How to Split: the Effect of Word Segmentation on Gender Bias in Speech Translation

Marco Gaido[1,2][†], Beatrice Savoldi[1,2][†], Luisa Bentivogli[1], Matteo Negri[1], Marco Turchi[1]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Italy
{mgaido,bsavoldi,bentivo,negri,turchi}@fbk.eu

Abstract

Having recognized gender bias as a major issue affecting current translation technologies, researchers have primarily attempted to mitigate it by working on the data front. However, whether algorithmic aspects concur to exacerbate unwanted outputs remains so far under-investigated. In this work, we bring the analysis on gender bias in automatic translation onto a seemingly neutral yet critical component: word segmentation. Can segmenting methods influence the ability to translate gender? Do certain segmentation approaches penalize the representation of feminine linguistic markings? We address these questions by comparing 5 existing segmentation strategies on the target side of speech translation systems. Our results on two language pairs (English-Italian/French) show that state-of-the-art sub-word splitting (BPE) comes at the cost of higher gender bias. In light of this finding, we propose a combined approach that preserves BPE overall translation quality, while leveraging the higher ability of character-based segmentation to properly translate gender.

Bias Statement.[1]   We study the effect of segmentation methods on the ability of speech translation (ST) systems to translate masculine and feminine forms referring to human entities. In this area, structural linguistic properties interact with the perception and representation of individuals (Gygax et al., 2019; Corbett, 2013; Stahlberg et al., 2007). Thus, we believe they are relevant gender expressions, used to communicate about the self and others, and by which the sociocultural and political reality of gender is negotiated (Hellinger and Motschenbacher, 2015).

[†]The authors contributed equally.
[1]As suggested by (Blodgett et al., 2020) and required for other venues (Hardmeier et al., 2021), we formulate our bias statement.

Accordingly, we consider a model that systematically and disproportionately favours masculine over feminine forms to be biased, as it fails to properly recognize women. From a technical perspective, such behaviour deteriorates models' performance. Most importantly, however, from a human-centered view, real-world harms are at stake (Crawford, 2017), as translation technologies are unequally beneficial across gender groups and reduce feminine visibility, thus contributing to misrepresent an already socially disadvantaged group.

This work is motivated by the intent to shed light on whether issues in the generation of feminine forms are also a by-product of current algorithms and techniques. In our view, architectural improvements of ST systems should also account for the trade-offs between overall translation quality and gender representation: our proposal of a model that combines two segmentation techniques is a step towards this goal.

Note that technical mitigation approaches should be integrated with the long-term multidisciplinary commitment (Criado-Perez, 2019; Benjamin, 2019; D'Ignazio and Klein, 2020) necessary to radically address bias in our community. Also, we recognize the limits of working on binary gender, as we further discuss in the ethic section (§8).

1 Introduction

The widespread use of language technologies has motivated growing interest on their social impact (Hovy and Spruit, 2016; Blodgett et al., 2020), with gender bias representing a major cause of concern (Costa-jussà, 2019; Sun et al., 2019). As regards translation tools, focused evaluations have exposed that speech translation (ST) – and machine translation (MT) – models do in fact overproduce masculine references in their outputs (Cho et al., 2019; Bentivogli et al., 2020), except for feminine asso-

ciations perpetuating traditional gender roles and stereotypes (Prates et al., 2020; Stanovsky et al., 2019). In this context, most works identified *data* as the primary source of gender asymmetries. Accordingly, many pointed out the misrepresentation of gender groups in datasets (Garnerin et al., 2019; Vanmassenhove et al., 2018), focusing on the development of data-centred mitigating techniques (Zmigrod et al., 2019; Saunders and Byrne, 2020).

Although data are not the only factor contributing to generate bias (Shah et al., 2020; Savoldi et al., 2021), only few inquiries devoted attention to other technical components that exacerbate the problem (Vanmassenhove et al., 2019) or to architectural changes that can contribute to its mitigation (Costa-jussà et al., 2020b). From an algorithmic perspective, Roberts et al. (2020) additionally expose how "taken-for-granted" approaches may come with high overall translation quality in terms of BLEU scores, but are actually detrimental when it comes to gender bias.

