Test Suites Task: Evaluation of Gender Fairness in MT with MuST-SHE and INES

Beatrice Savoldi, Marco Gaido, Matteo Negri, Luisa Bentivogli
Fondazione Bruno Kessler
{bsavoldi,mgaido,negri,bentivo}@fbk.eu

Abstract

As part of the WMT-2023 "Test suites" shared task, in this paper we summarize the results of two test suites evaluations: MuST-SHE^{WMT23} and INES. By focusing on the en-de and de-en language pairs, we rely on these newly created test suites to investigate systems' ability to translate feminine and masculine gender and produce gender-inclusive translations. Furthermore we discuss metrics associated with our test suites and validate them by means of human evaluations. Our results indicate that systems achieve reasonable and comparable performance in correctly translating both feminine and masculine gender forms for naturalistic gender phenomena. Instead, the generation of inclusive language forms in translation emerges as a challenging task for all the evaluated MT models, indicating room for future improvements and research on the topic.

We make MuST-SHE^{WMT23} and INES freely available, respectively at:

https://mt.fbk.eu/must-she/

https://mt.fbk.eu/ines/

1 Introduction

As Machine Translation (MT) has made strides in generic performance, there is an increasing recognition of the need to scrutinize finer, more nuanced aspects that defy assessment through traditional metrics computed on generic test sets. It is within this context that the WMT Test Suites shared task emerges, aiming to provide a dedicated evaluation framework to delve into specific dimensions of MT output with a laser focus. In particular, those representing well-known challenges within the current MT landscape.

In light of the above, our contribution is dedicated to the critical themes of gender bias and inclusivity in translation (Savoldi et al., 2021). Given the large-scale deployment of MT, such aspects are not only relevant from a technical perspective, where gender-related errors negatively impact the accuracy of automatic translation. Rather, biased and non-inclusive systems can pose the concrete risk of under/misrepresenting gender minorities by over-producing masculine forms, while reinforcing binary gendered expectations and stereotypes (Blodgett et al., 2020; Lardelli and Gromann, 2022).

Accordingly, in this paper we present the FBK participation in the Test Suites shared task by conducting evaluations on two newly-created test suites:

1. **MuST-SHE**^{WMT23} for en-de, created as a English→German extension of the already existing multilingual MuST-SHE corpus (Bentivogli et al., 2020). This dataset is designed to allow for fine-grained analysis of (binary) gender bias in MT.

2. **INES** for de-en, designed to assess the ability of MT systems to generate inclusive language forms over non-inclusive ones when translating from German into English.

The MuST-SHE^{WMT23} and INES datasets, as well as their corresponding metrics and evaluations, are respectively discussed in Section 2 and 3. In the evaluations presented in this paper, we obtained translations of our test suites by systems that are part of the standard General Translation Task of the Eighth Conference on Machine Translation (WMT-2023). In particular, we evaluated 11 systems for MuST-SHE^{WMT23} en-de and 13 systems for INES de-en.

2 MuST-SHE^{WMT23}: en-de Evaluation

MuST-SHE^{WMT23} is a test suite designed to evaluate the ability of MT systems to correctly translate gender. It is composed of 200 segments that require the translation of at least one English gender-neutral word into the corresponding

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 252–262
December 6–7, 2023.

Form		Category 1: *Ambiguous first-person references*	Speaker
Fem.	SRC	The other hat that I've worn in my work is as **an activist**...	She
	REF_{De}	Der andere Hut, den ich bei meiner Arbeit getragen habe, ist **der**<den> **Aktivistin**<Aktivist>...	
Masc.	SRC	I mean, I'm a **journalist**.	He
	REF_{De}	Ich meine, ich bin **Journalist** <Journalistin>.	
		Category 2: *Unambiguous references disambiguated by gender info*	
Fem.	SRC	A college classmate wrote me a couple weeks ago and she said ...	He
	REF_{De}	**Eine**<Ein> **Kommilitonin**<Kommiliton> hat mir vor ein paar Wochen geschrieben und gesagt...	
Masc.	SRC	I decided to pay a visit to **the manager** [...] and he pointed ...	She
	REF_{De}	Also entschied ich mich **den**<die>**Filialleiter**<Filialleiterin> zu besuchen [...]	

Table 1: MuST-SHE annotated segments organized per category. For each gender-neutral word referring to a human entity in the English source sentence (SRC), the reference translation (REF) shows the corresponding gender-marked (Fem./Masc.) forms, annotated with their wrong <gender-swapped> forms. The last column of the table provides information about the speaker's gender.

masculine or feminine target word(s) in German. The test suite is created as an extension of MuST-SHE, a multilingual, natural benchmark built on TED talks data (Bentivogli et al., 2020), which allows for a fine-grained analysis of gender bias in MT and ST. The original MuST-SHE corpus comprises ∼3,000 (*audio, transcript, translation*) triplets annotated with qualitatively differentiated gender-related phenomena for thee language pairs: English→ French/Italian/Spanish. Here, we introduce a newly created **English→ German** textual portion (*transcript, translation*) of the MuST-SHE corpus.

