en-es (12.5). For masculine PE, the median HTER values are systematically 0, although the number of not edited sentences is visibly higher for en-de.[21] Median temporal efforts based on the number of source words per second are also very close, i.e. always 0 for M; whereas in the feminine PE we find 1.6 (en-it) 1.2 (en-es) 1.1 (en-de) – see Figure 4b. Overall, **differences in efforts based on gender generalize across the considered language pairs**.

4.3 Multiuser Results

As a last step, we confront the PE activity of professional translators (P) with less experienced high-school students (S). Cumulative results in Table 2 (bottom) show that **in the student condition gender gaps widen significantly**. More specifically, percentage differences for MTGEN-A en-it go from +177.6% (TE) and +201.8% (HTER) – assessed with professionals – up to respectively +329.8% and +636.3% for students. Quite surprisingly, and also confirmed by the distributions in Figures 3c and 4c, students are overall quicker, and edit less across both F and M.

We explain these results by the lower familiarity with both the English language and the PE task itself. In fact, based on observations during the experiments, also confirmed by manual revision of

[20] We choose MTGEN-A also to include a non-romance language (de), since MusT-SHE is only available for en-it/fr/es.

[21] Based on a manual analysis, this is due to a lower incidence of *preferential edits* (i.e. not gender-related), suggesting that post-editors perceived the en-de output as of high quality.

the collected post-edits, students did not engage with the improvement of the overall accuracy of the translation. Rather, they almost exclusively focused on adjusting gendered words.[22] Thus, to a certain extent, students' results allow us to isolate even more neatly the sole effect of gender bias in MT with our human-centered measurements, an issue that might be further amplified should lay users be involved in similar experiments.

5 Discussion

We found strong evidence for the human-centered impact of bias in MT, with a quality of service disparity that can disproportionately affect women. Such allocative harm is evident in the extra time and energy required for feminine gender translation. Note that our results are likely conservative, involving experienced users with high language proficiency. Indeed, in less controlled conditions, or among individuals with lower proficiency in either target or source language, such a negative impact would likely be even greater. Misgendered references may go unnoticed, propagating errors in texts and communications, or necessitating the use of external resources such as dictionaries to be fixed. Due to experimental constraints (§3.1), such a scenario remains open to future analyses. To better frame the implications of our findings, we conclude with two critical reflections. First, individuals might rely on third-party language services to translate their text, thus raising the question: *Can gender bias imply a differential in economic cost*? Second, while informative assessments that center users are crucial to guide the field forward, *are current automatic evaluations able to capture such human-centered disparities*?

Someone has to pay for the cost of gender bias. We explore the economic costs of F and M post-editing considering two stakeholders: *i)* a *final user*, who buys the PE text from *ii)* a third-party *translator*. As a case study, we analyze the three en-it datasets edited by professionals (§4.1) – using averaged HTER and source words shown in Table 3. Note that pricing in the language industry is complex (Lambert and Walker, 2022) and can be based on various parameters (Scansani and Mhedhbi, 2020; Cid, 2020). Here, we consider two common payment scenarios – i.e. `HTER-Rate` and `Word-Rate`. For both payments, we use a baseline word-rate of €0.09 per source word, reflecting best market prices for en-it (Inbox-Translation, 2023).

	HTER	src-W	**HTER€**	**Word€**
FEM	10.92	5629	202.63	177.30
MAS	4.60	5629	177.30	177.30

Table 3: Economic costs of feminine and masculine en-it PE. We provide pricing based on technical effort (HTER€) and on source text lenght (Word€).

`HTER Rate`: With this method, prices are adjusted based on the *actual* technical effort required to post-edit, with lower edit rates leading to lower costs, and vice versa. Following existing price schemes (Localization, 2022),[23] HTER below 10 is paid at 35% of the word rate (i.e. €0.0315 per word), whereas HTER between 10-20 is paid at 40% (i.e. €0.036 per word). Hence, and as shown in Table 3 (HTER€), feminine PE would cost more. While translators are compensated for the additional effort, such a financial burden will inevitably fall on the final user purchasing the F translation.

`Word Rate`: This pricing is based on source text length, where the cost per word is decided *a priori*. For PE tasks, the word-rates vary depending on the content or the language (Sarti et al., 2022).[24] For en-it data from a general domain such as ours, a 35% word rate could be paid. Given that – to the best of our knowledge – this type of pricing does not consider gendered content, the same word-rate would be indiscriminately applied to both feminine and masculine PE. Thus, as shown in Table 3 (Word€), a final user buying their translation would pay the same price, regardless of gender. However, this would place the financial cost on the translator, whose additional effort required for feminine PE would be underestimated and under-compensated.

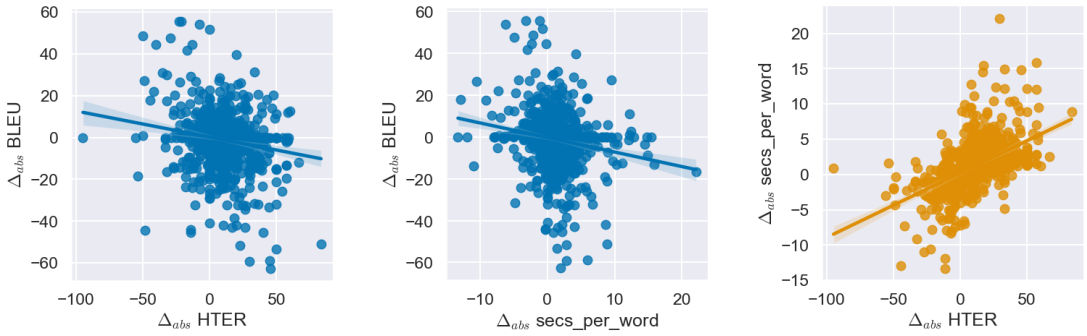
🔍 To sum up, this analysis shows that gender bias has an economic cost which can unfairly fall onto either one of the two PE stakeholders. Besides financial implications, unfair compensation could also invite less edits than necessary, thus compromising the quality of feminine PE. Analysing such potential quality-oriented implications is a crucial aspect for future research.

[22] Post-editing examples available in Appendix D.

[23] See Figure 8 in Appendix E.

[24] e.g. creative texts or certain languages are notably poorly handled by MT, thus corresponding to higher word-rates.

[25] Computed using `scipy 1.13.1`: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html



(a) Δ_{abs} HTER and BLEU (b) Δ_{abs} secs_per_word and BLEU (c) Δ_{abs} HTER and secs_per_word

Figure 5: Scatter plots with overlaid regression lines of the differences between F and M scores for all *datasets*, *languages* and *users*. Each point represents a sentence-level difference. The correlation between the different metrics is measured with the Pearson r coefficient, and all results are statistically significant (p-value < 0.05).[25]

Automatic bias measurements do not reliably correlate with human-centered measures. Methods to quantify bias are key to much research that seeks to monitor the creation of equitable technologies (Dev et al., 2022). In this context, growing evidence underscored how *intrinsic* metrics—focusing on models' representations—might not be a reliable bias indicator in downstream, real-world tasks, as assessed with *extrinsic* ones—focusing on models' output (Jin et al., 2021; Goldfarb-Tarrant et al., 2021; Cao et al., 2022; Orgad and Belinkov, 2022). Arguably, however, even extrinsic measures are model-centric (§2), and only assumed to reflect more reliably the downstream harms to individuals. We verify this assumption by comparing our human-centered measures of differential effort with the automatic evaluations associated with MT-GenEval and MuST-SHE (§3.1). As in the original papers, we use the set of contrastive F/M target references[26] to compute gender-related performance differences with BLEU[27] (Papineni et al., 2002), i.e. $BLEU_F - BLEU_M$. Scatter plots of the automatic (i.e., BLEU score) and human-centric metrics (i.e., HTER and TE) differences, in absolute values, are reported in Figure 5. We provide aggregate results for all languages, datasets and users.[28]

Looking at our results, we notice a Pearson-r of -0.19 for Δ_{abs} HTER and Δ_{abs} BLEU (Figure 5a), and -0.18 for Δ_{abs} secs_per_word and Δ_{abs} BLEU (Figure 5b). The negative correlation is expected since, while for BLEU the higher the better, the opposite is true for both HTER and TE. Still, the results clearly indicate that both temporal and technical efforts are in *weak* correlation (Schober et al., 2018) with automatic scores. On the one hand, it is known that time measures may not always have a linear relationship with textual differences measured by automatic metrics (Tatsumi, 2009; Macken et al., 2020), e.g. even small edits can require a high cognitive load and more time. On the other hand, given that both BLEU and HTER capture surface modifications and quantity of edits, their weak correlations are particularly noteworthy.[29] A *moderate* correlation (Person-r 0.54) is found only between the human-centered measures HTER and TE. As observed in Figure 5c, the higher the number of edits, the more time required.

