Gender Lost In Translation: How Bridging The Gap Between Languages Affects Gender Bias in Zero-Shot Multilingual Translation

Lena Cabrera[1], **Jan Niehues**[2]
[1]Department of Advanced Computing Sciences, Maastricht University, The Netherlands
[2]Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany
l.cabreraperez@student.maastrichtuniversity.nl,
jan.niehues@kit.edu

Abstract

Neural machine translation (NMT) models often suffer from gender biases that harm users and society at large. In this work, we explore how bridging the gap between languages for which parallel data is not available affects gender bias in multilingual NMT, specifically for zero-shot directions. We evaluate translation between grammatical gender languages which requires preserving the inherent gender information from the source in the target language. We study the effect of encouraging language-agnostic hidden representations on models' ability to preserve gender and compare pivot-based and zero-shot translation regarding the influence of the bridge language (participating in all language pairs during training) on gender preservation. We find that language-agnostic representations mitigate zero-shot models' masculine bias, and with increased levels of gender inflection in the bridge language, pivoting surpasses zero-shot translation regarding fairer gender preservation for speaker-related gender agreement.

1 Introduction

With the rapid proliferation of intelligent systems, machine learning models reflecting patterns of discriminatory behavior found in the training data is a growing concern of practitioners and academics. Neural machine translation (NMT) models have proven notoriously gender-biased, often resulting in harmful gender stereotyping or an under-representation of the feminine gender in their outputs. In recent years, several approaches to debias NMT have been proposed, including debiasing the data before model training, the models during training, or post-processing their outputs. However, to the best of the authors' knowledge, it has yet to be explored how the phenomenon of not observing enough data, if any, to model language accurately affects gender discrimination in multilingual NMT (MNMT).

To support translation between language pairs never seen during training (i.e., zero-shot directions), two widely-used approaches leverage the language resources (i.e., parallel data) available during training: *Pivot-based* translation uses an intermediate pivot/bridge language (as in source→pivot→target), whereas *zero-shot* translation learns to bridge the gap between unseen language pairs using cross-lingual transfer learning.[1]

In this work, we analyze gender bias in MNMT in the context of *gender preservation*, where gender information conveyed by the source language sentence needs to be preserved in the target language translation; in our experimental setting, source and target languages are grammatical gender languages that use a noun class system conforming with the *gender binary*, i.e., the classification of gender into the opposite forms of feminine and masculine, considered indicative of a person's biological sex.[2] We examine translations

[1] We use "zero-shot *directions*" to refer to language pairs unseen during training, whereas "zero-shot *translation*" is NMT capable of zero-shot inference, relying on a model's generalizability to conditions unseen during training.

[2] While gender, as opposed to biological sex, is viewed as a non-binary spectrum, many languages have not (yet) evolved beyond the male-female gender binary regarding linguistic gender when it ideally should correlate with biosocial gender.

in terms of differences in gender preservation between both genders, which, if found, are evidence of gender-biased machine translation (MT). More precisely, we focus on the impact that *bridging the gap between unseen language pairs* has on the MT models' ability to preserve the feminine and masculine gender, unambiguously indicated by the source sentence, equally well in their outputs. Our research questions are:

RQ1 How do zero-shot and pivot-based translation compare regarding gender-biased outputs for zero-shot directions?

RQ2 Does the bridge language affect the gender biases perpetuated by zero-shot and pivot-based translations?

RQ3 Do translation quality improvements of zero-shot models reduce their gender biases?

The remainder of this paper is structured as follows. Section 2 introduces the task of gender preservation in translation with relevant terminology and reviews related work on gender bias in NMT. Section 3 describes our experimental design, tailored toward investigating cause-and-effect relationships of gender bias in MNMT. Section 4 presents the data used and the evaluative procedure followed in our experiments. Section 5 presents the experimental setup and results, and Section 6 concludes with our summarized findings, limitations, and future research directions.

2 Terminology & Related Work

In a large-scale analysis of the plethora of existing research addressing gender bias in NMT, Savoldi et al. (2021) categorize them based on two conceptualizations of the problem: research works focusing on the weight of prejudice and stereotypes in NMT, and studies assessing whether gender is preserved in translation. In this paper, we analyze gender bias in MNMT in the context of gender preservation, where for translation into a gender-sensitive target language, the gender information conveyed by the source language needs to be retained in the target language translation.