2.1 MuST-SHE^{WMT23} Dataset

Phenomena of Interest. Following the MuST-SHE original design, MuST-SHE^{WMT23} is intended to evaluate the translation of a source English neutral word into its corresponding target gender-marked one(s) in the context of human referents, e.g. en: ***the*** *good* ***friend***, de: ***der/die*** *gute* ***Freund/in***.

To allow revealing a potential gap across the generation of feminine/masculine gender forms, the corpus includes a balanced number of feminine (F) and masculine (M) translation phenomena. Also, the corpus features two categories of phenomena, which differ in the presence/lack of a gender cue to disambiguate the translation. Namely, *i)* **CAT1**: consisting of first-person singular references (i.e. to the speaker), which are to be translated according to the speaker's linguistic expression of gender, e.g., *I am a good friend*. Then, *ii)* **CAT2** consisting of references to any participant, which are be translated according to explicit gender information available in the sentence, like lexically gendered words (*sister*, *Mr*), or pronouns (*He/she is a good friend*). These categories allow differentiating systems' behaviour across ambiguos vs. unambiguos cases.

Dataset creation. In order to create MuST-SHE^{WMT23} we collected a pool of English-German candidate segments by exploiting the same TED-based data sources used to create the other MuST-SHE datasets, namely: the Dev and Common Test sets of the MuST-C corpus, and other parallel sentences extracted from additional TED talks. Then, to target those segments that represented our phenomena of interest, we followed the same automatic procedure used for the original MuST-SHE benchmark, which was aimed to quantitatively and qualitatively maximize the extraction of an assorted variety of gender-marked phenomena. Regular expressions were employed to transform German gender-agreement rules into search patterns to be applied to our pool of candidate sentences. Also, to specifically match a differentiated range of gender-marked lexical items, we also compiled two series of 50 human-referring adjectives in English and German.

Once the automatic step was concluded, the pool of retrieved sentence pairs underwent a manual inspection to: *i)* remove any noise and keep only pairs containing at least one gender phenomenon; *ii)* ensure that the final (*transcript, translation*) pairs were not affected by misalignments resulting from the automatic procedure used to create

MuST-C and the new TED Talks data. Also, we examined the remaining pairs to verify that those to be included in MuST-SHE featured a a balanced distribution of categories, F/M forms, and speakers. Accordingly, since the MuST-C corpus presents a well-known gender imbalance[1], we excluded all of the extracted masculine segments that exceeded the feminine counterpart. Across categories, instead, we were not able to ensure a balanced distribution, as fewer instances from CAT1 could be identified.[2]

The resulting dataset – whose statistics are given in Table 2 – was then manually enriched with different types of information. For each segment, the annotation includes: category (1 and 2), gender form (F and M), and speaker's gender information.[3] Also, for each target gender-marked word in MuST-SHE^{WMT23}, we created a corresponding gender-swapped counterpart in the opposite gender form. As shown in Table 1, these word forms were paired and annotated in the reference translations. As we will describe in more detail in the upcoming Section 2.2, such annotated target gender-marked words are key features of MuST-SHE, which enable gender-sensitive, fine-grained analyses focusing solely on the correct generation of target gender-marked words.

The manual selection of appropriate sentences and their annotation was carried out by two annotators, both students proficient in the German language and with a background in Applied Linguistics.[4] Each annotator worked on half of the corpus independently and then revised the work done by the other. Finally, all the differences found were reconciled to get to the final corpus.

	CAT1	CAT2
Fem.	23	77
Masc.	23	77
Tot.	200	

Table 2: MuST-SHE^{WMT23} sentence-level statistics.

[1]As reported in MuST-Speakers, ∼70% of the speakers in MuST-C are referred to by *He* pronouns.

[2]This is most likely due to the gendered features of the German language, which – unlike es, fr, and it – does not carry gender markings on verbs (e.g., *I went* → de: *Ich bin gegangen* vs it: *Sono andat**a/o***) nor adjective in the nominative case (e.g., *I am good* → de: *Ich bin gut* vs. es: *Soy buen**a/o***.

[3]Such an information is migrated from the MuST-Speakers resource (Gaido et al., 2020), where gender information for each speaker in MuST-C has been labeled based on the personal pronouns the speakers used to describe themselves in their publicly available personal TED section.

[4]Their work was carried out as part of an internship at FBK.