🔍Overall, our results suggest that existing *model-centric* measures of gender bias in MT might not reliably reflect the downstream harms to users. While the contrastive evaluation approaches explored here have been used to reveal gender gaps (Bentivogli et al., 2020; Currey et al., 2022), they do not correlate with or accurately reflect the magnitude of disparities found through more concrete, human-centered measures.[30] To ensure that advancements in the field prioritize impacted individuals, future research should explore both the metrics and the data used to compute them (Orgad and Belinkov, 2022). This includes investigating how automatic

[26] e.g. *I am a **friend*** → M-es: *soy **amigo***, F-es: *soy **amiga***.

[27] `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp`

[28] We also compute separate statistics for each sample, and with other metrics (COMET-22 and TER). The hereby discussed trends are confirmed. See Appendix F.3. Details on the automatic metrics computation are provided in Appendix B.3.

[29] As a matter of fact, additional results reported in Appendix F.3.1 show that COMET – despite its attested higher degree of correlation with human assessments for overall MT quality – exhibit a *very weak* correlation with human-centered measures of bias.

[30] See also the contrastive, automatic bias scores reported in Table 9 in Appendix F.2.

metrics relate to human-centered measures and how they can be translated into more transparent, user-relevant evaluations (Liao and Xiao, 2023).

6 Conclusion

From cars' safety measures more effective for men, (Ulfarsson and Mannering, 2004), to virtual reality headsets that are too big to wear (Robertson, 2016), evidence of social and technological advances being less functional for women, or even harmful, abounds (Criado-Perez, 2019). While it is increasingly acknowledged that also language technologies can contribute to broader patterns of gender bias, still little is known about their tangible impact on people. Our study represents a novel effort to empirically examine the implications of gender bias in MT with a human-centered perspective. Previous research has often inferred the downstream impact of bias based on automatic, model-centric scores. In contrast, we provide concrete empirical evidence showing that gender bias in MT leads to tangible service disparities, which can disproportionately affect women. Also, we quantify these disparities using measures that are more meaningful to impacted individuals, such as workload and economic costs.Our study advocates for an understanding of bias and its impact that centers on the actual users of this technology to guide the field. To this aim, we make our collected data and metadata publicly available for future studies on the topic.

7 Limitations

Experimental construct. To foreground the impact of gender bias, our study employs datasets that include at least one gender translation phenomenon per sentence. While these data more closely simulate our scenarios of interest like the translation of biographies or CVs – where human gender references are common – in other contexts such phenomena may be more sparse. Despite potential variations in bias magnitude across different types of text, however, our findings would not change: gender bias would simply be more difficult to detect. Also, while women would likely be the main target of bias-related issues, the found costs and disparities could actually fall on anyone attempting to use feminine expressions, e.g. current attempts to avoid "masculine default" expressions for generic referents, and rather rely on generically intended feminine forms (Merkel et al., 2017). Overall, since we rely on two widely recognized MT gender bias benchmarks, the density of gender phenomena in our study is actually the same density that is automatically evaluated with current bias metrics.

MT system. We prioritize the type of languages, participants and datasets as variables of interest over including MT system comparisons. This choice is also motivated by the fact that gender bias is a widespread issue in generic MT models (Savoldi et al., 2023), and attested with limited variation in commercial MT applications (Rescigno et al., 2020a; Troles and Schmid, 2021). Despite being a commercial system that can limit reproducibility, we pick Google Translate as it represents one of the most used MT engines by the public. Also, we exclude experiments based on instruction-tuned models such as ChatGPT given that the language industry as well as end-users mostly rely on standard MT for core translation tasks (Fishkin, 2023).[31] Also, while "gender-specific translation prompts" could help in the future (Sánchez et al., 2024), they are currently less realistic as they require users to craft them and – before that – to be aware of the presence, and thus the need to control for gender bias in MT.

Languages. Our study focuses on the translation of English sentences into grammatical gender languages that distinguish between masculine and feminine forms to express the gender of human referents (Gygax et al., 2019). As such, we should be cautious in generalizing our findings to languages that mark gender differently, or not at all. Also, we focus on three language pairs (en-it/es/de) that are well-supported by current MT. Hence, it remains open to future investigation if the human-centered impact of gender bias could vary for languages with overall lower MT quality.

ACL query. The review of prior work on gender (bias) in MT considers only literature from the ACL Anthology. While searching other sources could have enriched our analysis, the Anthology represents the main historical reference point in the field and serves as a good and sufficiently comprehensive litmus test for examining the main trends in NLP.

Finally, we discuss the limitations of our binary gender setup in the upcoming section.

8 Ethical Statement

Our study is limited to binary, *feminine* and *masculine*, linguistic expressions of gender. Indeed,

[31] This was also confirmed by our study participants.

this choice, as well as the use of gender as a variable, warrants some ethical reflections. First of all, we stress that – by working on binary linguistic forms – we do not imply a binary vision on the extra-linguistic reality of gender and gender identities (D'ignazio and Klein, 2023). The motivation behind our binary design was to ensure comparable conditions between gendered post-edits. While non-binary forms and neutral expressions are increasingly emerging in the target languages of our study (Bonnin and Coronel, 2021; Mirabella et al., 2024; Daems, 2023; Piergentili et al., 2024), the attitude towards these forms, as well as their level of usage can widely vary among speakers (Bonnin and Coronel, 2021; Piergentili et al., 2023b). Given that non-binary and neutral expressions are not standardized like masculine and feminine terms, incorporating them would necessitate controlling for participants' prior familiarity with these forms. This additional variable could introduce cognitive effort complicating the measurement of post-editing effort. By focusing solely on binary gender expressions, we aim to isolate the effort and costs that are exclusively due to the system's bias without confounding it with the potential cognitive load associated with producing non binary language (Lardelli and Gromann, 2023; Paolucci et al., 2023). While by all means of utmost importance for future research, we were not able for the time being to also account for this cognitive dimension, which would have required additional tools and costs.

Acknowledgements

Beatrice Savoldi is supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. The work presented in this paper is also funded by the Horizon Europe research and innovation programme, under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People), and the ERC Consolidator Grant No 101086819. This research was made possible by the participation of several bodies and individuals that took part in our human-centered study. We thank the Directorate-General for Translation (DGT) of the European Commission and the DGT translators that kindly agreed to participate in the activity for en-it. We also tank the independent professional translators that worked with us across all language pairs, as well as the high-school students that participated in our laboratories, thus contributing to the multiuser experiments. Finally, we thank Jasmijn Bastings for the insightful discussion on and contribution to the gender bias papers' review.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022a. The Arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France. European Language Resources Association.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022b. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Bashar Alhafni, Ossama Obeid, and Nizar Habash. 2023. The user-aware Arabic gender rewriter. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 3–11, Tampere, Finland. European Association for Machine Translation.

Sultan Alrowili and Vijay Shanker. 2022. Generative approach for gender-rewriting task with ArabicT5. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 491–495, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada. Association for Computational Linguistics.

Giuseppe Attanasio, Flor Miriam 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.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *SIGCIS Conference*, Philadelphia, Pennsylvania.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing*

Workshop, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Rachel Bawden. 2017. Machine translation of speech-like texts: Strategies for the inclusion of context. In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, pages 1–14, Orléans, France. ATALA.

Rachel Bawden, Guillaume Wisniewski, and Hélène Maynard. 2016. Investigating gender adaptation for speech translation. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Posters)*, pages 490–497, Paris, France. AFCP - ATALA.

Emily M. Bender. 2019. A Typology of Ethical Risks in Language Technology with an Eye Towards where Transparent Documentation might help. In *CRAASH. The future of Artificial Intelligence: Language, Ethics, Technology*, Cambridge, UK.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Juan Eduardo Bonnin and Alejandro Anibal Coronel. 2021. Attitudes toward gender-neutral spanish: Acceptability and adoptability. *Frontiers in Sociology*, 6.

Vicent Briva-Iglesias, Sharon O'Brien, and Benjamin R Cowan. 2023. The impact of traditional and interactive post-editing on machine translation user experience, quality, and productivity. *Translation, Cognition & Behavior*, 6(1):60–86.