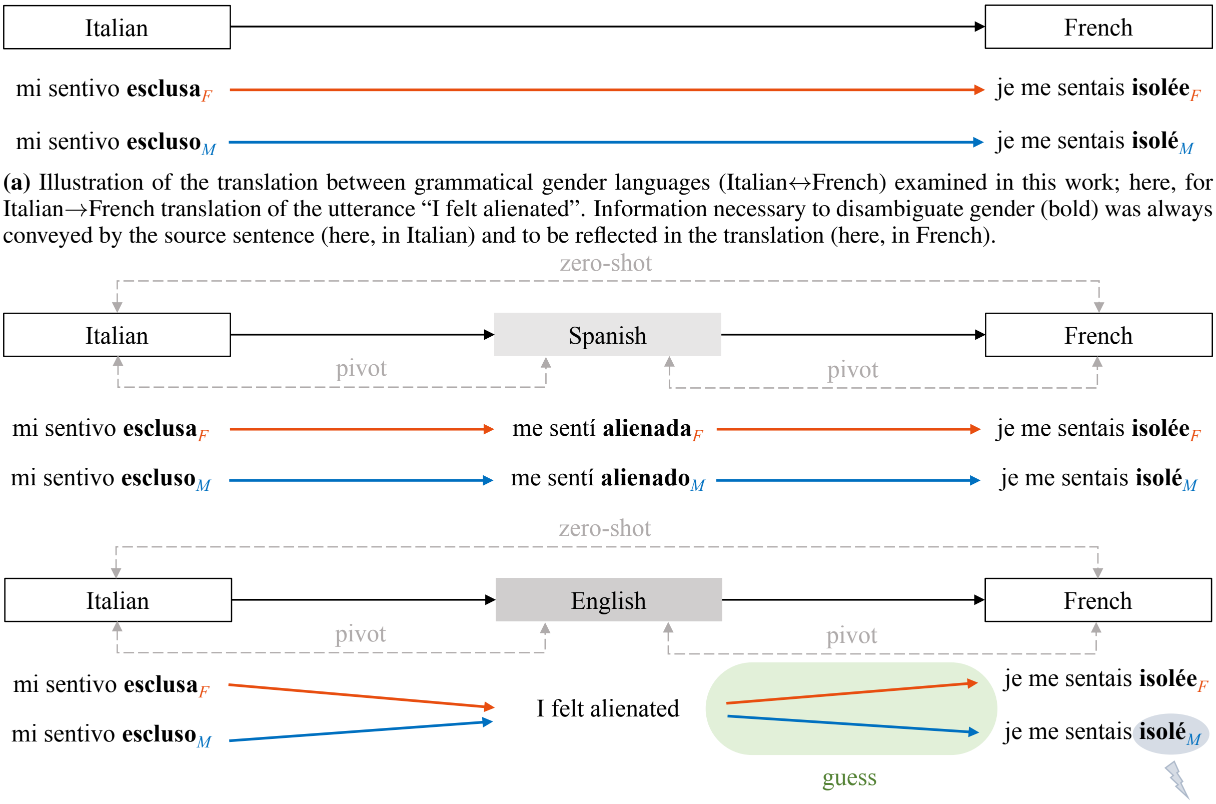
Gender in Lingustics: In our gender bias evaluation we consider *referential gender*, which, according to Cao and Daumé III (2021), only exists when an entity (i.e., a human) is mentioned and their gender (or sex) is realized linguistically. Moreover, we focus on the translation between languages using *grammatical gender*, a way of classifying nouns, assigning them gender categories (e.g., masculine, feminine, neuter, etc.) that may be independent of the real-world biosocial genders associated with referents; however, there is a tendency for languages to correlate grammatical gender with the gender of a referent, especially if human (Corbett, 1991; Ackerman, 2019).

For example, talking about a specific doctor (e.g., "the doctor loves *her*$_F$ job"), the word choice of the female anaphoric pronoun is not determined by grammatical gender but only by referential gender. The same sentence translated into German ("*die*$_F$ *Ärztin*$_F$ liebt *ihren*$_F$ Job$_M$.") requires the article ("die" = the) and pronoun ("ihren" = her) to agree with the feminine grammatical gender category the noun is assigned ("Ärztin" = female doctor).[3] On the other hand, the sentence "the doctor helps the nurse" without any further context information does not indicate the gender of either of the two mentioned entities; for the German translation, the gender of both the doctor ("Arzt$_M$"/"Ärztin$_F$") and the nurse ("Krankenpfleger$_M$"/"Krankenschwester$_F$") needs to be considered for the correct syntactic build-up of the sentence. For details on the many differences in the manifestation of gender in languages, we refer the interested reader to related works such as that of Cao and Daumé III (2021).

Gender Preservation: Translation into a gender-sensitive language, e.g., a grammatical gender language, involves gender agreement between nominal properties—e.g. grammatical and referential gender of a (pro)noun—and a determiner, adjective, verb, etc., depending on the target language agreement rules. Whenever the source language is (largely) genderless, i.e., the gender of the noun is unspecified, and context information is unavailable, gender preservation is a non-trivial task for machines and humans alike.

In recent years, several approaches have been proposed to address the challenge of gender preservation. Vanmassenhove et al. (2018) leverage additional gender information by prepending a gender tag to each source sentence, both at training and inference time, to improve the generation of speakers' referential markings. Avoiding the need

[3] Note, in German, the abstract noun "Job" is assigned the masculine grammatical gender category, while in English, "job" has no grammatical gender.



(a) Illustration of the translation between grammatical gender languages (Italian↔French) examined in this work; here, for Italian→French translation of the utterance "I felt alienated". Information necessary to disambiguate gender (bold) was always conveyed by the source sentence (here, in Italian) and to be reflected in the translation (here, in French).

(b) The richness of the gender-inflectional system of the bridge language, used to facilitate translation for unseen language pairs, affects models' ability to preserve the gender information from the source sentence. Scarcity of gender inflection in the bridge language (e.g., English) causes models to miss gender clues from the source and to resort to guessing the gender; when making the wrong guess, i.e., choosing the wrong gender as presented in the source, the model exhibits gender hallucination.

Figure 1: Overview of our investigated translation scenario (here, for the utterance meaning "I felt alienated"): At inference, we translated between unseen gender-inflected source-target language pairs (i.e., Italian↔French) by bridging, implicitly (zero-shot) and explicitly (pivot-based), using bridge languages with different gender-inflectional systems (e.g., Spanish or English).

for additional context information for training or inference, Basta et al. (2020) concatenate each sentence with its predecessor to achieve slight improvements in gender translation. Moryossef et al. (2019) inject context information as they prepend a short phrase, e.g., "*she* said to *them*", to the source sentence, translate the sentence with the prefix, and afterward remove the prefix translation from the model's output. Specifying gender inflection in this way improves models' ability to generate feminine target forms, but it relies on (not always available) metadata about speakers and listeners. Furthermore, different gender-specific translations in terms of word choices can be an arguably non-desirable side-effect.