2.2 MuST-SHE^{WMT23} Evaluation

Following the original MuST-SHE evaluation protocol described in Gaido et al. (2020), MuST-SHE^{WMT23} evaluation allows to focus on the gender realization of the target gender-marked forms, which are annotated in the reference translations together with their *wrong*, gender-swapped form (see Table 1). The evaluation is carried out in two steps, and by matching the annotated (*correct*/*wrong*) gender-marked words against the MT output. Accordingly, we first calculate the **Term Coverage** as the proportion of gender-marked words annotated in MuST-SHE (either in the correct or wrong form) that are actually generated by the system, on which the accuracy of gender realization is therefore *measurable*. Then, we define **Gender Accuracy** as the proportion of correct gender realizations among the words on which it is *measurable*. This evaluation method[5] has several advantages. On one side, *term coverage* unveils the precise amount of words on which systems' gender realization is measurable. On the other, *gender accuracy* directly informs about systems' performance on gender translation and related gender bias: scores below 50% indicate that the system produces the wrong gender more often than the correct one, thus signalling a particularly strong biased behaviour.

2.3 MuST-SHE^{WMT23} Results

In Table 3 we present the MuST-SHE^{WMT23} results for the 11 en-de systems that were submitted to the WMT-2023 standard General Translation Task. Starting from coverage results, the scores range between 67.34% (`AIRC`) and 77.07% (`ONLINE-G`), with only 3 systems under 70%. Hence, overall all models produce a good amount of gender-marked words that can be evaluated with regards to the accuracy of their gender realization. Moving onto the overall accuracy scores (All-Acc), we can see that – while there is still room for improvement – all of the evaluated MT systems are reasonably good at translating gender, with `ONLINE-M` emerging as the best model, able to correctly translate gender in 80% of the generated instances. If we go more fine-grained into results disaggregated across gender forms (F and M) and categories (1 and 2), however, we can unveil subtle differences. Indeed, for unambiguous

[5]The evaluation script is publicly available at: https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/gender/mustshe_gender_accuracy.py.

	All-Cov	All-Acc	1F-Acc	1M-Acc	2F-Acc	2M-Acc
ONLINE-M	75.07	80.07	50.00	84.00	86.08	80.00
ONLINE-Y	73.35	79.65	30.43	96.15	86.96	78.51
NLLB_MBR_BLEU	71.92	79.43	36.00	92.31	87.27	78.51
ONLINE-W	67.91	79.32	23.81	90.91	86.11	80.87
ONLINE-G	77.07	78.87	16.00	95.15	87.39	79.69
ONLINE-B	72.20	78.64	14.28	100.00	83.92	81.25
ONLINE-A	74.78	78.00	25.00	92.30	84.34	79.36
GPT4-5shot	69.63	77.94	10.53	95.83	83.33	80.17
ZenhHuiMT	73.07	77.35	19.23	95.65	84.82	79.37
Lan-BridgeMT	71.92	75.79	16.67	92.31	83.19	77.05
AIRC	67.34	73.98	10.53	87.50	81.25	74.56

Table 3: MuST-SHE^{WMT23} results for en-de. Systema are ranked based on overall Gender Accuracy (All-Acc).

gender translation from CAT2, systems perform basically on par across gender forms, with actually slightly higher results for feminine translation. Instead, results on CAT1 unveil a huge gender gap, with systems achieving almost perfect results for masculine translation, whereas feminine accuracy can be as low as 10.53%. In fact, the best ranked systems `ONLINE-M` generates the correct feminine form in 50% of the cases, namely at a random rate.

Overall, results on MuST-SHE^{WMT23} show that the evaluated MT systems are reasonably good at translating gender under realistic conditions, achieving comparable results across feminine and masculine gender translation. However, for ambiguous cases where the input sentence does not inform about the gender form to be used in translation, we confirm a strong skew where all systems favour masculine generation almost by default. This finding calls for further research endeavours and evaluation initiatives to counter gender bias in MT and measure future advances.

3 INES: de-en Evaluation

The INclusive Evaluation Suite (INES) is a test set designed to assess MT systems ability to produce gender-inclusive translations for the German→English language pair. By design, each German source sentence in INES includes an expression that can be rendered by means of either an *inclusive* (`IN`) or *non-inclusive* (`N-IN`) expression in the English target language.

Overall, INES comprises 162 manually curated German sentences, which are annotated with their corresponding (`IN`/`N-IN`) English expressions. As such, it allows to evaluate to what extent MT systems favor the generation of non-inclusive solutions over alternative, valid inclusive translation in their output.

3.1 INES Dataset

Here, we first describe the phenomena of interest included in INES. Then, we proceed by describing its creation methodology.