Lena Cabrera and Jan Niehues. 2023. Gender lost in translation: How bridging the gap between languages affects gender bias in zero-shot multilingual translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 25–35, Tampere, Finland. European Association for Machine Translation.

Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2350–2361, Online. Association for Computational Linguistics.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Sheila Castilho. 2022. How much context span is enough? examining context-related issues for document-level MT. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3017–3025, Marseille, France. European Language Resources Association.

Sheila Castilho, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. 2021. DELA corpus - a document-level corpus annotated with context-related issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 566–577, Online. Association for Computational Linguistics.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? what about a GPT model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.

Amanda Cercas Curry, Judy Robertson, and Verena Rieser. 2020. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Barcelona, Spain (Online). Association for Computational Linguistics.

Gary Charness, Uri Gneezy, and Michael A Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, 81(1):1–8.

Jennifer Chien and David Danks. 2024. Beyond behaviorist representational harms: A plan for measurement and mitigation. *arXiv preprint arXiv:2402.01705*.

Won Ik Cho, 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. Association for Computational Linguistics.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara Ginovart Cid. 2020. Report about a survey-based research on machine translation post-editing: Common ground and gaps between lscs, linguists, and trainers.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ander Corral and Xabier Saralegi. 2022. Gender bias mitigation for NMT involving genderless languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 165–176, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023a. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023b. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2022. Evaluating gender bias in speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147, Marseille, France. European Language Resources Association.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France. European Language Resources Association.

Caroline Criado-Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage Books.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Joke Daems and Janiça Hackenbuchner. 2022. DeBiasByUs: Raising awareness and creating a database of MT bias. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 289–290, Ghent, Belgium. 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.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2022. On measures of biases and harms in nlp. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267.

Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism*.

Tu Anh Dinh and Jan Niehues. 2023. Perturbation-based QE: An explainable, unsupervised word-level quality estimation method for blackbox machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 59–71, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Carlos Escolano, Graciela Ojeda, Christine Basta, and Marta R. Costa-jussa. 2021. Multi-task learning for improving gender accuracy in neural machine translation. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 12–17, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of*

the First Workshop on Gender Bias in Natural Language Processing, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Rand Fishkin. 2023. We analyzed millions of chatgpt user sessions: Visits are down 29 https://sparktoro.com/blog/we-analyzed-millions-of-chatgpt-user-sessions-visits-are-down-29-since-may-programming-assistance-is-30-of-use/. Accessed: 2024-06-14.

Dennis Fucci, Marco Gaido, Sara Papi, Mauro Cettolo, Matteo Negri, and Luisa Bentivogli. 2023. Integrating language models into direct speech translation: An inference-time solution to control gender inflection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11505–11517, Singapore. Association for Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. How to split: the effect of word segmentation on gender bias in speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online. Association for Computational Linguistics.

Harritxu Gete and Thierry Etchegoyhen. 2023. An evaluation of source factors in concatenation-based context-aware neural machine translation. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 399–407, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria, and Maite Martin. 2022. TANDO: A corpus for document-level machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France. European Language Resources Association.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare Voss, Marine Carpuat, and Hal Daumé III. 2023. What else do i need to know? the effect of background information on users' reliance on qa systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3330.

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.

Ana Guerberof-Arenas and Joss Moorkens. 2023. Ethics and machine translation: The end user perspective. In *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 113–133. Springer.

Pascal Mark Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology*, 10.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Bar Iluz, Tomasz Limisiewicz, Gabriel Stanovsky, and David Mareček. 2023. Exploring the impact of training data distribution and subword tokenization on gender bias in machine translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 885–896, Nusa Dua, Bali. Association for Computational Linguistics.

Inbox-Translation. 2023. [research] freelance translator survey 2023: See how you compare to colleagues when it comes to rates, cpd, and business practices. Last updated 11 December 2023.

Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating gender augmented data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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.

Hans P Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.

Joseph Lambert and Callum Walker. 2022. Because we're worth it: Disentangling freelance translation, status, and rate-setting in the united kingdom. *Translation Spaces*, 11(2):277–302.

Manuel Lardelli and Dagmar Gromann. 2023. 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.

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.

Ngoc Tan Le, Oussama Hansal, and Fatiha Sadat. 2023. Challenges and issue of gender bias in under-represented languages: An empirical study on Inuktitut-English NMT. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 89–97, Remote. Association for Computational Linguistics.

Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. 2023. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16825–16839, Singapore. Association for Computational Linguistics.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Danni Liu and Jan Niehues. 2024. How transferable are attribute controllers on pretrained multilingual translation models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 334–348, St. Julian's, Malta. Association for Computational Linguistics.

Localization. 2022. A Fair and Workable Remuneration Model for Machine Translation Post-Editing Effort? (The Story of TER).

Tianshuai Lu, Noëmi Aepli, and Annette Rios. 2023. Reducing gender bias in NMT with FUDGE. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 61–69, Tampere, Finland. European Association for Machine Translation.

Lieve Macken, Daniel Prou, and Arda Tezcan. 2020. Quantifying the effect of machine translation in a high-quality human translation production process. *Informatics*, 7(2).

Michal Měchura. 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. Association for Computational Linguistics.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4:725911.

Elisa Merkel, Cristina Cacciari, Martina Faralli, and Anne Maass. 2017. It only needs one man or can mixed groups be described by feminine generics? *It only needs one man or can mixed groups be described by feminine generics?*, pages 45–59.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Marta Mirabella, Claudia Mazzuca, Chiara De Livio, Bianca Di Giannantonio, Fau Rosati, Maric Martin Lorusso, Vittorio Lingiardi, Anna M Borghi, and Guido Giovanardi. 2024. The role of language in nonbinary identity construction: Gender words matter. *Psychology of Sexual Orientation and Gender Diversity*.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. What do compressed multilingual machine translation models forget? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4308–4329, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

Mary Nurminen and Niko Papula. 2018. Gist MT users: A snapshot of the use and users of one online MT tool. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 219–228, Alicante, Spain.

Joseph Olive. 2005. Global autonomous language exploitation (gale). *DARPA/IPTO Proposer Information Pamphlet*.

Parmy Olson. 2018. The Algorithm That Helped Google Translate Become Sexist. `https://bit.ly/olson_google_sexist`. Accessed: 2023-06-20.

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Sharon O'Brien. 2022. How to deal with errors in machine translation: Postediting. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 18:105.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Marjolene Paulo, Vera Cabarrão, Helena Moniz, Miguel Menezes, Rachel Grewcock, and Eduardo Farah. 2023. Context-aware and gender-neutral translation memories. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 437–444, Tampere, Finland. European Association for Machine Translation.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *HERMES - Journal of Language and Communication in Business*, (63):209–225.

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.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large

language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Jeff Pitman. 2021. Google translate: One billion installs, one billion stories. https://blog.google/products/translate/new-features-make-translate-more-accessible-for-its-1-billion-users/. Engineering Manager, Google Translate.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Maja Popovic and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Paul C Price, RS Jhangiani, IA Chiang, Dana C Leighton, and Carrie Cuttler. 2017. Research methods in psychology (3rd american edition). *Washington: PressBooksPublications*.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Krithika Ramesh, 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. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Argentina Anna Rescigno, Vanmassenhove Eva, Johanna Monti, Way Andy, et al. 2020a. A case study of natural gender phenomena in translation-a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *CEUR Workshop Proceedings*, pages 359–364. AILC-Associazione Italiana di Linguistica Computazionale.

Argentina Anna Rescigno, Johanna Monti, Andy Way, and Eva Vanmassenhove. 2020b. A case study of natural gender phenomena in translation: A comparison of Google Translate, bing Microsoft translator and DeepL for English to Italian, French and Spanish. In *Workshop on the Impact of Machine Translation (iMpacT 2020)*, pages 62–90, Virtual. Association for Machine Translation in the Americas.

Denise Rey and Markus Neuhäuser. 2011. *Wilcoxon-Signed-Rank Test*, pages 1658–1659. Springer Berlin Heidelberg, Berlin, Heidelberg.

Adi Robertson. 2016. Building for virtual reality? don't forget about women. *The Verge*.

Samantha Robertson, Wesley Hanwen Deng, Timnit Gebru, Margaret Mitchell, Daniel J Liebling, Michal Lahav, Katherine Heller, Mark Díaz, Samy Bengio, and Niloufar Salehi. 2021. Three directions for the design of human-centered machine translation. *Google Research*.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023a. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023b. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada. Association for Computational Linguistics.