A different approach is to post-process the output using counterfactual data augmentation. Saunders and Byrne (2020) use a lattice rescoring module that maps gender-marked words in the output to all possible inflectional variants and rescores all paths in the lattice corresponding to the different sentences with a model that has been gender-biased at the cost of lower translation quality. Choosing the sentence with the highest score as the final translation results in increased accuracy of gender selection. A downside is that data augmentation is very demanding for complex sentences with a variety of gender phenomena, such as those typically occurring in natural language scenarios.

3 Analyzing Gender Bias in MNMT

In our experimental setting, information necessary to disambiguate gender was *always* conveyed by the source sentence (cf. Figure 1a) and, thus, available to the models. Motivated by our research inquiry, we focused our investigation on the effect of bridging on gender preservation in MNMT between unseen language pairs, as illustrated broadly in Figure 1b, exploring three influencing factors to learn about the cause-and-effect relationship of gender bias in MNMT: *i)* the approach taken to bridge unseen language pairs (i.e., using continuous representations for zero-shot translation or discrete pivot language representations); *ii)* the choice of bridge language; and *iii)* language-agnostic model hidden representations.

Zero-Shot Translation Vs. Pivoting: To bridge the gap between an unknown source-target language pair at inference, we took two different approaches using the same trained translation model. For *pivot-based translation*, we cascaded a model to perform source→pivot and pivot→target translation. As such, pivoting used the pivot language as an explicit bridge between the unknown language pair. For *zero-shot translation*, we used the same model to translate directly between the unknown language pair, relying on the model's learned semantic space where sentences with the same meaning are mapped to similar regions regardless of the language. Compared to pivoting, zero-shot translation circumvents error propagation and reduces computation time, but achieving high-quality zero-shot translations is challenging. In light of our inquiry, we analyzed each approach's ability to preserve gender, comparing their performances for the feminine and the masculine gender.[4]

Bridge Language: English often participates in most, if not all, language pairs in a training corpus, making English, a language limited to pronominal gender (with a few exceptions), the most reasonable choice for a bridge language. When translating into a genderless language (e.g., Hungarian), the potential loss of gender information conveyed by the source sentence is unproblematic as it is evidently without detrimental consequence. However, when translating into a language with a *higher* gender-inflected system than English (e.g., French or Italian), the loss of gender information poses a significant problem since the information necessary to disambiguate gender is virtually no longer existent (cf. bottom in Figure 1b).

As preserving non-existent gender information is inherently impossible, also for humans, it is fair to assume that MT models have difficulty when encountering this phenomenon of gender ambiguity; the simplest solution is to resort to *random guessing*, with a 50% chance of choosing one gender over the other. Any other gender distribution ($\neq$ 50:50%) is not reflective of random guessing but instead indicative of *educated guessing* based on knowledge or observations *assumed* to be true that can, however, include biases.

Against this background, we studied the role of the bridge language in gender preservation, focusing on the gender bias differences between pivot-based and zero-shot translation, using bridge languages with different gender-inflectional systems, including English (low gender inflection), German and Spanish (high(er) gender inflection). German and English are both Germanic languages. Whereas in German, all noun classes require masculine, feminine, or neuter[5] inflection, English lacks a similar grammatical gender system. In German, the gender of the noun is reflected in determiners like articles, possessives, and demonstratives. On the other hand, Spanish is a Romance language with a binary grammatical gender system, differentiating masculine and feminine nouns; from a grammatical point of view, there are no gender-neutral nouns. The gender of nouns agrees with (some) determiners and, more often than in German, adjectives, making gender a pervasive feature in Spanish.

Language-Agnostic Hidden Representations: Since languages are characterized by different linguistic features, including those related to gender, it is reasonable to assume that language-*specific* representations, tailored to the language pairs included during training, *impair* gender preservation for unseen language pairs. Because of this, we explored the effect of three modifications to (the training of) a baseline Transformer (Vaswani et al., 2017) to encourage language-*agnostic* hidden representations, which have proven to cause performance gains for zero-shot translation. We

- removed a residual connection in a middle Transformer encoder to *lessen positional correspondences to the input tokens* and, thereby, reduce dependencies to language-specific word order (R) as proposed by Liu et al. (2021),
- encouraged *similar (i.e., closer) source and target language representations* through an auxiliary loss (AUX_{SIM}) similar to Pham et al. (2019) and Arivazhagan et al. (2019), and
- performed joint adversarial training *penalizing recovery of source language signals* in the

[4] In the presentation of our results, we use ZS and PV, short for zero-shot and pivot-based translation when space is limited.

[5] In German, neuter gender inflection does not apply to nouns identifying people (cf. referential gender).

representations (ADV_{LAN}) as done by Arivazhagan et al. (2019).

In our experiments, we examined the effect of these three modifications in isolation and tested some combinations; in total, we compared five different models to our baseline (B)—which we refer to as B+AUX_{SIM}, B+ADV_{LAN}, R, R+AUX_{SIM}, and R+ADV_{LAN}—to determine whether they mitigated models' gender biases.

4 Evaluation Data & Procedure

For our evaluation, we built on the work of Bentivogli et al. (2020) regarding the data and procedure used for our gender bias evaluation.