Phenomena of interest. Despite its comparatively restricted gender grammar, English has traditionally relied on the use of marked forms that treat the masculine gender as the conceptually generic, default human prototype, i.e. as *masculine generics* (Silveira, 1980; Bailey et al., 2022). Exemplary cases of such a phenomenon are man-derivates (e.g., *man-made*, *freshman*) and the use of masculine personal pronouns for generic referents (e.g., "each student must submit *his* form"). Besides, expressions such as "*man* and *wife*" have been identified as depicting skewed representation of genders and gender roles (Stahlberg et al., 2007). Toward the adoption of fairer language for all genders, alternative and inclusive solutions are increasingly promoted by institutions (Höglund and Flinkfeldt, 2023) and recommended in writing (APA, 2020). These include the use of unmarked forms (e.g. *human-made*, *first-year student*) and neutral pronouns (e.g. "each student must submit *their* form") for generic and under-specified referents, as well as more symmetrical formulations that cast men and women in the same role (e.g. "*husband* and *wife*").

On this basis, INES represents translation phenomena where, given a source German sentence, systems are confronted with the generation of a corresponding inclusive or non-inclusive solution. As shown by the examples in Table 4, the German sentences can entail the use of either *i)* a generic masculine form, e.g. *Der Polizist*, or *ii)* a term that does not convey gender, e.g. *Die Menschheit*. Regardless of the source German term, the expected ideal behaviour of the MT system always entails

	German src	English pair
a.	**Der Polizist** half der alten Dame, die Straße sicher zu überqueren.	**police officer**, **policeman**
b.	**Die Menschheit** hat das Potenzial, die Welt zu einem besseren Ort zu machen.	**humankind**, **mankind**
c.	Die **fachmännische** Arbeit des Teams führte zum erfolgreichen Abschluss des Projekts.	**skillful**, **workmanlike**
d.	Die **geschickte** Arbeit des Teams führte zum erfolgreichen Abschluss des Projekts.	**skillful**, **workmanlike**

Table 4: INES source German example sentences with their corresponding annotated English IN and N-IN terms.

the generation of inclusive target words.

Dataset Creation. Since the focus of the INES test suite is to evaluate the ability of MT systems to generate inclusive English translations, we started by compiling a list of well-established pairs of English IN/N-IN terms and expressions. This list was created based on existing collections of paired terms (Vanmassenhove et al., 2021; Amrhein et al., 2023) and integrated with few additional terms retrieved from other inclusive language guidelines from international institutions[6] and universities.[7][8] As a result, we obtained 48 IN/N-IN English pairs, which are shown in Table 5.

Starting from this list, we created the source German sentences that compose INES following a two-step semi-automatic procedure.

In the first step, for each English IN/N-IN term of the pairs, GPT[9] was prompted to generate 3 German sentences containing such term translated into German, for a total of 6 sentences for each English pair.

In the second step, the initial pool of 288 synthetic sentences was manually revised by a linguist proficient in German.[10] The revision was aimed to *i)* correct generation errors and *ii)* select a balanced amount of German sentences for each phenomenon of interest. To this purpose:

- when all the 6 German sentences generated for the two (IN/N-IN) terms of the English pair contained only gender-marked terms (e.g. *police officer –> Der Polizist / policeman –> Der Polizist*) or only gender-neutral terms (e.g. *humankind –> Die Menschheit / mankind –> Die Menschheit*), only 3 sentences out of 6 were kept (see examples a. and b. in Table 4);

IN vs N-IN for job titles	
anchor	anchorman
anchors	anchormen
bartender	barman
bartenders	barmen
business person	businessman
business persons	businessmen
chairpeople	chairmen
chairperson	chairman
firefighter	fireman
firefighters	firemen
flight attendant	steward
flight attendants	stewards
mail carrier	postman
mail carriers	postmen
member of congress	congressman
members of congress	congressmen
police officer	policeman
police officers	policemen
principal	headmaster
principals	headmasters
salesperson	salesman
salespersons	salesmen
spokesperson	spokesman
spokespeople	spokesmen
supervisor	foreman
supervisors	foremen
weather reporter	weatherman
weather reporters	weathermen
IN vs N-IN for generic man	
average person	average man
average people	average men
best people for the job	best men for the job
best person for the job	best man for the job
human-made	man-made
humankind	mankind
husband and wife	man and wife
intermediaries	middlemen
intermediary	middleman
skillful	workmanlike
laypeople	laymen
layperson	layman
workforce	manpower
first-year student	freshman
first-year students	freshmen
IN vs N-IN pronouns	
their	his
theirs	his
them	him
themself	himself
they	he

Table 5: INES pairs of English Inclusive (IN) vs non-inclusive (N-IN) expressions.