Along this line, we focus on ST systems and inspect a core aspect of neural models: word segmentation. Byte-Pair Encoding (BPE) (Sennrich et al., 2016) represents the *de-facto* standard and has recently shown to yield better results compared to character-based segmentation in ST (Di Gangi et al., 2020). But does this hold true for gender translation as well? If not, why?

Languages like French and Italian often exhibit comparatively complex feminine forms, derived from the masculine ones by means of an additional suffix (e.g. en: *professor*, fr: *professeur* M vs. *professeure* F). Additionally, women and their referential linguistic expressions of gender are typically under-represented in existing corpora (Hovy et al., 2020). In light of the above, purely statistical segmentation methods could be unfavourable for gender translation, as they can break the morphological structure of words and thus lose relevant linguistic information (Ataman et al., 2017). Indeed, as BPE merges the character sequences that co-occur more frequently, rarer or more complex feminine-marked words may result in less compact sequences of tokens (e.g. en: *described*, it: *des@@critto* M vs. *des@@crit@@ta* F). Due to such typological and distributive conditions, may certain splitting methods render feminine gender less probable and hinder its prediction?

We address such questions by implementing different families of segmentation approaches em-

ployed on the decoder side of ST models built on the same training data. By comparing the resulting models both in terms of overall translation quality and gender accuracy, we explore whether a so far considered irrelevant aspect like word segmentation can actually affect gender translation. As such, **(1)** we perform the first comprehensive analysis of the results obtained by 5 popular segmentation techniques for two language directions (en-fr and en-it) in ST. **(2)** We find that the target segmentation method is indeed an important factor for models' gender bias. Our experiments consistently show that BPE leads to the highest BLEU scores, while character-based models are the best at translating gender. Preliminary analyses suggests that the isolation of the morphemes encoding gender can be a key factor for gender translation. **(3)** Finally, we propose a multi-decoder architecture able to combine BPE overall translation quality and the higher ability to translate gender of character-based segmentation.

2 Background

Gender bias. Recent years have seen a surge of studies dedicated to gender bias in MT (Gonen and Webster, 2020; Rescigno et al., 2020) and ST (Costa-jussà et al., 2020a). The primary source of such gender imbalance and adverse outputs has been identified in the training data, which reflect the under-participation of women – e.g. in the media (Madaan et al., 2018), sexist language and gender categories overgeneralization (Devinney et al., 2020). Hence, preventive initiatives concerning data documentation have emerged (Bender and Friedman, 2018), and several mitigating strategies have been proposed by training models on *ad-hoc* gender-balanced datasets (Saunders and Byrne, 2020; Costa-jussà and de Jorge, 2020), or by enriching data with additional gender information (Moryossef et al., 2019; Vanmassenhove et al., 2018; Elaraby and Zahran, 2019; Saunders et al., 2020; Stafanovičs et al., 2020).

Comparatively, very little work has tried to identify concurring factors to gender bias going beyond data. Among those, Vanmassenhove et al. (2019) ascribes to an algorithmic bias the loss of less frequent feminine forms in both phrase-based and neural MT. Closer to our intent, two recent works pinpoint the impact of models' components and inner mechanisms. Costa-jussà et al. (2020b) investigate the role of different architectural de-

signs in multilingual MT, showing that language-specific encoder-decoders (Escolano et al., 2019) better translate gender than shared models (Johnson et al., 2017), as the former retain more gender information in the source embeddings and keep more diversion in the attention. Roberts et al. (2020), on the other hand, prove that the adoption of beam search instead of sampling – although beneficial in terms of BLEU scores – has an impact on gender bias. Indeed, it leads models to an extreme operating point that exhibits zero variability and in which they tend to generate the more frequent (masculine) pronouns. Such studies therefore expose largely unconsidered aspects as factors contributing to gender bias in automatic translation, identifying future research directions for the needed countermeasures.