4.1 Multilingual Gender Preservation Dataset

In our experiments, we used the publicly available TED-based corpora MuST-C (Di Gangi et al., 2019) for model training (cf. Section 5.1 for details) and evaluated our models on a subset of MuST-SHE (Bentivogli et al., 2020), a gender-annotated benchmark. MuST-SHE is a subset of MuST-C and is available for English-French, English-Italian, and English-Spanish translations, where at least one English gender-neutral word in a sentence needs to be translated into the corresponding masculine/feminine target word(s).

The target languages included in MuST-SHE allowed us to investigate gender preservation for sentences where *the source language always provides enough information to disambiguate gender*; with this research inquiry, two main criteria needed to be met by the evaluation data: First, we wanted to evaluate gender translation *between* grammatical gender languages. Therefore, we formed a many-to-many subset from MuST-SHE, keeping only true-parallel data and realigning it to support evaluating translation between the three initial target languages. Second, we wanted to investigate the gender biases in translation between language pairs unseen during training (i.e., zero-shot directions). Using training corpora comprising different language pairs, we built models with different supervised translation directions. Accordingly, the models did not share the same zero-shot directions. For instance, a model trained on Spanish-X data had seen examples for language pairs that included Spanish. Therefore, we discarded the Spanish examples and only used French-Italian examples in our evaluation to ensure equal zero-shot directions across all models considered in our experiments.

We obtained 278 sentences with detailed statistics presented in Table 1. The included French↔Italian directions left us with 556 translations for evaluation.

	Feminine	(Female/Male)	Masculine	(Female/Male)	**Total**	(Female/Male)
Cat. 1	64	(64/0)	56	(0/56)	120	(64/56)
Cat. 2	72	(58/14)	86	(27/59)	158	(85/73)
Total	136	(122/14)	142	(27/115)	**278**	(149/129)

Table 1: Statistics of the MuST-SHE data used, broken down by referent gender (Feminine/Masculine), gender agreement (Cat. 1/2: speaker-related/speaker-independent), and speaker gender (Female/Male).

The composition of this dataset, comprising French-Italian parallel data, provides different evaluative dimensions that can be considered for gender bias evaluation of MT models.

Referent Gender: Grammatical gender agreement determines the modification of certain words to express gender congruent with the other words they relate to, which, in our case, were the words designating a *referent*—a person the speaker mentioned. Consequently, the gender of a referent (cf. referential gender) determined the gender of gender-marked words relating to the referent (i.e., for a female referent, feminine inflected words, and for a male referent, masculine inflections). All gender-marked words in a sentence did agree with the same (referent) gender. As MuST-SHE is TED-based data, a referent was either the speaker, or a person not identified as the speaker (nor the addressee(s)/audience in our data).

Speaker Gender: Due to the evaluation data stemming from TED talks, examples are transcribed utterances spoken by different speakers of both feminine or masculine gender. Depending on the type of gender agreement occurring in an utterance, the speaker's gender and referents' gender did or did not correlate.

Gender Agreement: Whenever the speaker was the referent, i.e., the speaker was referring to him- or herself, there is speaker-*related* gender agreement among those gender-marked words referring to the speaker. Languages with a less pronounced inflection of gender, such as English, can encounter syntactic structures that do not indicate a speaker's gender (cf. bottom in Figure 1b). In contrast, syntactic structures of languages with rich gender-inflected systems typically encode enough information to unambiguously classify a speaker's gender (cf. top in Figure 1b). Consequently, we hypothesized that using English as a bridge language results in the loss of gender information for sentences with speaker-related gender agreement; meanwhile, the higher gender-inflected grammatical gender languages, German and Spanish, were hypothesized to preserve the gender information when used as a bridge language.

Whenever a person other than the speaker was the referent, i.e., the speaker was talking about someone else (e.g., “mi *padre* se sentía alienado$_M$” = “my dad felt alienated” uttered by a *female* speaker), there is speaker-*independent* gender agreement among those gender-marked words referring to the referent. For these examples in our data, meaning construction typically does not require the integration of semantic information about the speaker for correct syntactic processing and translation. The gender inflection of words is therefore often purely based on syntactic agreement with a formally marked subject (here, the referent), making the referent's gender identity explicit in those utterances for all three considered bridge languages, English, German, and Spanish.

4.2 Method of Measurement

Similar to Bentivogli et al. (2020), we used the concept of gender-swapping to measure how often a model preserved the gender compared to how often it produced the opposite gender form, thus opting for the wrong instead of the correct gender, which, if frequently done, signaled models' acting on gender biases.

Following this idea, models' generated translations of gender-marked words belonged to one of three categories, which we exemplify using Figure 2. First, the *expected translation*, for which we measured how often the *correct* translation (ground truth)—specified by a reference translation C-REF—was produced (e.g., “isolée” in the exemplary model output in Figure 2). Second, the *gender-reversed translation*, for which we measured how often the translation was *wrong*, but only regarding the gender inflection of gender-marked words—specified by a reference W-REF—i.e., instead of the required correct gender realization as per ground truth (e.g., the feminine adjective “intimidée”), the model produced the opposite gender form (e.g., the masculine adjective “intimidé”). Third, a *translation different from both reference translations*, e.g., instead of “jugée” (C-REF) or “jugé” (W-REF), the model produced the adjective “condamnée”, or any other word not matching C-REF or W-REF; in this case, we had no reference as to whether the gender inflection, regardless of the predicted word base, was correct or wrong, forcing us to exclude these translations from our gender bias evaluation.