- on the contrary, when the 6 German sentences generated for the two (IN/N-IN) English terms included both gender-marked and gender-neutral forms (e.g. *firefighters –> Feuerwehrleute / firemen –> Feuerwehrmänner*), they were all kept, so as to have a richer representation of the phenomenon of interest

[6] https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf

[7] https://writingcenter.unc.edu/tips-and-tools/gender-inclusive-language/.

[8] https://www.gsws.pitt.edu/resources/faculty-resources/gender-inclusive-non-sexist-language-guidelines-and-resources.

[9] gpt-3.5-turbo.

[10] One of the authors of the paper.

in the source (see c. and d. in Table 4).

Unfortunately, we found only very few instances of double German realizations, and thus at the end of the manual revision, we remained with 162 German sentences: 21 with an inclusive source term, and 141 with a non-inclusive masculine generic in the source. All the 162 manually-curated German source sentences are included in INES, and provided with their corresponding English IN/N-IN term pair so as to allow for focused evaluations.

3.2 INES Evaluation

To evaluate systems against INES, we can leverage the annotated pairs of English IN/N-IN expressions and match them against the MT generated output. Accordingly, we can perform our evaluation by adopting the same evaluation protocol and metrics defined for MuST-SHE in 2.2. Namely, by *i)* first computing **Term Coverage** as the proportion of IN/N-IN generated by a system, and then *ii)* calculating **Inclusivity Accuracy** as the proportion of IN generated expressions, among all of the generated ones. As a result, all the *out of coverage words* (OOC) are necessarily left unevaluated.

While prior manual assessments of the terms left unevaluated by such an automatic method have been able to confirm the robustness and validity of the accuracy results in the context of binary gender translation (Savoldi et al., 2022b), here we hypothesise a potential limit for evaluating inclusivity in English outputs. Our hypothesis lies on the fact that English, a notional gender language (McConnell-Ginet, 2013), has a restricted repertoire of gender-marked – potentially N-IN – words, whereas most English nouns simply do not convey any gender distinctions (e.g., *doctor*, *secretary*, *president*). In other words, there might be many potential inclusive alternatives and synonyms (e.g. *presenter* and *host* for <*anchor*>) for a single N-IN term (e.g. <*anchorman*>). Thus, whereas OOC words in the context of binary gender present the same distribution assessed automatically in terms of accuracy, this metric might be stringent for inclusivity in English, and overly penalize the generation of alternative terms that differ from those annotated in INES.

In light of the above, we also propose the **Inclusivity Index** metric, defined as:

$$\text{Inclusivity Index} = 1 - \frac{n_{\text{N-IN}}}{N} \qquad (1)$$

where $n_{\text{N-IN}}$ is the number of non-inclusive terms annotated in INES that are generated by a system, and N is the size of INES (i.e. total number of sentences to be evaluated).

In what follows, we thus carry out both **Inclusivity Accuracy** and **Inclusivity Index** evaluations,[11] and assess which one better correlates with human judgments.

3.3 INES Results

In this section (Table 6), we present the results obtained on INES by the 13 de-en systems that were submitted to the WMT-2023 standard General Translation Task. Such results are computed and discussed for Inclusivity Accuracy (Table 6a) and Inclusivity Index (Table 6b). Then, based on a manual analysis, we compare such automatic results against the systems ranking obtained with human evaluations (Table 6c).

Automatic Evaluation Results. Table 6a presents coverage and accuracy-based results. Based on such scores, the INES dataset emerges as quite a challenging test suite for current de-en systems. In fact, with the sole exception of the GPT4-5SHOT – which emerges as the best performing system (but see also Sec. 5) – all systems obtain scores that are below 50%, thus suggesting that they generate undesirable N-IN forms in more than half of the (measurable) cases. The lowest accuracy is obtained by NLLB_MBR_BLEU, amounting to 29.41% only.

Moving onto the Inclusivity Index results in Table 6b, from a bird's eye view we can immediately unveil some differences. On the one hand, GPT4-5SHOT and NLLB_MBR_BLEU still emerge as, respectively, the best and worst performing systems. On the other hand, however, there are discrepancies in the overall ranking. For instance, AIRC results as the system that generates the second-best level of inclusive translation according to the Inclusivity Index metrics, whereas it was ranked 7th in terms of accuracy.

Manual Evaluation Results. To verify which of the two automatic metrics yields more reliable results, we proceed with a manual analysis of all MT output sentences that defied the automatic evaluation procedure. Namely, we performed a human evaluation of all OOC terms to determine whether

[11] Evaluation script available at: https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/gender/INES_eval.py.