Danielle Saunders 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*, pages 7724–7736, Online. Association for Computational Linguistics.

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.

Danielle Saunders, 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). Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland. Association for Computational Linguistics.

Beatrice Savoldi, 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.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022a. On the dynamics of gender learning in speech translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–111, Seattle, Washington. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022b. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland. Association for Computational Linguistics.

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.

Beatrice Savoldi, 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)*, pages 256–267, St. Julian's, Malta. Association for Computational Linguistics.

Randy Scansani and Lamis Mhedhbi. 2020. How do lsps compute mt discounts? presenting a company's pipeline and its use. In *Proceedings of the 22nd annual conference of the European Association for Machine Translation*, pages 393–401.

Patrick Schober, Christa Boer, and Lothar A. Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126:1763–1768.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1968–1984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Pushpdeep Singh. 2023. Gender inflected or bias inflicted: On using grammatical gender cues for bias evaluation in machine translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 17–23, Nusa Dua, Bali. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Celia Soler Uguet, Fred Bane, Mahmoud Aymo, João Pedro Fernandes Torres, Anna Zaretskaya, and

Tània Blanch Miró. 2023. Enhancing gender representation in neural machine translation: A comparative analysis of annotating strategies for English-Spanish and English-Polish language pairs. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 171–172, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Ingason. 2022. Mean machine translations: On gender bias in Icelandic machine translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3113–3121, Marseille, France. European Language Resources Association.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in Natural Language Processing. *arXiv preprint arXiv:2112.14168*.

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.

Dario Stojanovski, Benno Krojer, Denis Peskov, and Alexander Fraser. 2020. ContraCAT: Contrastive coreference analytical templates for machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4732–4749, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tony Sun, 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. Association for Computational Linguistics.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Gender-specific machine translation with large language models.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of Machine Translation Summit XII: Posters*.

Bertille Triboulet and Pierrette Bouillon. 2023. Evaluating the impact of stereotypes and language combinations on gender bias occurrence in NMT generic systems. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 62–70, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jonas-Dario Troles 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*, pages 531–541, Online. Association for Computational Linguistics.

Gudmundur F Ulfarsson and Fred L Mannering. 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident Analysis & Prevention*, 36(2):135–147.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2016. Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2571–2581, Osaka, Japan. The COLING 2016 Organizing Committee.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021a. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove 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. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021b. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. A rewriting approach for gender inclusivity in Portuguese. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759, Singapore. Association for Computational Linguistics.

Sebastian Vincent. 2021. Towards personalised and document-level machine translation of dialogue. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 137–147, Online. Association for Computational Linguistics.

Sebastian Vincent, Robert Flynn, and Carolina Scarton. 2023. MTCue: Learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8210–8226, Toronto, Canada. Association for Computational Linguistics.

Sebastian T. Vincent, Loïc Barrault, and Carolina Scarton. 2022. Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 121–130, Ghent, Belgium. European Association for Machine Translation.

Eric Peter Wairagala, Jonathan Mukiibi, Jeremy Francis Tusubira, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, and Ivan Ssenkungu. 2022. Gender bias evaluation in Luganda-English machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 274–286, Orlando, USA. Association for Machine Translation in the Americas.

Angelina Wang, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. 2024. Measuring machine learning harms from stereotypes: requires understanding who is being harmed by which errors in what ways.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi, and Zhaopeng Tu. 2023. A survey on zero pronoun translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3325–3339, Toronto, Canada. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.

Guillaume Wisniewski, Lichao Zhu, Nicolas Bailler, and François Yvon. 2021a. Screening gender transfer in neural machine translation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 311–321, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2021b. Biais de genre dans un système de traduction automatiqueneuronale : une étude préliminaire (gender bias in neural translation : a preliminary study). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 11–25, Lille, France. ATALA.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2022a. Analyzing gender translation errors to identify information flows between the encoder and decoder of a NMT system. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 153–163, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2022b. Biais de genre dans un système de traduction automatique neuronale : une étude des mécanismes de transfert cross-langue [gender bias in a neural machine translation system: a study of crosslingual transfer mechanisms]. In *Traitement Automatique des Langues, Volume 63, Numéro 1 : Varia [Varia]*, pages 37–61, France. ATALA (Association pour le Traitement Automatique des Langues).

Lichao Zhu, Guillaume Wisniewski, Nicolas Ballier, and François Yvon. 2022. Flux d'informations dans les systèmes encodeur-décodeur. application à l'explication des biais de genre dans les systèmes de traduction automatique. (information flow in encoder-decoder systems applied to the explanation of gender bias in machine translation systems). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, pages 10–18, Avignon, France. ATALA.

A Details on ACL Anthology Search

Our ACL search is based on the combination of keywords displayed in Table 4. Note that we also include terms such as "rewriters"', which several works apply to the output of MT models as a bias mitigation strategy to offer double feminine and masculine outputs. To avoid retrieving unrelated works that only marginally mentioned *MT* or *gender* in the main body, the searches parsed only the title and abstract of the queried papers.

	Keywords	# Papers
main	translation, NMT, MT, rewriter	
+	gender	138
+	bias	113
+ +	manual, survey, human, participant, expert, qualitative, user, people, annotat*, linguist, professional	96

Table 4: Number of search results for each specific keyword combinations on the ACL anthology. In total, we find 347 results comprising 251 unique articles, of which 146 were discarded as out of scope.

Manual selection We retrieved a total of 251 unique articles. Of those, we discarded all unrelated papers that refer to e.g. *inductive bias*, *bias lenght*, or "*translation*", but not in relation to the MT task. We thus arrive at a total of 105 papers. The whole selection was carried out manually, and we annotated both papers that that matched the query focusing on human assessment as well as those that did not, so not ensure not to overlook any paper involving humans. We defined the papers to be considered *in-scope* as follows:

- **MT application**: we only keep those works that primarily focused on MT, whereas those that relied on MT as an intermediate tool (e.g. to automatically translate a set of data) are discarded.[32]
- **Modality**: while limited in number, we keep also MT beyond the text-to-text modality.
- **Gender (bias)**: we include in our selection all works that focus on gender translation in the context of human entities. This includes works that do not explicitly engage with the notion of social bias – especially prior to 2018. Papers more broadly addressing gender fairness and inclusivity are also included.

The full list of extracted papers that made our final selection is provided below.

The first *in-scope* papers date back to 2016, whereas the latest two are from 2024. As of April, in fact, only few papers had been included in the Anthology. These 2024 papers are thus not shown in the figure to avoid incomplete views on approaches for the present year.

[32]Two papers Daems (2023); Paolucci et al. (2023) that focused on gender (bias) translation, but did not focus on MT were discarded, too.

MT gender bias papers, no human assessment van der Wees et al. (2016); Rabinovich et al. (2017); Bawden (2017); Popel (2018); Michel and Neubig (2018); Vanmassenhove et al. (2018); Moryossef et al. (2019); Escudé Font and Costa-jussà (2019); Cho et al. (2019); Habash et al. (2019); Stafanovičs et al. (2020); Basta et al. (2020); Costa-jussà and de Jorge (2020); Saunders et al. (2020); Gonen and Webster (2020); Stojanovski et al. (2020); Rescigno et al. (2020b); Bentivogli et al. (2020); Saunders and Byrne (2020); Hovy et al. (2020); González et al. (2020); Costa-jussà et al. (2020); Troles and Schmid (2021); Savoldi et al. (2021); Wisniewski et al. (2021b); Ciora et al. (2021); Escolano et al. (2021); Ramesh et al. (2021); Levy et al. (2021); Gaido et al. (2021); Vanmassenhove et al. (2021b); Vincent (2021); Renduchintala et al. (2021); Castilho et al. (2021); Wisniewski et al. (2021a); Vanmassenhove and Monti (2021); Wisniewski et al. (2022b); Costa-jussà et al. (2022); Castilho (2022); Gete et al. (2022); Sólmundsdóttir et al. (2022); Savoldi et al. (2022a); Měchura (2022); Corral and Saralegi (2022); Mohammadshahi et al. (2022); Saunders et al. (2022); Karpinska et al. (2022); Zhu et al. (2022); Sharma et al. (2022); Wisniewski et al. (2022a); Vincent et al. (2022); Wang et al. (2022); Renduchintala and Williams (2022); Alrowili and Shanker (2022); Alhafni et al. (2022a); Gete and Etchegoyhen (2023); Dinh and Niehues (2023); Singh (2023); Iluz et al. (2023); Alhafni et al. (2023); Sandoval et al. (2023); Wicks and Post (2023); Piergentili et al. (2023a); Saunders and Olsen (2023); Kostikova et al. (2023); Cabrera and Niehues (2023); Fucci et al. (2023); Lu et al. (2023); Castilho et al. (2023); Paulo et al. (2023); Le et al. (2023); Sarti et al. (2023b); Vincent et al. (2023); Costa-jussà et al. (2023a); Attanasio et al. (2023); Lee et al. (2023); Wang et al. (2023); Veloso et al. (2023); Sarti et al. (2023a)

MT gender bias papers, manual evaluation Bawden et al. (2016); Stanovsky et al. (2019); Gaido et al. (2020); Kocmi et al. (2020); Caglayan et al. (2020); Choubey et al. (2021); Popović (2021); Jain et al. (2021); Vamvas and Sennrich (2021); Vanmassenhove et al. (2021a); Currey et al. (2022); Wairagala et al. (2022); Savoldi et al. (2022b); Alhafni et al. (2022b); Savoldi et al.