We used two metrics to evaluate our models: BLEU (similar to Bentivogli et al. (2020)) and accuracy. For the accuracy on feminine and masculine word forms, we measured how often a model was able to produce the correct gender (C) for those words that matched either the correct or the wrong reference set (C+W); we refer to this as *gender preservation* (α_{correct}). As we only relied on correct and wrong “matches” (C+W)—excluding words that did not match any reference set (N)—the larger in size this set was, i.e., the larger the sample size, the more significant our findings; therefore, we weighted α_{correct} by the size of C+W in relation to the number of all translations (C+W+N), matching a reference (C+W) or not matching any reference (N); we refer to this weighting factor as *sample size* (ρ). Formally, we defined the accuracy γ to measure the *gender preservation performance weighted by the sample size* as follows:

$$\gamma = \frac{C}{\underbrace{C+W}_{\alpha_{\text{correct}}}} \cdot \frac{C+W}{\underbrace{C+W+N}_{\rho}} = \frac{C}{C+W+N}$$

To compare the performances for the two genders, we computed the *gender gap* δ between results for feminine and the masculine word forms:

$$\delta = 1 - \frac{\min(\gamma^{\text{F}}, \gamma^{\text{M}})}{\max(\gamma^{\text{F}}, \gamma^{\text{M}})}$$

As a reflection of gender biases, gender gaps should be as small as possible and ideally zero due to minimal differences between the results for the feminine and the masculine gender. Furthermore, we analyzed the difference between scores for the correct and the wrong references to determine whether translations were gender-biased.

5 Experiments & Results

The code and scripts used for our experimental evaluation are available on GitHub.[6]

[6] https://github.com/lenacabrera/gb_mnmt

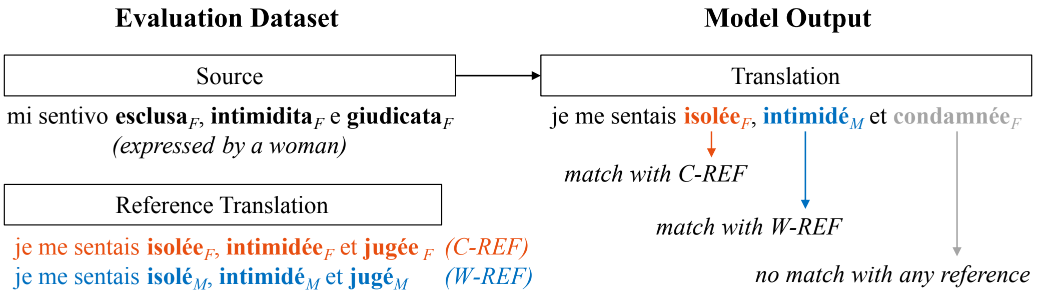

Evaluation Dataset
Model Output
Source
Translation
mi sentivo **esclusa**$_F$, **intimidita**$_F$ e **giudicata**$_F$
(expressed by a woman)
Reference Translation
je me sentais **isolée**$_F$, **intimidée**$_F$ et **jugée** $_F$ *(C-REF)*
je me sentais **isolé**$_M$, **intimidé**$_M$ et **jugé**$_M$ *(W-REF)*
je me sentais **isolée**$_F$, **intimidé**$_M$ et **condamnée**$_F$
match with C-REF
match with W-REF
no match with any reference

Figure 2: Illustration of the three possible translation outcomes of required gender preservation for Italian→French translation of the utterance "I felt *alienated*, *intimidated*, and *judged*": The translation of a gender-inflected word either matched the correct reference translation *C-REF* (here, "isolée" = alienated), the wrong reference translation *W-REF* (here, "intimidé" = intimidated), or neither (here, "condamnée" = condemned).

5.1 Experimental Setup

Training Data: In our experiments, we used the publicly available corpora MuST-C (Di Gangi et al., 2019) for model training. To investigate the impact of the bridge language, determined by the language pairs included during training, we formed three training corpora that are subsets of MuST-C (X),[7] with language pairs en↔X\en, de↔X\de, and es↔X\es, where X\en is the language set X excluding English (en), German (de), or Spanish (es). On each of the three corpora, we trained a model and afterward evaluated the three trained models on our evaluation data. Since only a portion (~10%) of MuST-C is true-parallel data, the training corpora differed in size, as specified in Table 2.

Language Pairs	# Sentences per Direction
en ↔ X\en	125,000–267,000
de ↔ X\de	103,000–223,000
es ↔ X\es	102,000–258,000

Table 2: Overview of the three MuST-C subsets used.

Preprocessing: MuST-C comes with partitioned training and validation sets which we kept unchanged in our experiments, except for the modifications described above. For the training and validation data, we first performed tokenization and truecasing using the Moses[8] tokenizer and truecaser. Afterward, we learned byte pair encoding (BPE) using subword-nmt[9] (Sennrich et al., 2016). We performed 20 thousand merge operations and only used tokens occurring in the training set with a minimum frequency of 50 times. Our evaluation data was preprocessed in a similar way using the BPE-learned vocabulary.