To the best of our knowledge, no prior work has taken into account if it may be the case for segmentation methods as well. Rather, prior work in ST (Bentivogli et al., 2020) compared gender translation performance of cascade and direct systems using different segmentation algorithms, disregarding their possible impact on final results.

Segmentation. Although early attempts in neural MT employed word-level sequences (Sutskever et al., 2014; Bahdanau et al., 2015), the need for open-vocabulary systems able to translate rare/unseen words led to the definition of several word segmentation techniques. Currently, the statistically motivated approach based on byte-pair encoding (BPE) by Sennrich et al. (2016) represents the *de facto* standard in MT. Recently, its superiority to character-level (Costa-jussà and Fonollosa, 2016; Chung et al., 2016) has been also proved in the context of ST (Di Gangi et al., 2020). However, depending on the languages involved in the translation task, the data conditions, and the linguistic properties taken into account, BPE greedy procedures can be suboptimal. By breaking the surface of words into plausible semantic units, linguistically motivated segmentations (Smit et al., 2014; Ataman et al., 2017) were proven more effective for low-resource and morphologically-rich languages (e.g. agglutinative languages like Turkish), which often have a high level of sparsity in the lexical distribution due to their numerous derivational and inflectional variants. Moreover, fine-grained analyses comparing the grammaticality of character, morpheme and BPE-based models exhibited different capabilities. Sennrich (2017) and Ataman et al. (2019) show the syntactic advantage of BPE

in managing several agreement phenomena in German, a language that requires resolving long range dependencies. In contrast, Belinkov et al. (2020) demonstrate that while subword units better capture semantic information, character-level representations perform best at generalizing morphology, thus being more robust in handling unknown and low-frequency words. Indeed, using different atomic units does affect models' ability to handle specific linguistic phenomena. However, whether low gender translation accuracy can be to a certain extent considered a by-product of certain compression algorithms is still unknown.

3 Language Data

As just discussed, the effect of segmentation strategies can vary depending on language typology (Ponti et al., 2019) and data conditions. To inspect the interaction between word segmentation and gender expressions, we thus first clarify the properties of grammatical gender in the two languages of our interest: French and Italian. Then, we verify their representation in the datasets used for our experiments.

3.1 Languages and Gender

The extent to which information about the gender of referents is grammatically encoded varies across languages (Hellinger and Motschenbacher, 2015; Gygax et al., 2019). Unlike English – whose gender distinction is chiefly displayed via pronouns (e.g. *he/she*) – fully grammatical gendered languages like French and Italian systematically articulate such semantic distinction on several parts of speech (gender agreement) (Hockett, 1958; Corbett, 1991). Accordingly, many lexical items exist in both feminine and masculine variants, overtly marked through morphology (e.g. en: *the tired kid sat down*; it: *il bimbo stanco si è seduto* M vs. *la bimba stanca si è seduta* F). As the example shows, the word forms are distinguished by two morphemes (*–o, –a*), which respectively represent the most common inflections for Italian masculine and feminine markings.[2] In French, the morphological mechanism is slightly different (Schafroth, 2003), as it relies on an additive suffixation on top of masculine words to express feminine gender (e.g. en: *an expert is gone*, fr: *un expert est allé* M vs.

---

[2]In a fusional language like Italian, one single morpheme can denote several properties as, in this case, gender and singular number (the plural forms would be *bimbi* vs. *bimbe*).

une experte est allée F). Hence, feminine French forms require an additional morpheme. Similarly, another productive strategy – typical for a set of personal nouns – is the derivation of feminine words via specific affixes for both French (*e.g. –eure, –euse*)[3] and Italian (*–essa, –ina, –trice*) (Schafroth, 2003; Chini, 1995), whose residual evidence is still found in some English forms (e.g. *heroine, actress*) (Umera-Okeke, 2012).

In light of the above, translating gender from English into French and Italian poses several challenges to automatic models. First, gender translation does not allow for one-to-one mapping between source and target words. Second, the richer morphology of the target languages increases the number of variants and thus data sparsity. Hereby, the question is whether – and to what extent – statistical word segmentation differently treats the less frequent variants. Also, considering the morphological complexity of some feminine forms, we speculate whether linguistically unaware splitting may disadvantage their translation. To test these hypotheses, below we explore if such conditions are represented in the ST datasets used in our study.