(2023); Triboulet and Bouillon (2023); Costa-jussà et al. (2023b); Soler Uguet et al. (2023); Savoldi et al. (2024); Liu and Niehues (2024);

MT gender bias papers, survey Daems and Hackenbuchner (2022) (Lardelli and Gromann, 2023); Piergentili et al. (2023b); Lauscher et al. (2023); Amrhein et al. (2023)

MT gender bias papers, participatory Gromann et al. (2023)

B Experimental details

B.1 Data Details

Here we provide additional information concerning the selection of the data used in our experiments (§B.1.1). Also, some minor corrections were made on the MTGEN-A reference translation (§B.1.2).

B.1.1 Data selection

MTGenEval-A selection The 250 sentences used in our en-it experiments represent a randomly selected sample of the "ambiguous" section of the original MTGenEval dataset (Currey et al., 2022). For the multilanguage experiments, we also maximize the overlap between en-it/es/de subsets. Overall, we retrieve 76 sentences which are common across all languages, whereas the remaining are randomly extracted within each monolingual portion of the original dataset.

MTGenEval-UN selection The MTGEN-UN sample used in our experiments was randomly extracted from the "unambiguous" section of the original MTGenEval corpus. Note that, by being a subset with unambiguous gender in the English source, for this sample we extract 250 *pairs* of sentences, for a total of 500. To exemplify, each pair corresponds to *i)* a *feminine* <source-target> segment (e.g. en: "Sarandon has **appeared** in two episodes of The Simpsons, once as herself and...", it: "Sarandon è **apparsa** in due episodi dei Simpson, una volta interpretando se **stessa**..."), and *ii)* a *masculine* <source-target> segment (e.g. en: "Sarandon has **appeared** in two episodes of The Simpsons, once as himself and...", it: "Sarandon è **apparso** in due episodi dei Simpson, una volta interpretando se **stesso**..."). We automatically translate with GT the total 500 English sentences and create the corresponding feminine and masculine samples of 250 sentences each to be post-edited.

MuST-SHE selection For MUST-SHE, which by design contains an higher variety of gender phenomena for several parts of speech we relied on preliminary filters to ensure a less noisy experimental environment. Namely, we excluded sentences that in the original dataset are annotated as "FREEREF", and for which the human reference translation is known to be quite creative and less literal. Also, prior work based on this dataset has shown that – due to its higher variability – a good amount of gendered words available in the reference translation might not be actually generated in the MT output for a range of reasons, i.e. errors, synonyms etc (Savoldi et al., 2022b). Thus, first we translated the whole corpus with Google Translate. Then, we only retained those sentences where the MT output contained at least one gendered word annotated in the corresponding reference translations. To do so, we relied on the *coverage* evaluation script[33] made available with the corpus. Overall, these filters ensured *i)* the presence of gender phenomena to revise during the PE task, *ii)* less creative reference translation that eased more reliable assessments with automatic metrics. The final 250 sentences were randomly extracted from this pre-filtered MuST-SHE subset.

B.1.2 MTGenEval-A reference translations

For MTGEN-A, we find that for some English sentences not all ambiguous human entities are translated with masculine or feminine gender in the corresponding reference of the M/F contrastive pair. We thus manually revised all reference translations for for the 3 en-it/es/de datasets. This is necessary to align the results of our PE activity – where *all* entities whose gender is ambiguous in English are post-edited either as masculine or feminine – with the automatic bias evaluation method presented in Section 5, which is based on the reference translations. To exemplify, see the following en-es segment:

src-en: **The doctor** and **some** of **the** patients had signed off to purchase it
tgt-es-F: **La doctora** y **algunos** de **los** pacientes se habían apuntado para comprarlo.
tgt-es-M: **El doctor** y **algunos** de **los** pacientes se habían apuntado para comprarlo.

While "doctor" is respectively translated as masculine or feminine in the corresponding references,

[33] https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/gender/mustshe_gender_accuracy.py

the equally ambiguos "some of the patients" is not, and rather remains masculine in both references. To fix these instances, for each of the 250 source sentences included in the en-it, en-es and en-de datasets, we manually revised both reference translations.

This was carried out by a linguist with expertise in all languages pairs. Overall, 40 segments were modified for en-it, 15 for en-es, and 28 for en-de.

B.2 Matecat tool and settings

To work in Matecat,[34] we created two separate projects for each dataset: one for the feminine setting and one for the masculine setting. For each project, we followed the same procedure. We uploaded the input English text and created a corresponding dedicated Translation Memory (TMX). The TMX contains the translations produced by GT, which are shown to the translators as suggestions to post-edit. Crucially, we ensured our settings as follows: *i)* each translator had access to the dedicated TMX in a "lookup-only" mode, meaning that they could not update it with their post-edits – which would have otherwise become visible to the other translators and make the experiment ineffective; also, *ii)* the general Matecat TMX was disabled, so as to avoid that translators had access to additional suggestions other than the GT outputs.; then, *iii)* to ensure that the Matecat tool would maintain the original sentence division of the dataset, we activated the *paragraph* setting, which does not re-segment the input text. Finally, each M/F project was split into sub-projects of around 15 sentences each to be assigned to participants (14 splits for MTGEN-A, 16 for MTGEN-UN, and 16 for MUST-SHE). Each participant received two links to work on both an M and an F sub-project, for a total of around 30 sentences to post-edit.

B.3 Automatic Metrics

The automatic metrics used to evaluate translation quality are BLEU, (Papineni et al., 2002), based on n-gram matching, TER (Olive, 2005), based on edit rates, and the neural-based COMET (Rei et al., 2020). BLEU and TER are computed with the well-established tool for evaluating machine translation outputs, sacrebleu v2.4.0 (Post, 2018).[35] COMET is computed using the official GitHub repository[36] with the `Unbabel/wmt22-comet-da`[37] model.

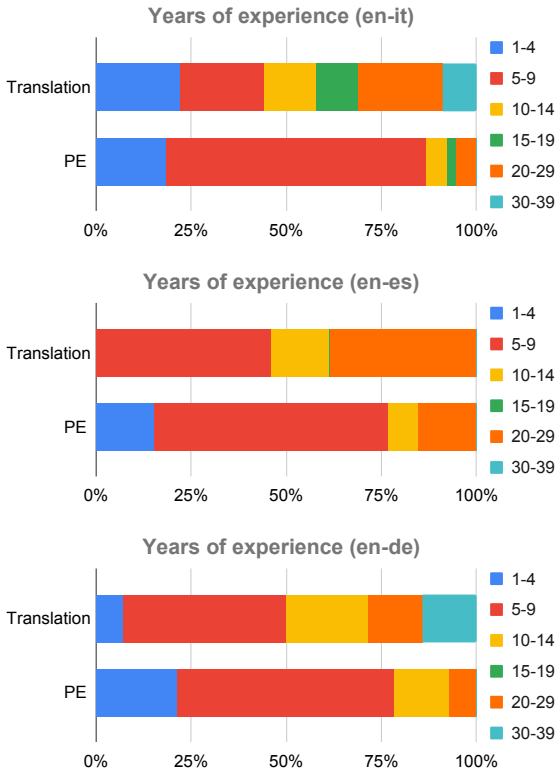


Figure 6: Professional translators' years of experience as translators, and as MT post-editors. Results are shown for each language pair.

C Study participants

We relied on two types of participants in our experiments: *professional translators* and *high school students*. As for translators, the experiment include professionals who participated on a voluntary basis as well as paid professionals. To ensure comparability, we replicated the same settings and used the same guidelines across all conditions. For students, we added a warm-up phase to introduce them to MT, the PE task, and the Matecat tool.