Training & Inference Details: Our baseline was a Transformer with 5 encoder and 5 decoder layers with 8 attention heads, an embedding size of 512, and an inner size of 2048. For regularization, we used dropout with a rate of 0.2 and performed label smoothing with a rate of 0.1. Moreover, we used the learning rate schedule from Vaswani et al. (2017) with 8,000 warmup steps (WUS). The source and target word embeddings were shared. To specify the output language, we used a target-language-specific beginning-of-sentence token. As part of our model modifications, we removed a residual connection (R) in the third encoder layer (Liu et al., 2021). We trained each model for 64 epochs and averaged the weights of the five best checkpoints ordered by the validation loss. For the auxiliary similarity loss (AUX_{SIM}) and the adversarial language classifier (ADV_{LAN}), we resumed training of the baseline and the model with removed residual connections for 10 additional epochs (400 WUS). By default, we only included supervised directions in the validation set. To compute BLEU scores, we used sacreBLEU (Post, 2018), which provides a fair and reproducible evaluation, as it operates on detokenized text.

5.2 Results

In Figure 3, we present the BLEU scores indicative of the similarity of the generated translations of MuST-SHE utterances to the *Correct* references and their gender-reversed counterparts (*Wrong* references) regardless of the referent gender, as well as the difference (delta) between *Cor-*

[7] From release version 1.2, we included 10 of the 15 available languages: Czech, Dutch, English, French, German, Italian, Portuguese, Romanian, Russian, and Spanish.

[8] https://github.com/moses-smt/mosesdecoder

[9] https://github.com/rsennrich/subword-nmt

rect and *Wrong* scores for zero-shot models only.[10]

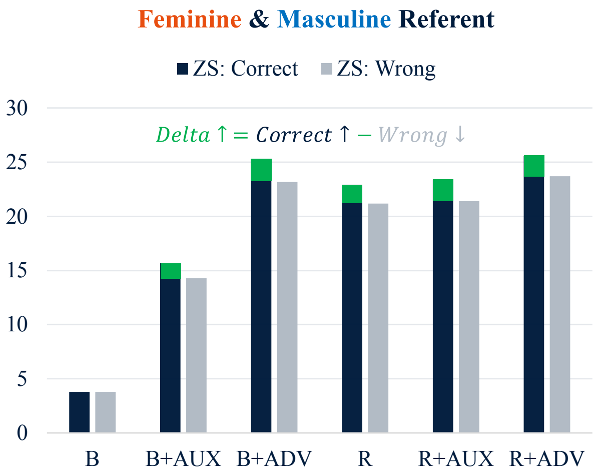


Figure 3: Average BLEU scores for *Correct* (left bar, higher ↑ is better) and *Wrong* (right bar, lower ↓ is better) MuST-SHE references of our six evaluated zero-shot models, complemented with the delta (green bar, higher ↑ is better) between both. Results are for the feminine and masculine referent gender.[10]

The bar graph illustrates that modifying our baseline B to encourage language-agnostic representations improves the poor gender preservation performance of B noticeably when performing zero-shot translation. While the delta between *Correct* and *Wrong* scores for B is zero, we consistently observe positive deltas (cf. green bars) that signal more correct than wrong gender translations; hence, through more language-agnostic hidden representations the modified zero-shot models more often can recover information (conveyed by the source language sentence) necessary to preserve the gender in the target language translation which, in turn, reduces the number of translations produced based on reflecting learned gender biases (in response to RQ3). It shows that $R+ADV_{LAN}$, closely followed by $B+ADV_{LAN}$, yields the highest *Correct* BLEU scores (higher is better) and one of the largest deltas between *Correct* and *Wrong* scores (higher is better); therefore, we take a closer look at the performance of $R + ADV_{LAN}$.

Complementary to the BLEU-based evaluation, we examine $R + ADV_{LAN}$ accuracies (γ), where better or worse performance measured is reliably attributed to better or worse translation of *gender-inflected words only*. From Figure 4, we can observe very similar performances for zero-shot and pivot-based translation using $R + ADV_{LAN}$ (RQ1). While both approaches achieve similar

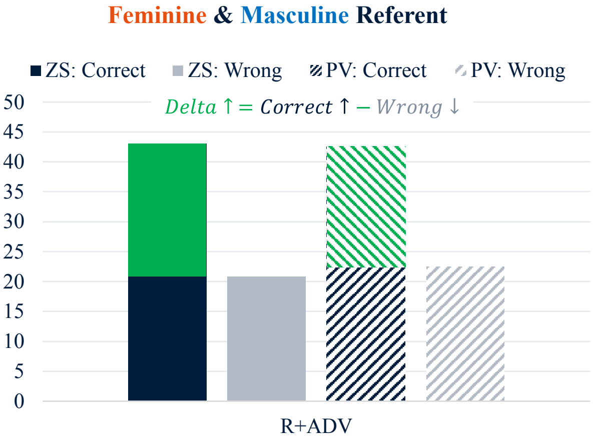


Figure 4: Average accuracy scores of zero-shot translation (full bars) and pivoting (hatched) for *Correct* (left bar, higher ↑ is better) and *Wrong* (right bar, lower ↓ is better) MuST-SHE references complemented with the delta (green bars, higher ↑ is better) between both for the model $R + ADV_{LAN}$. Results are for the feminine and masculine referent gender.[10]

Correct accuracy scores (43.0 for ZS and 42.5 for PV), we observe slightly lower *Wrong* scores for zero-shot translation (20.8) than for pivoting (22.5). As a result, the delta for zero-shot is higher (better) than for pivot-based translation (22.2 vs. 20.2).

To gain better insight into the difference in gender preservation between both approaches, we break down the accuracies and compare them for the feminine and masculine gender; the corresponding results are depicted in Figure 5. The large differences between the accuracies for feminine and masculine referents clearly show that the model is acting according to a *masculine bias* that detriments feminine and benefits masculine preservation of gender signals conveyed by the source sentence. The *Correct* accuracies in the masculine case are almost twice as high as their feminine counterparts. Furthermore, comparing the *Wrong* accuracies, we see an even bigger difference, as masculine *Wrong* scores are much smaller (by a factor of 5), whereas feminine *Wrong* scores are almost identical to their *Correct* counterparts.