	Cov	Acc (↑)
GPT4-5shot	64.81	**65.71**
ONLINE-W	**75.31**	48.36
ONLINE-Y	74.07	45.83
ZenhHuiMT	73.46	44.54
ONLINE-A	74.69	42.98
ONLINE-B	70.99	41.74
AIRC	53.70	37.93
Lan-BridgeMT	68.52	36.94
ONLINE-M	70.37	36.84
ONLINE-G	74.07	35.00
GTCOM_Peter	74.69	33.06
NLLB_Greedy	74.07	31.67
NLLB_MBR_BLEU	73.46	29.41

(a) Coverage and Accuracy results

	In.Idx. (↑)
GPT4-5shot	77.78
AIRC	66.67
ONLINE-W	61.11
ONLINE-Y	59.88
ZenhHuiMT	59.26
ONLINE-B	58.64
ONLINE-A	57.41
Lan-BridgeMT	56.79
ONLINE-M	55.56
ONLINE-G	51.85
GTCOM_Peter	50.00
NLLB_Greedy	49.38
NLLB_MBR_BLEU	48.15

(b) Inclusivity Index results

	Human (↑)
GPT4-5shot	76.73
ONLINE-W	60.25
AIRC	59.03
ONLINE-Y	58.13
ZenhHuiMT	56.60
ONLINE-B	56.25
ONLINE-A	55.28
ONLINE-M	52.53
Lan-BridgeMT	52.26
ONLINE-G	48.45
NLLB_MBR_BLEU	46.25
GTCOM_Peter	48.13
NLLB_Greedy	44.03

(c) Human judgment – *Official ranking*

Table 6: INES evaluation results (percentage). Per each metric, systems are ranked based on their performance.

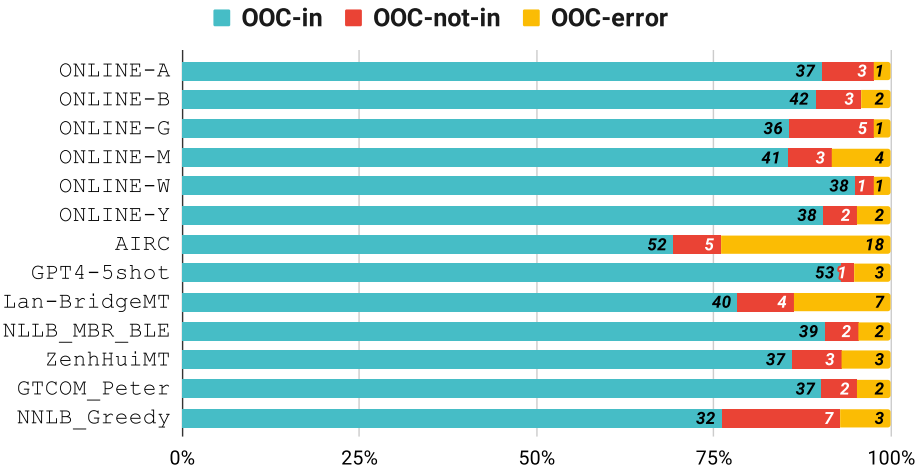

Figure 1: INES manual analysis results for out-of-coverage (OOC) terms.

Metric	Pearson (r)	Kendall (τ)	Spearman (ρ)
Acc	0.9601	0.8205	0.9285
In.Idx.	**0.9738**	**0.9231**	**0.9835**

Table 7: Correlation Coefficients with Human Judgment

the generated expression entailed *i)* an inclusive expression (OOC-in), which simply differed from the `IN` term annotated in INES but was completely acceptable; *ii)* a non-inclusive expression (OOC-not-in) different from the `N-IN` term annotated in INES; and finally *iii)* a translation error (OOC-error), which was not possible to judge in terms of inclusivity.[12] The results of such an analysis across all systems are reported in Figure 1. Such results show that, of all the OOC terms, the vast majority is represented by inclusive terms (e.g., <*business person*>/<*busissnessman*>→ *entrepeneur*). Errors, instead, are quite rare, just like non-inclusive OOC terms, which all correspond to the INES annotated `N-IN` term, but in a different number (e.g., <*freshmen*> → *freshman*).

In light of the above, our initial hypothesis – outlined in Sec. 3.2 – is thus reinforced: we do not find the same inclusivity distribution between evaluated cases in terms of accuracy (see Table 6a) and the OCC instances left unevaluated. Having now collected a complete evaluation of all the sentences, we leverage such information to obtain our official system ranking, which is shown in Table 6c. Results are computed as the proportion of inclusive (`IN` + OOC-in) terms generated by a system among all the terms that could be assessed (i.e. OOC-errors are not measurable, hence excluded).

[12]We underscore that such an analysis only concerns the terms representing the phenomena of our interest, whereas the overall judgement of the whole sentence is not accounted for.