All the experiments were agreed upon with all participants. The **privacy protection** of the involved participants is guaranteed by the complete anonymity of the whole collected data, which make it impossible to identify the involved subjects.

C.1 Recruitment and Task organization

Professional translators (volunteers) For en-it, a first round of experiments was carried out with professional translators from the European

[34] https://www.matecat.com/

[35] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp

[36] https://github.com/Unbabel/COMET

[37] https://huggingface.co/Unbabel/wmt22-comet-da

Commission, Directorate-General for Translation, Italian-language Department. These participated on a voluntary basis as part of an educational lab held by the authors of this paper. As such, no compensation was involved.

To carry out experiments on MTGEN-A, MTGEN-UN, and MUST-SHE, we needed data from 14 + 16 + 16 participants, respectively, for a total of 46 participants. However, eventually 22 blocks of sentences (corresponding to the activity of 11 participants) were not carried out or completed. This was due to several reasons: some expected participants were absent, others experienced internet connection problems that hindered them to properly carry out the PE activity, and one participant decided not to take part in the experiment. Thus, in order to complete our data collection, we resorted to paid professional translators.

Professional translators (paid) The remaining en-it data and all en-es and en-de data were post-edited by paid professionals, who were recruited via a translation agency. The only eligibility criterion we required was that the en-* pair assigned to them represented one of their main language direction in their professional work, and that they were native speakers of the target language (i.e. the same working condition of volunteers). The experiments where carried out via online meetings, in groups of around 8 translators. To avoid introducing any confounding effect that could influence their PE work, all post-editors were requested to remain in the meeting for its entire duration of 50 minutes, and compensation was time-based. The total cost (translation agency recruitment and translator's work) amounted to €50 per post-editor, taxes excluded.

The similarities of the work carried out by the two types of professional translators, verified as discussed in Appendix C.2, allowed us to merge all en-it data coming from professionals and carry out aggregated dataset-level analyses.

Students (volunteers) The activity of the students was carried out during a laboratory as a part of their school activity. These students were from a school offering a foreign language specialization, thus ensuring that they had a good (B2 level) proficiency in English. They were all part of the same class, attending the penultimate year of high school. All the activities were allowed under the consensus of their school supervisor and under the supervision of their regular teachers. For this task setting, we also included a warm-up phase to introduce the students to MT, the PE task, and the Matecat tool before starting the experiments.

C.1.1 Participant Statistics

For each pair of languages, in Figure 6 we provide the years of experience of the involved professionals, both as translators (i.e. translating from scratch) as well as MT post-editors. In line with overall statistics in field,[38] women make up the majority of involved translators (77%). We did not enforce balanced gender distributions in the recruitment process and did not deem the gender of the translators as a significant variable. Indeed, feminine and masculine lexical terms are equally standard, grammatical forms used to refer to human referents, which are part of the current language. This is also confirmed by prior work (Popovic and Lapshinova-Koltunski, 2024), which did not find translator's gender to be an indicative factor in gender translation. Participants were only instructed to use them in translation according to the provided gender information for each sentence.

No personal information was collected for students.

C.2 PE effort across voluntary and paid professionals

Given that the PE activity for en-it is carried out by both paid and non-paid professionals (see Appendix C.1), we want to ensure that the two conditions are comparable. For this reason, we collected a *control subset of sentences* – edited by both paid professionals and voluntary professionals – to compare the PE results across these two potentially different types of subjects. To do so, we have paid translators redo 125 sentences for MTGEN-A, which is the dataset upon which most of our experiments are based. Hence, we collect an additional set of 300 post-edited sentences (i.e. the same 125 source sentences correspond to 125 F post-edits and 125 M post-edits).

Results are reported in Table 5. As we can see, the type of professional does not appear as a significant confounding variable. In absolute numbers, the two sets are highly comparable, with only a 6-minute difference in TE, and less than 1 HTER score (Δabs).

[38] https://www.linkedin.com/pulse/lets-talk-gender-equality-translation-industry-josephine-matser/

	TE				HTER			
	FEM	MASC	Δabs	Δrel	FEM	MASC	Δabs	Δrel
VOLUNTARY PROFESSIONALS	1:13	0:27	0:46	170.40	14.31	2.39	12.78	259.23
PAID PROFESSIONALS	1:07	0:26	0:40	150.95	17.71	4.93	11.92	498.74

Table 5: Comparative post-editing results for 125 sentences en-it on MTGEN-A, carried out by the group of voluntary professional translators and the second setting of paid professional translators. We provide time to edit (TE, i.e. hour:minutes) and HTER.

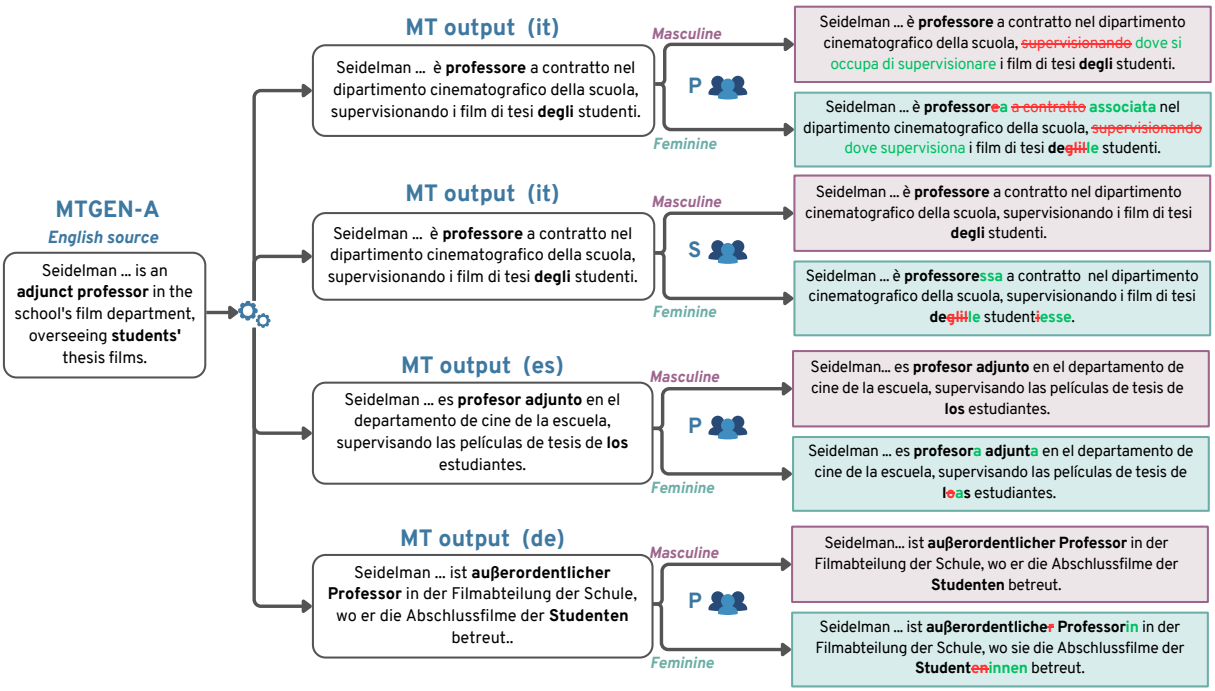


Figure 7: Post-editing example for a MTGEN-A source English sentence, which is common across all language pairs. Given the source English sentence, we show the GT automatic translation, and its associated feminine and masculine post-edits. For en-it, we show post-editing by both professionals (P) and students (S). In bold, we show **gender-related words** in the source, output, and post-edited sentences. For the post-edits, we show deletions and insertions.

Given the results of this analysis, we could safely merge the data coming from both types of translators to compose the final en-it datasets. For MTGEN-A, the 125 common sentences that we decided to keep for the main experiments are those post-edited by the professional translators, so as to allow for higher comparability with the fully "paid" en-es/de data samples.

D Post-editing

In Figure 7, we show an example of the PE activity carried our for the MTGEN-A dataset. We provide an English sentence which is common to all language pairs, associated with its corresponding GT output, and both masculine and feminine post-edits showing the PE activity. As we can see from the figure, all GT outputs consistently translate human referents with masculine gender forms, which are then adjusted for the feminine PE.