3.2 Gender in Used Datasets

MuST-SHE (Bentivogli et al., 2020) is a gender-sensitive benchmark available for both en-fr and en-it (1,113 and 1,096 sentences, respectively). Built on naturally occurring instances of gender phenomena retrieved from the TED-based MuST-C corpus (Cattoni et al., 2020),[4] it allows to evaluate gender translation on qualitatively differentiated and balanced masculine/feminine forms. An important feature of MuST-SHE is that, for each reference translation, an almost identical "wrong" reference is created by swapping each annotated gender-marked word into its opposite gender. By means of such wrong reference, for each target language we can identify ~2,000 pairs of gender forms (e.g. en: *tired*, fr: *fatiguée* vs. *fatigué*) that we compare in terms of *i)* length, and *ii)* frequency in the MuST-C training set.

As regards frequency, we asses that, for both language pairs, the types of feminine variants are less frequent than their masculine counterpart in over 86% of the cases. Among the exceptions, we find words that are almost gender-exclusive

(e.g. *pregnant*) and some problematic or socially connoted activities (e.g. *raped, nurses*). Looking at words' length, 15% of Italian feminine forms result to be longer than masculine ones, whereas in French this percentage amounts to almost 95%. These scores confirm that MuST-SHE reflects the typological features described in §3.1.

4 Experiments

All the direct ST systems used in our experiments are built in the same fashion within a controlled environment, so to keep the effect of different word segmentations as the only variable. Accordingly, we train them on the MuST-C corpus, which contains 492 hours of speech for en-fr and 465 for en-it. Concerning the architecture, our models are based on Transformer (Vaswani et al., 2017). For the sake of reproducibility, we provide extensive details about the ST models and hyper-parameters' choices in the Appendix §*A*.[5]

4.1 Segmentation Techniques

To allow for a comprehensive comparison of word segmentation's impact on gender bias in ST, we identified three substantially different categories of splitting techniques. For each of them, we hereby present the candidates selected for our experiments.

Character Segmentation. Dissecting words at their maximal level of granularity, **character-based** solutions have been first proposed by Ling et al. (2015) and Costa-jussà and Fonollosa (2016). This technique proves simple and particularly effective at generalizing over unseen words. On the other hand, the length of the resulting sequences increases the memory footprint, and slows both the training and inference phases. We perform our segmentation by appending "@@ " to all characters but the last of each word.

Statistical Segmentation. This family comprises data-driven algorithms that generate statistically significant subwords units. The most popular one is **BPE** (Sennrich et al., 2016),[6] which proceeds by merging the most frequently co-occurring characters or character sequences. Recently, He et al. (2020) introduced the Dynamic Programming Encoding (**DPE**) algorithm, which performs

---

[3]French also requires additional modification on feminine forms due to phonological rules (e.g. en: *chef/spy*, fr: *cheffe*/*espionne* vs. *chef*/*espion*).

[4]Further details about these datasets are provided in §8.

[5]Source code available at https://github.com/mgaido91/FBK-fairseq-ST/tree/acl_2021.

[6]We use SentencePiece (Kudo and Richardson, 2018): https://github.com/google/sentencepiece.

competitively and was claimed to accidentally produce more linguistically-plausible subwords with respect to BPE. DPE is obtained by training a mixed character-subword model. As such, the computational cost of a DPE-based ST model is around twice that of a BPE-based one. We trained the DPE segmentation on the transcripts and the target translations of the MuST-C training set, using the same settings of the original paper.[7]

Morphological Segmentation. A third possibility is linguistically-guided tokenization that follows morpheme boundaries. Among the unsupervised approaches, one of the most widespread tools is **Morfessor** (Creutz and Lagus, 2005), which was extended by Ataman et al. (2017) to control the size of the output vocabulary, giving birth to the **LMVR** segmentation method. These techniques have outperformed other approaches when dealing with low-resource and/or morphologically-rich languages (Ataman and Federico, 2018). In other languages, they are not as effective, so they are not widely adopted. Both Morfessor and LMVR have been trained on the MuST-C training set.[8]

	en-fr	en-it
# tokens[9]	5.4M	4.6M
# types	96K	118K
BPE	8,048	8,064
Char	304	256
DPE	7,864	8,008
Morfessor	26,728	24,048
LMVR	21,632	19,264

Table 1: Resulting dictionary sizes.