Correlation between Automatic and Human evaluation. On this basis, and to finally verify our hypothesis, in Table 7 we report the correlation coefficients between the automatic metrics and human judgements. Accordingly, while both the Inclusivity Accuracy and Index show a satisfactory correlation with human judgements, the latter consistently emerges as a more reliable indicator of inclusivity. As such, Inclusivity Index is confirmed as the most suited measure to quantify gender-inclusive translation into English.

To conclude, our results in Tables 6 consistently indicate that current MT systems still struggle with the generation of inclusive translations. Within this landscape, GPT4-5SHOT consistently results as the model achieving the highest level of inclusivity, whereas all other models generate a ∼40% or more of non-inclusive translations in their output. This finding highlights that, while on the (binary) gender *bias* side (Section 2.3) MT systems still struggle with specific and particularly challenging ambiguous cases, the limitations of most of them on the gender *inclusion* side are evident and the problem emerges as an urgent topic for future research.

4 Related work

The last few years have witnessed and increasing attention toward (binary) gender bias in NLP (Sun et al., 2019; Stanczak and Augenstein, 2021; Savoldi et al., 2022a). Concurrently, emerging research has highlighted the importance of reshaping gender in NLP technologies in a more inclusive manner (Dev et al., 2021), also through the representation of non-binary identities in language (Lauscher et al., 2022; Ovalle et al., 2023). Foundational works in this area have included several applications, such as coreference resolution systems (Cao and Daumé III, 2020; Brandl et al., 2022) and fair rewriters (Vanmassenhove et al., 2021; Amrhein et al., 2023).

In MT, the research agenda has mainly focused on the improvement of masculine/feminine gender translation into grammatical gender languages (Savoldi et al., 2021). Along this line, different strategies have been devised to improve gender translation and mitigate masculine bias (Costa-jussà and de Jorge, 2020; Gaido et al., 2021; Choubey et al., 2021; Saunders et al., 2022). To test these methods and inspect systems' behaviour, several template-based datasets have been made available – such as WinoMT (Stanovsky et al., 2019) or SimpleGEN (Renduchintala and Williams, 2022) – which are especially intended to target occupational stereotyping. Instead, natural datasets such as the Arabic Parallel Gender Corpus (Alhafni et al., 2022) and GATE (Rarrick et al., 2023) allow for evaluation of gender bias under more naturalistic conditions. Among such type corpora, MuST-SHE (Bentivogli et al., 2020) represents the only multilingual, natural test set designed to evaluate gender bias for both MT and ST. Already available for English→French/Italian/Spanish, here we have contributed to its expansion for the English→German language pair.

As far as the topic of inclusivity and neutral language translation is concerned, research in MT is quite in its infancy. A notable exception is the work by Saunders et al. (2020), who created parallel test and fine-tuning data to develop MT systems able to generate non-binary translations for English→German/Spanish. However, their target sentences are artificial – created by replacing gendered morphemes and articles with synthetic placeholders – thus serving only as a proof-of-concept. Piergentili et al. (2023), instead, are the first to advocate for the use of target gender-neutral rephrasings and synonyms as a viable paradigm toward more inclusive MT when gender is unknown or simply irrelevant. Cho et al. (2019) and Ghosh and Caliskan (2023) investigate the preservation of gender-neutral pronouns for Korean/Bengali→English. Their results, however, show that current MT systems still face serious difficulties on relying on the inclusive, neutral pronoun *they* in translation. Along this line of work, INES – to the best of our knowledge – represents the first test suite designed to asses the use of neutral, inclusive forms beside pronouns for translating into English.

5 Conclusion

This paper summarizes the results of our WMT-2023 Test Suites evaluations, which focus on gender bias and inclusivity in translation. To this aim, we have introduced the en-de expansion of the multilingual MuST-SHE test set (Bentivogli et al., 2020) and the newly created INES dataset for de-en. The former is designed to assess gender bias and translation across a qualitatively differentiated selection of feminine/masculine gender phenomena. INES, instead, measures systems' ability to generate inclusive English translations that do not

rely on the use of masculine generics. Results on MuST-SHE^{WMT23} show that the evaluated MT systems are reasonably good at translating gender under realistic conditions, achieving comparable results across feminine and masculine gender translation. However, for ambiguous cases where the input sentence does not inform about the gender form to be used in translation, we confirm a strong skew where all systems tend to generate masculine forms almost by default. Results on INES, instead, indicate that providing inclusive translations still represents a quite challenging task for current MT systems, in spite of the increasingly widespread use and preference for inclusive language forms in English.