In the masculine case, performances by both approaches are very similar, with pivoting achieving slightly higher *Correct* and *Wrong* scores (54.5 vs. 53.4 and 10.6 vs. 10.4). In the feminine case, we see that zero-shot translation is more accurate regarding feminine gender preservation: The delta between *Correct* and *Wrong* accuracies is small but positive (0.5), whereas for pivoting, we observe a negative delta (-4.9) that signals more wrong (masculine) than correct (feminine) trans-

[10] Results are for models trained on en↔X\en data.

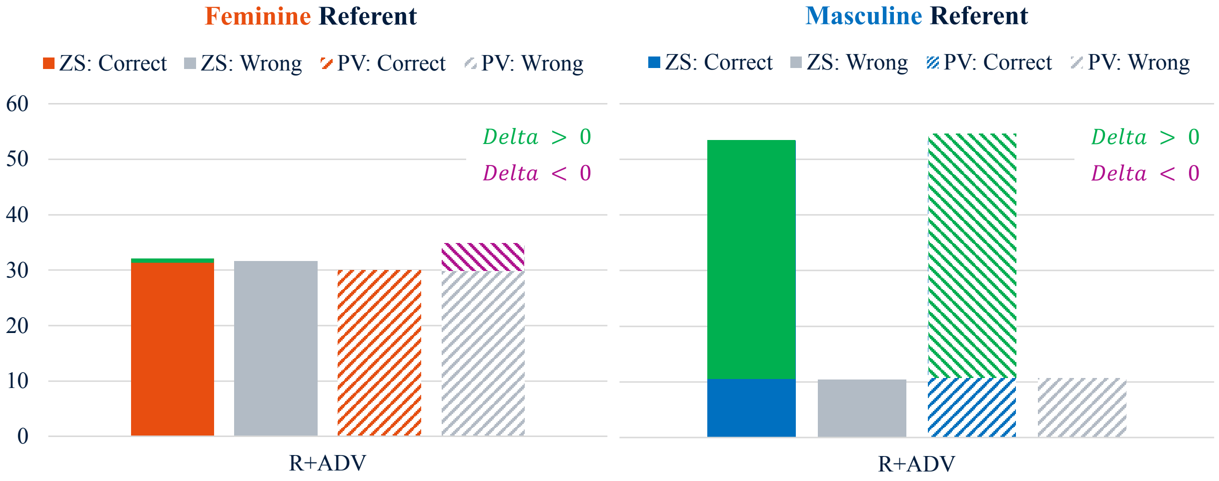


Figure 5: Average accuracy scores of zero-shot translation (full bars) and pivoting (hatched) for *Correct* (left bar, higher ↑ is better) and *Wrong* (right bar, lower ↓ is better) MuST-SHE references, complemented with the delta (green [*Delta* > 0] and magenta [*Delta* < 0] bars, higher ↑ is better) between both for the model $R + ADV_{LAN}$. Results are broken down by referent gender (feminine [left] vs. masculine [right]).[10]

lations for words where the required gender realization is feminine. Accordingly, it turns out that zero-shot translation performs noticeably better for feminine gender preservation—which is generally poorer than masculine gender preservation—compared to pivoting and, as a consequence, mitigates the masculine biases to a larger extent, producing more balanced gender outputs (RQ1).

As we assumed the bridge language to play an important role in gender preservation, we compare the model's performance for zero-shot and pivot-based translation when trained using different training corpora that enabled the use of different bridge languages, namely English (for the results presented so far) and the grammatical gender languages German and Spanish (in response to RQ2). As we expected to see differences between the three languages regarding sentences with and without speaker-related gender agreement, we present the *Correct* accuracies broken down by referent gender and complemented with the gender gap (δ) between feminine and masculine accuracies for either utterance category in Table 3.

Bridge Language	Feminine ↑		Masculine ↑		**Gender Gap ↓**	
	ZS	PV	ZS	PV	ZS	PV
Speaker-*Independent* Gender Agreement						
English	42.8	39.8	56.7	**58.3**	0.25	0.32
German	40.4	43.6	50.1	55.6	0.19	0.22
Spanish	**49.6**	45.3	57.7	55.0	**0.14**	0.18
Speaker-*Related* Gender Agreement						
English	20.2	19.2	48.2	48.7	0.58	0.61
German	15.1	18.4	**51.1**	49.8	0.70	0.63
Spanish	23.8	**29.4**	50.6	45.7	0.53	**0.36**

Table 3: Average accuracy scores for *Correct* (higher ↑ is better) references with speaker-related and speaker-independent gender agreement when bridging via English, German or Spanish using the model $R + ADV_{LAN}$. Results are broken down by referent gender and complemented with the gender gap (lower ↓ is better) between feminine and masculine accuracies. Underlined scores are the best of both approaches, and bold scores are the best across languages.

It shows that the performances for speaker-independent gender agreement are noticeably better (i.e., higher accuracies and smaller gender gaps) than for speaker-related gender agreement, which can be attributed to reduced gender ambiguity due to more explicit gender clues provided by source sentences in the former case. It shows that the poorer performance for speaker-related gender agreement affects the feminine gender more than the masculine gender when considering the much smaller difference in results for masculine word forms compared to a significant drop in scores for feminine word forms for speaker-related gender agreement (again, this very prominently highlights the model's masculine bias). Consequently, it shows that the feminine discrimination found throughout all models' performances is more prominent in cases of high gender ambiguity, confirming the notion of models making "educated" gender guesses that are tainted by gender biases.