For fair comparison, we chose the optimal vocabulary size for each method (when applicable). Following (Di Gangi et al., 2020), we employed 8k merge rules for BPE and DPE, since the latter requires an initial BPE segmentation. In LMVR, instead, the desired target dimension is actually only an upper bound for the vocabulary size. We tested 32k and 16k, but we only report the results with 32k as it proved to be the best configuration both in terms of translation quality and gender accuracy. Finally, character-level segmentation and Morfessor do not allow to determine the vocabulary size. Table 1 shows the size of the resulting dictionaries.

---

[7]See https://github.com/xlhex/dpe.

[8]We used the parameters and commands suggested in https://github.com/d-ataman/lmvr/blob/master/examples/example-train-segment.sh

[9]Here "tokens" refers to the number of words in the corpus, and not to the unit resulting from subword tokenization.

	en-fr			en-it		
	M-C	M-SHE	Avg.	M-C	M-SHE	Avg.
BPE	**30.7**	25.9	**28.3**	21.4	**21.8**	21.6
Char	29.5	24.2	26.9	21.3	20.7	21.0
DPE	29.8	25.3	27.6	21.9	21.7	**21.8**
Morfessor	29.7	25.7	27.7	21.7	21.4	21.6
LMVR	30.3	**26.0**	28.2	**22.0**	21.5	**21.8**

Table 2: SacreBLEU scores on MuST-C tst-COMMON (M-C) and MuST-SHE (M-SHE) for en-fr and en-it.

4.2 Evaluation

We are interested in measuring both *i)* the overall translation quality obtained by different segmentation techniques, and *ii)* the correct generation of gender forms. We evaluate translation quality on both the MuST-C *tst-COMMON* set (2,574 sentences for en-it and 2,632 for en-fr) and MuST-SHE (§3.2), using SacreBLEU (Post, 2018).[10]

For fine-grained analysis on gender translation, we rely on gender accuracy (Gaido et al., 2020).[11] We differentiate between two categories of phenomena represented in MuST-SHE. Category (1) contains first-person references (e.g. *I'm a student*) to be translated according to the speakers' preferred linguistic expression of gender. In this context, ST models can leverage speakers' vocal characteristics as a gender cue to infer gender translation.[12] Gender phenomena of Category (2), instead, shall be translated in concordance with other gender information in the sentence (e.g. *she/he is a student)*.

5 Comparison of Segmentation Methods

Table 2 shows the overall **translation quality** of ST systems trained with distinct segmentation techniques. BPE comes out as competitive as LMVR for both language pairs. On averaged results, it exhibits a small gap (0.2 BLEU) also with DPE on en-it, while it achieves the best performance on en-fr. The disparities are small though: they range within 0.5 BLEU, apart from Char standing ∼1 BLEU below. Compared to the scores reported by Di Gangi et al. (2020), the Char gap is however smaller. As our results are considerably higher than theirs, we believe that the reason for such differences lies in a sub-optimal fine-tuning of their hyper-parameters. Overall, in light of the trade-off between computational cost (LMVR and DPE require a dedicated training phase for data segmen-

---

[10]BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.3.

[11]Evaluation script available with the MuST-SHE release.

[12]Although they do not emerge in our experimental settings, the potential risks of such capability are discussed in §8.

tation) and average performance (BPE achieves winning scores on en-fr and competitive for en-it), we hold BPE as the best segmentation strategy in terms of general translation quality for direct ST.