As a final remark, we acknowledge that the phenomena subject to our analysis (gender bias and gender inclusion) are not yet part of the repertoire of phenomena for which MT systems are currently designed. These systems are indeed primarily built with the goal of maximising translation quality in general rather than aspects of the problem, specifically fairness, for which sensitivity is still limited. All in all, however, this experience has allowed us to shed light on these issues, raise the awareness of the MT community and, hopefully, favour future developments.

Limitations

Naturally, this work comes with some limitations. First, both test suites are limited in size and number of language pairs considered. Despite their restricted size, however, both test suites provide a first glimpse into understanding and monitoring systems' behaviour with respect to gender and inclusivity. Additionally, rather than a limitation per se, both INES and MuST-SHE^{WMT23} are designed based on the specific linguistic features of the source and target language taken into account. As such, results in our evaluations intentionally do not aspire to scale and generalize to any language direction. Indeed, such linguistic specificity is also openly accounted for in the introduction of the new Inclusivity Index metric, which considers the morphology of English for a better-suited evaluation of gender inclusivity in MT. We also note that such a metric results as the best one for evaluating inclusivity under the given experimental conditions of this paper, where all the scrutinized systems (those submitted to the WMT General Translation task) are expected to feature generally good overall translation quality and to make few translation errors. As such, future work might be needed to further validate the stability of the Inclusivity Index metric under less optimal conditions and for different target languages, possibly proposing tailored metrics for each case. Finally, to generate the initial pool of sentences in INES we relied on the GPT (gpt-3.5-turbo) closed-source model. This has holds two types of implications. On the one hand, the use of proprietary models such as GPT has reproducibility consequences, since this model is regularly updated, thus potentially yielding future results that differ from those reported in this paper. On the other hand, relying on – even though only partially and post-edited – artificially generated data for testing models, might lead to contamination issues. Indeed, in Sec. 3.2 (Table 6) the GPT4-5SHOT model resulted as the most promising one, achieving the best results for inclusive translation. However, it remains to further verified whether our specific experimental settings and INES benchmark – where we use GPT-derived test data – have advantaged the performance of GPT4-5SHOT.

Ethics Statement

By addressing bias and inclusivity in MT, this work presents an inherent ethical component. It builds from concerns toward the societal impact of widespread translation technologies that reflect and propagate male-grounded and exclusionary language. Still, our work is not without risks either and thus warrants some ethical considerations. In particular, MuST-SHE^{WMT23} only focuses on traditional binary feminine/masculine gender forms. Also, INES investigates neutral, inclusive language in the context of generic, unknown referents and based on inclusive solutions encouraged by institutional guidelines. As such, we do not account for other non-binary solutions (e.g., neopronouns and neomorphemes) that are emerging from grassroots efforts. It should be stressed that the gendered and inclusive strategies incorporated in this MT work are not prescriptively intended. Rather, they are orthogonal to other attempts and non-binary expressions for inclusive language (technologies) (Lauscher et al., 2023; Ginel and Theroine, 2022).

Acknowledgements

This work is part of the project "Bias Mitigation and Gender Neutralization Techniques for Automatic Translation", which is financially supported

by an Amazon Research Award AWS AI grant. Also, we acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. Also, we would like to thank the 2022 FBK internship students Sabrina Raus and Abess Benissmail from the University of Bolzano: the creation of MuST-SHE^{WMT23} was made possible by their work.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. 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.

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.

APA. 2020. *Publication Manual of the American Psychological Association*, 7th edition. American Psychological Association.

April H Bailey, Adina Williams, and Andrei Cimpian. 2022. Based on billions of words on the internet, people= men. *Science Advances*, 8(13):eabm2463.

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.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

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, IT. Association for Computational Linguistics.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. Improving gender translation accuracy with filtered self-training. *arXiv preprint arXiv:2104.07695*.

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.

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.

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, 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.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages.

María Isabel Rivas Ginel and Sarah Theroine. 2022. Neutralising for equality: All-inclusive games machine translation. In *Proceedings of New Trends in Translation and Technology*, pages 125–133. NeTTT.

Frida Höglund and Marie Flinkfeldt. 2023. Degendering parents: Gender inclusion and standardised language in screen-level bureaucracy. *International Journal of Social Welfare*.

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

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Debora Nozza, Archie Crowley, Ehm Miltersen, and Dirk Hovy. 2023. What about em? how commercial machine translation fails to handle (neo-)pronouns.

Sally McConnell-Ginet. 2013. Gender and its Relation to Sex: The Myth of 'Natural' Gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton, Berlin, DE.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. 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.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples.

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.

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.

Jeanette Silveira. 1980. Generic Masculine Words and Thinking. *Women's Studies International Quarterly*, 3(2-3):165–178.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. *Social communication*, pages 163–187.

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.

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, IT. Association for Computational Linguistics.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. 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.