Student and Professional PE Still in Figure 7, we show a typical behavioural difference that we attest between types of users for en-it. Namely, between professional translators (P) and less experienced students (S). As discussed in §4.3, we find that students post-edited less (i.e. lower number of edits and in less time) compared to professionals. As a matter of fact, students did not engage with the improvement of the overall quality of the sentence, most likely due to their lower English proficiency, and rather mainly looked at the Italian target to fix gendered translation. In fact, in the provided example (the en-it blocks at the top), the GT output provided a poor translation for "*overseeing*" – rendered as "supervisionando", which

TER % (x)	% of payment from full per source word rate	MT Discount
0	30%	70%
0<x<=10	35%	65%
10<x<=20	40%	60%
20<x<=25	50%	50%
25<x<=30	70%	30%
30<x<=40	90%	10%
x>40	100%	0%

Figure 8: HTER Pricing matrix

is suboptimal in terms of fluency, overall also impacting the adequacy and readability of the sentence. Indeed, for both feminine and masculine PE, professionals carried out a light post-editing that also ensured an alternative translation for that portion of the sentence, whereas it was overlooked by students. Overall, since the adjustments made by students were basically only gender-related, the attested gender disparities measured with HTER and TE become even more visible.

E HTER Payment Rates

To calculate HTER-based payments, we rely on the discount rates reported in Figure 8. The matrix is publicly available and based on Localization (2022). Note that discount rates can vary across companies. We compare the matrix with the HTER discounts used by other major language service providers. Such rates however cannot be divulged as they are internal to the company and reserved. Overall, we find that the used scheme is highly aligned with those from other private companies and – if anything – it is more conservative, with a limited number of HTER ranges.

F Additional Results

F.1 Overall Translation Quality

In Table 6 we report overall translation quality results obtained by Google Translate for all datasets and languages. We used the original target reference translation to compute the results.

Details on automatic metrics computation are available in Appendix B.3.

F.2 Automatic gender bias results

We report contrastive, reference-based gender bias results computed with different metrics in Table 9. For details on the metrics computation, please refer to Appendix B.3.

		BLEU (↑)	TER (↓)	COMET (↑)
en-it	MUST-SHE	40.64	47.54	84.56
en-it	MTGEN-UN	43.92	42.92	82.31
en-it	MTGEN-A	35.77	50.44	84.75
en-es	MTGEN-A	49.72	34.2	85.29
en-de	MTGEN-A	36.04	49.35	84.28

Table 6: Overall quality translation results per each dataset and language.

		BLEU (↑)		TER (↓)		COMET (↑)	
		FEM	MAS	FEM	MAS	FEM	MAS
en-it	MUST-SHE	37.15	43.51	50.38	45.18	83.59	85.43
en-it	MTGEN-UN	42.9	44.94	43.94	41.91	84.25	84.86
en-it	MTGEN-A	30.63	39.8	55.05	46.79	83.02	86.52
en-es	MTGEN-A	43.53	54.56	38.95	30.72	83.75	86.64
en-de	MTGEN-A	30.29	40.52	53.99	45.87	82.82	85.77
		Δabs	Δrel	Δabs	Δrel	Δabs	Δrel
en-it	MUST-SHE	-6.36	-14.62	5.20	11.51	-1.84	-2.15
en-it	MTGEN-UN	-2.04	-4.54	2.03	4.84	-0.60	-0.71
en-it	MTGEN-A	-9.17	-23.04	8.26	17.65	-3.50	-4.05
en-es	MTGEN-A	-11.03	-20.22	8.23	26.79	-2.89	-3.34
en-de	MTGEN-A	-10.23	-25.25	8.12	17.70	-2.95	-3.44

Table 9: Contrastive reference-based evaluation results for each language and dataset (Top), as computed with different metrics. Below, we show absolute difference (Δabs) and percentage difference (Δrel) values between feminine and masculine scores.

As expected, and in line with our post-editing results discussed in §4, the unambiguous dataset MTGEN-UN obtains the smallest difference in scores. Overall, by looking at the differences in score computed against the feminine and masculine references (Δ) also automatic evaluation methods confirm that GT exhibits gender bias, leading to a higher generation of masculine forms. However, we immediately see that the *magnitude* of such differences is notably small compared to our human-centered results reported in the main experiments of the paper (see §4). This is particularly true for COMET, which is less sensitive to surface differences, such morphological gender-related differences. Overall, however, none of these metrics appear particularly sensitive at capturing gender differences, which are at best framed as +26.79 percentage difference as measured with TER (see MTGEN-A for en-es). To further investigate this point, in the upcoming Appendix F.3 we verify the correlation between automatic scores and our human-centered measures.

F.3 Correlation with automatic metrics

F.3.1 Aggregated results with COMET and TER scores

As already discussed in Section 5, performance differences in automatic metrics show a weak correlation with differences in human-centric metrics. This trend is reconfirmed by both COMET and TER scores, as shown in Figure 9. Here, we still present aggregate results computed for all datasets, languages, and types of users.

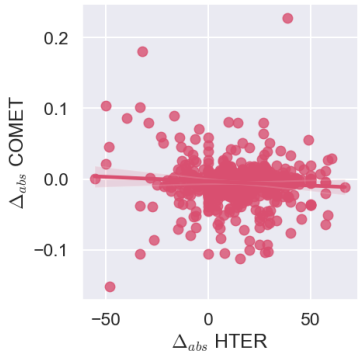
For the differences in COMET, we observe a relatively sparse distribution in Figure 9.a, with a Pearson-r coefficient of -0.12, meaning a *very weak* negative correlation, against HTER. Similarly, the Pearson-r coefficient against temporal effort (seconds per word) is -0.17, which is slightly higher but still represents a *very weak* correlation. Even in the case of COMET, the correlation is negative because lower scores are better, while the opposite is true for HTER and sec_per_word. Moreover, when compared to Figure 5, we observe a very similar behavior of BLEU (§5, Figure 5) with the one shown by COMET in Figure 9.a and 9.b, resembling similar distributions. Looking at TER differences, the samples of the distributions are slightly more squeezed towards the regression line. This means that the correlation is slightly higher but, however, still reaming *very weak*, both considering HTER ($r = 0.14$), and secs_per_word ($r = 0.18$). In this case, the correlations are positive since the higher TER scores the better, similar to human-centric metrics.

F.3.2 BLEU Results per dataset

We report language, users, and dataset-wise results of the correlations between the automatic metric BLEU and the human-centric metrics HTER and secs_per_word. Similar trends are also shown for COMET and TER, as discussed in Appendix F.3.1.

Pearson correlation coefficients for each combination are shown in Table 10. Language-wise correlations on MTGEN-A are shown in Figure 7 while dataset-wise correlations on MTGENEVAL_UN and MUST-SHE for *en-it* are shown in Figure 8.

In Section 5 , we elaborated on the weak correlations between automatic metrics such as BLEU scores and temporal and technical effort metrics such as HTER and seconds per word (SPW). When looking at the correlation results for each dataset, we observe similar trends: only HTER and SPW are moderately correlated while automatic and tem-



(a) Δ_{abs} HTER and COMET

(b) Δ_{abs} secs_per_word and COMET

(c) Δ_{abs} HTER and TER

(d) Δ_{abs} secs_per_word and TER

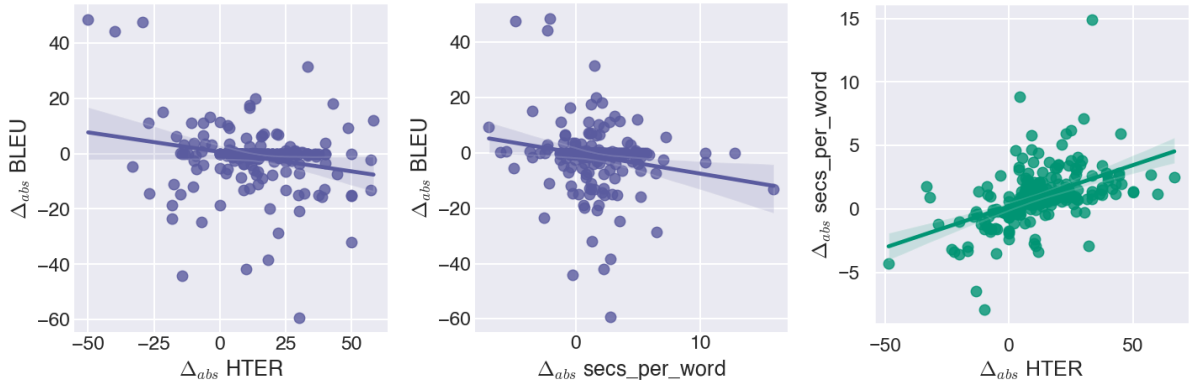
Figure 9: Scatter plots with overlaid regression lines on all datasets and languages for differences between feminine and masculine scores.