Turning to **gender translation**, the gender accuracy scores presented in Table 3 exhibit that all ST models are clearly biased, with masculine forms (M) disproportionately produced across language pairs and categories. However, we intend to pinpoint the relative gains and losses among segmenting methods. Focusing on overall accuracy (ALL), we see that Char – despite its lowest performance in terms of BLEU score – emerges as the favourite segmentation for gender translation. For French, however, DPE is only slightly behind. Looking at morphological methods, they surprisingly do not outperform the statistical ones. The greatest variations are detected for feminine forms of Category 1 (1F), where none of the segmentation techniques reaches 50% of accuracy, meaning that they are all worse than a random choice when the speaker should be addressed by feminine expressions. Char appears close to such threshold, while the others (apart from DPE in French) are significantly lower.

These results illustrate that target segmentation is a relevant parameter for gender translation. In particular, they suggest that Char segmentation improves models' ability to learn correlations between the received input and gender forms in the reference translations. Although in this experiment models rely only on speakers' vocal characteristics to infer gender – which we discourage as a cue for gender translation for real-world deployment (see §8) – such ability shows a potential advantage for Char, which could be better redirected toward learning correlations with reliable gender meta-information included in the input. For instance, in a scenario in which meta-information (e.g. a gender tag) is added to the input to support gender translation, a Char model might better exploit this information. Lastly, our evaluation reveals that, unlike previous ST studies (Bentivogli et al., 2020), a proper comparison of models' gender translation potentialities requires adopting the same segmentation. Our question then becomes: What makes Char segmentation less biased? What are the tokenization features determining a better/worse ability in generating the correct gender forms?

Lexical diversity. We posit that the limited generation of feminine forms can be framed as an issue of data sparsity, whereas the advantage of

	en-fr				
	ALL	1F	1M	2F	2M
BPE	65.18	37.17	75.44	61.20	80.80
Char	**68.85**	48.21	74.78	65.89	**81.03**
DPE	68.55	**49.12**	70.29	**66.22**	80.90
Morfessor	67.05	42.73	75.11	63.02	80.98
LMVR	65.38	32.89	**76.96**	61.87	79.95
	en-it				
BPE	67.47	33.17	88.50	60.26	81.82
Char	**71.69**	**48.33**	85.07	**64.65**	**84.33**
DPE	68.86	44.83	81.58	59.32	82.62
Morfessor	65.46	36.61	81.04	56.94	79.61
LMVR	69.77	39.64	**89.00**	63.85	83.03

Table 3: Gender accuracy (%) for MuST-SHE Overall (ALL), Category 1 and 2 on en-fr and en-it.

Char-based segmentation ensues from its ability to handle less frequent and unseen words (Belinkov et al., 2020). Accordingly, Vanmassenhove et al. (2018); Roberts et al. (2020) link the loss of *linguistic diversity* (i.e. the range of lexical items used in a text) with the overfitted distribution of masculine references in MT outputs.

To explore such hypothesis, we compare the lexical diversity (LD) of our models' translations and MuST-SHE references. To this aim, we rely on Type/Token ratio (**TTR**) – (Chotlos, 1944; Templin, 1957), and the more robust Moving Average TTR (**MATTR**) – (Covington and McFall, 2010).[13]

As we can see in Table 4, character-based models exhibit the highest LD (the only exception is DPE with the less reliable TTR metric on en-it). However, we cannot corroborate the hypothesis formulated in the above-cited studies, as LD scores do not strictly correlate with gender accuracy (Table 3). For instance, LMVR is the second-best in terms of gender accuracy on en-it, but shows a very low lexical diversity (the worst according to MATTR and second-worst according to TTR).

	en-fr		**en-it**	
	TTR	MATTR	TTR	MATTR
M-SHE Ref	*16.12*	*41.39*	*19.11*	*46.36*
BPE	14.53	39.69	17.46	44.86
Char	**14.97**	**40.60**	17.75	**45.65**
DPE	14.83	40.02	**18.07**	45.12
Morf	14.38	39.88	16.31	44.90
LMVR	13.87	39.98	16.33	44.71

Table 4: Lexical diversity scores on en-fr and en-it

Sequence length. As discussed in §3, we know that Italian and French feminine forms are, al-