Moreover, our results reveal clear differences in gender preservation between languages for both types of gender agreement: For speaker-independent gender agreement (e.g., “mi *padre* se sentía alienado$_M$” = “my dad felt alienated”), we find that zero-shot translation produces smaller gender gaps compared to pivoting for all three bridge languages. For the English bridge, the difference between zero-shot translation and pivoting is most pronounced, albeit small. For speaker-related gender agreement (e.g., “me sentí alienada$_F$” = “I felt alienated”), it turns out that zero-shot translation achieves a slightly smaller gender gap compared to pivoting using the English bridge language (where gender information is likely lost); for the German and the Spanish bridge languages, we observe better pivoting results regarding smaller gender gaps and, thus, more balanced correct gender outputs. This outcome confirms our hypothesis that for languages where gender inflection is relatively low, zero-shot translation is not as much affected by a loss of gender information (which impairs gender preservation for pivoting using discrete language representations), as it relies on more language-agnostic gender clues likely found in the continuous representations. Moreover, the outcomes suggest that with an increased level of gender inflection in the bridge language, pivoting surpasses zero-shot translation regarding fairly balanced gender preservation for speaker-related gender agreement.

6 Conclusion

In this paper, we explored gender bias in MNMT in the context of gender preservation for zero-shot translation directions, i.e., unseen language pairs (French↔Italian), compared the performances of pivoting and zero-shot translation using discrete and continuous representations respectively, studied the influence the bridge language has on both approaches, and examined the effect language-agnostic representations have on zero-shot models' gender biases. Based on our experimental results, we addressed three research questions.

RQ1 How do zero-shot and pivot-based translation compare regarding gender-biased outputs for zero-shot directions?

We find that zero-shot translation and pivoting achieve similar gender preservation performances, but zero-shot translation better preserves the feminine gender, which mitigates the masculine bias—the consistently worse feminine than masculine results across all evaluated models and both approaches—more than pivoting when bridging via English.

RQ2 Does the bridge language affect the gender biases perpetuated by zero-shot and pivot-based translations?

Our experiments revealed that the bridge language affects gender biases in MNMT. For English, a language limited to pronominal gender (with a few exceptions), we find that zero-shot translation performs better than pivoting regarding a more fairly balanced preservation of feminine and masculine gender. Using two richer gender-inflected bridge languages, Spanish and German, revealed that with an increased level of gender inflection in the bridge language, pivoting surpasses zero-shot translation regarding fewer gender-biased outputs for utterances with speaker-related gender agreement.

RQ3 Do translation quality improvements of zero-shot models reduce their gender biases?

All three evaluated modifications encouraging language-agnostic hidden representations (cf. Section 3) improved zero-shot models' ability to preserve the feminine and masculine gender and reduced the gap between better masculine and worse feminine results; they improved zero-shot models' performances to the point where they outperformed pivoting regarding more fairly balanced preservation of both genders when bridging via English.

Besides our findings, this work also features some limitations that can be addressed in future work. First, the data used in our experimental evaluation limited the scenarios to those examined. Future work can examine the translation of sentences with mixed gender (i.e., sentences including feminine *and* masculine word forms) and directions, including languages from different language families and with different gender systems, to further study language differences. Second, developing a large-scale gender-annotated corpus suitable for MNMT training could most likely be used to improve models' gender preservation performance. A well-performing gender classifier could be used to annotate the MuST-C dataset with token- or word-level gender labels. Third, we believe that the metrics currently used to evaluate models' gender biases are not ideal. For instance, model outputs mismatching the reference translations used for evaluation are discarded, despite potentially being appropriate translations (e.g., synonyms); future work could explore using additional morphological analysis tools to include those translations in the gender bias evaluation. Generally, inquiring about the phenomenon of gender bias in translation requires appropriate and established metrics; the lack thereof currently leaves room for improvement in evaluative procedures.

While there is a lot of potential for further research on this topic, it is crucial to acknowledge that, ultimately, translation technology is bound by the principles of language, which subtly reproduces societal asymmetries and embeds signs of sexism, including masculine defaults and more subtle conventions by which expressions referring to females are grammatically more complex in many languages. Consequently, combating gender biases in translation technology requires awareness of language use, as it is one of the most powerful means through which sexism and gender discrimination are perpetrated and reproduced.

References

Ackerman, Lauren M. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa*.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Basta, Christine, Marta R Ruiz Costa-jussà, and José Adrián Rodríguez Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the 4th Widening Natural Language Processing Workshop*, pages 99–102, Online.

Bentivogli, Luisa, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? Evaluating speech translation technology on the MuST-SHE corpus. *arXiv preprint arXiv:2006.05754*.

Cao, Yang Trista and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Computational Linguistics*, 47(3):615–661, November.

Corbett, Greville. 1991. *Gender*. Cambridge University Press.

Di Gangi, Mattia A., Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 2012–2017, Minneapolis, Minnesota.

Liu, Danni, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. 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 1259–1273, Online.

Moryossef, Amit, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. *arXiv preprint arXiv:1903.03467*.

Pham, Ngoc-Quan, Jan Niehues, Thanh-Le Ha, and Alex Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the 4th Conference on Machine Translation. Vol. 1. Ed.: O. Bojar*, pages 13–23.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*.

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

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1715–1725, Berlin, Germany.

Vanmassenhove, Eva, 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.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.