	Pearson-r	BLEU-HTER	BLEU-SPW	HTER-SPW
en-it	MTGEN-UN	0.18	0.03$^{\times}$	0.50
en-it	MUST-SHE	-0.14	-0.22	0.48
en-it	MTGEN-A (P)	-0.22	-0.18	0.54
en-it	MTGEN-A (S)	-0.24	-0.31	0.51
en-es	MTGEN-A	-0.44	-0.27	0.49
en-de	MTGEN-A	0.19	0.03$^{\times}$	0.50

Table 10: Pearson R Coefficients of correlations between Δ_{abs} BLEU, Δ_{abs} HTER and Δ_{abs} SPW (secs_per_word), for the different datasets and languages analyzed in the paper. Non-statistically significant results are indicated with $^{\times}$.

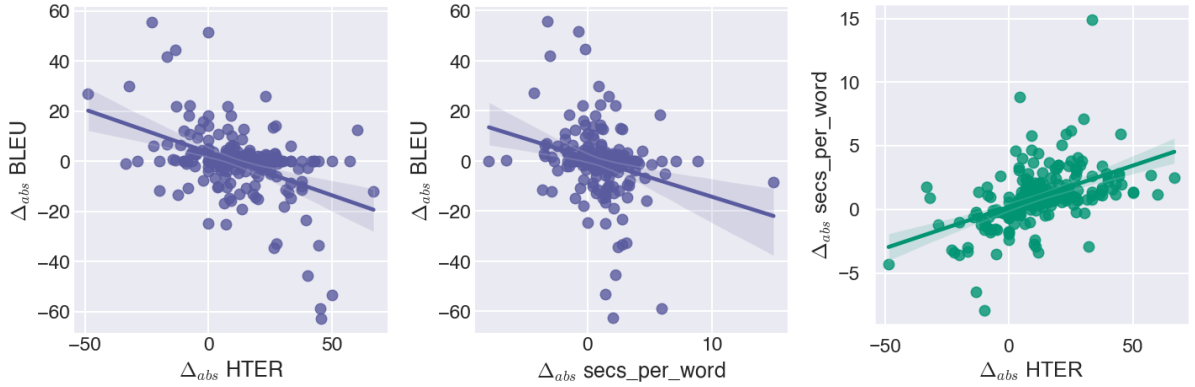
poral/technical effort metrics exhibit no or weak correlation, with also some non-statically significant results. Therefore, the conclusions drawn when looking at aggregated statistics are similar to those obtained individually for each dataset.

MTGEN-A *en-it* (P)



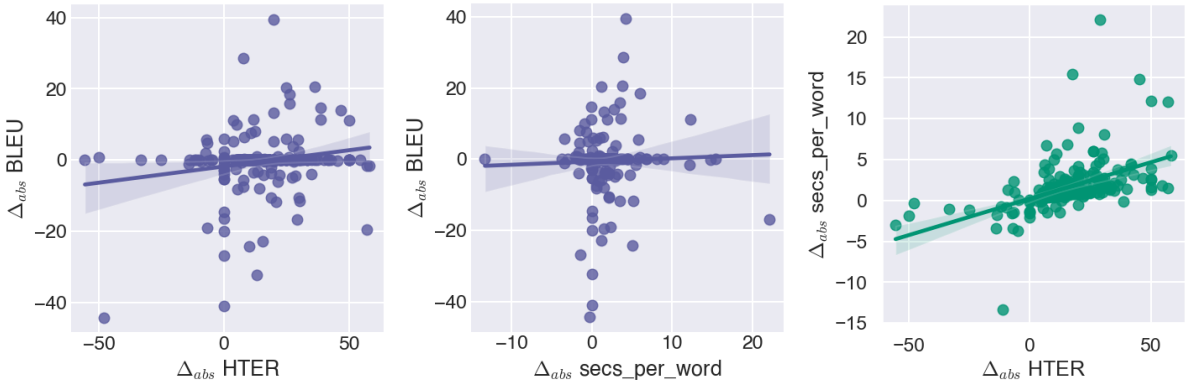
(a) Δ_{abs} HTER and BLEU (b) Δ_{abs} secs_per_word and BLEU (c) Δ_{abs} HTER and secs_per_word

MTGEN-A *en-es*



(a) Δ_{abs} HTER and BLEU (b) Δ_{abs} secs_per_word and BLEU (c) Δ_{abs} HTER and secs_per_word

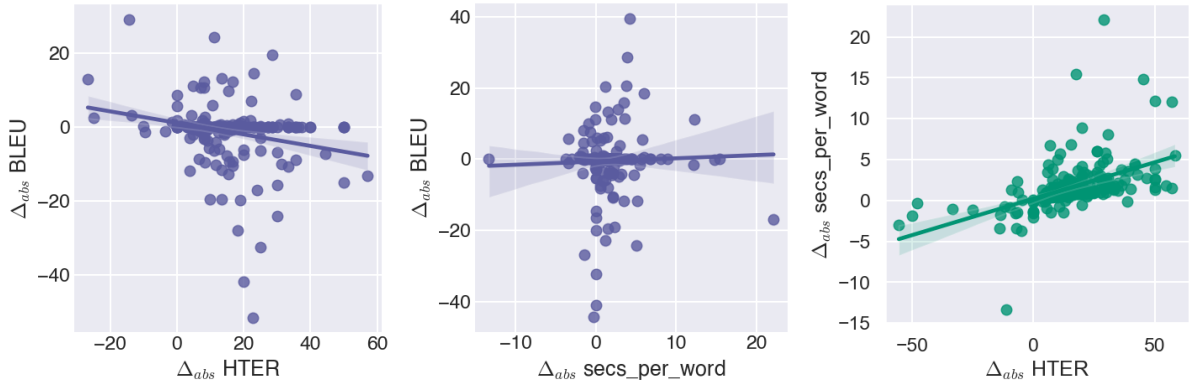
MTGEN-A *en-de*



(a) Δ_{abs} HTER and BLEU (b) Δ_{abs} secs_per_word and BLEU (c) Δ_{abs} HTER and secs_per_word

Figure (7): Scatter plots with overlaid regression lines for all languages on MTGEN-A.

MTGEN-A *en-it* (S)

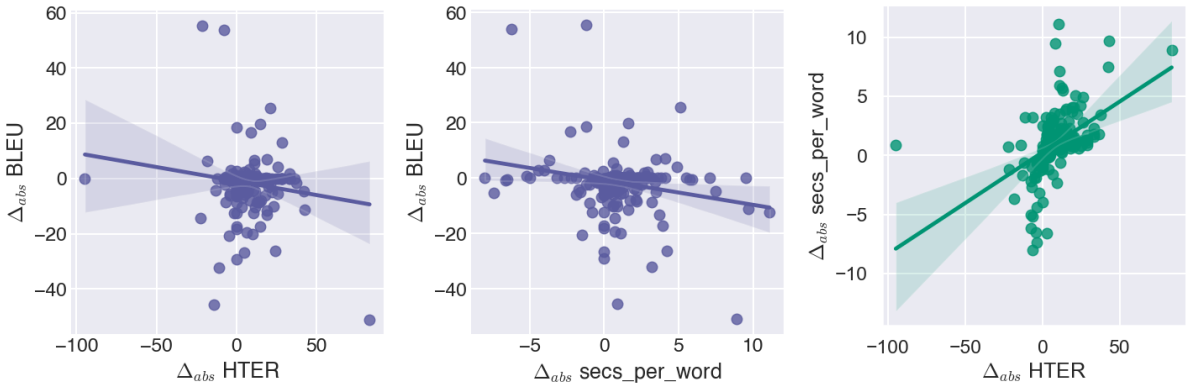


(a) Δ_{abs} HTER and BLEU (b) Δ_{abs} secs_per_word and BLEU (c) Δ_{abs} HTER and secs_per_word

MTGEN-UN

(a) Δ_{abs} HTER and BLEU (b) Δ_{abs} secs_per_word and BLEU (c) Δ_{abs} HTER and secs_per_word

MUST-SHE



(a) Δ_{abs} HTER and BLEU (b) Δ_{abs} secs_per_word and BLEU (c) Δ_{abs} HTER and secs_per_word

Figure (8): Scatter plots with overlaid regression lines on MTGEN-UN (S), MTGEN-UN and MUST-SHE for *en-it*.