---

[13]Metrics computed with software available at: https://github.com/LSYS/LexicalRichness. We set 1,000 as `window_size` for MATTR.

	en-fr (%)	en-it (%)
BPE	1.04	0.88
Char	1.37	0.38
DPE	2.11	0.77
Morfessor	1.62	0.45
LMVR	1.43	0.33

Table 5: Percentage increase of token sequence's length for feminine words over masculine ones.

	en	Segm.	F	M	Freq. F/M
a)	asked	BPE	chie–sta	chiesto	36/884
b)		DPE	chie–sta	chiesto	36/884
c)	friends	BPE	a–miche	amici	49/1094
d)		DPE	a–miche	amici	49/1094
e)	adopted	BPE	adop–tée	adop–té	30/103
f)		DPE	adop–t–é–e	adop–t–é	30/103
g)	sure	Morf.	si–cura	sicuro	258/818
h)	grown up	LMVR	cresci–uta	cresci–uto	229/272
i)	celebrated	LMVR	célébr–ées	célébr–és	3/7

Table 6: Examples of word segmentation. The segmentation boundary is identified by ”–”.

though to different extent, longer and less frequent than their masculine counterparts. In light of such conditions, we expected that the statistically-driven BPE segmentation would leave feminine forms unmerged at a higher rate, and thus add uncertainty to their generation. To verify if this is the actual case – explaining BPE's lower gender accuracy – we check whether the number of tokens (characters or subwords) of a segmented feminine word is higher than that of the corresponding masculine form. We exploit the coupled "wrong" and "correct" references available in MuST-SHE, and compute the average percentage of additional tokens found in the feminine segmented sentences[14] over the masculine ones. Results are reported in Table 5.

At a first look, we observe opposite trends: BPE segmentation leads to the highest increment of tokens for feminine words in Italian, but to the lowest one in French. Also, DPE exhibits the highest increment in French, whereas it actually performs slightly better than Char on feminine gender translation (see Table 3). Hence, even the increase in sequence length does not seem to be an issue on its own for gender translation. Nonetheless, these apparently contradictory results encourage our last exploration: *How* are gender forms actually split?

Gender isolation. By means of further manual analysis on 50 output sentences per each of the 6 systems, we inquire if longer token sequences for feminine words can be explained in light of the different characteristics and gender productive mechanisms of the two target languages (§3.1). Table 6 reports selected instances of coupled feminine/masculine segmented words, with their respective frequency in the MuST-C training set.

Starting with Italian, we find that BPE sequence length increment indeed ensues from greedy splitting that, as we can see from examples **(a)** and **(c)**, ignores meaningful affix boundaries for both same length and different-length gender pairs, respec-

---

[14] As such references only vary for gender-marked words, we can isolate the difference relative to gender tokens.

tively. Conversely, on the French set – with 95% of feminine words longer than their masculine counterparts – BPE's low increment is precisely due to its loss of semantic units. For instance, as shown in **(e)**, BPE does not preserve the verb root (*adopt*), nor isolates the additional token (*-e*) responsible for the feminine form, thus resulting into two words with the same sequence length (2 tokens). Instead DPE, which achieved the highest accuracy results for en-fr feminine translation (Table 3), treats the feminine additional character as a token *per se* (**f**).

Based on such patterns, our intuition is that the proper splitting of the morpheme-encoded gender information as a distinct token favours gender translation, as models learn to productively generalize on it. Considering the high increment of DPE tokens for Italian in spite of the limited number of longer feminine forms (15%), our analysis confirms that DPE is unlikely to isolate gender morphemes on the en-it language pair. As a matter of fact, it produces the same kind of coarse splitting as BPE (see **(b)** and **(d)**).

Finally, we attest that the two morphological techniques are not equally valid. Morfessor occasionally generates morphologically incorrect subwords for feminine forms by breaking the word stem (see example **(g)** where the correct stem is *sicur*). Such behavior also explains Morfessor's higher token increment with respect to LMVR. Instead, although LMVR (examples **(h)** and **(i)**) produces linguistically valid suffixes, it often condenses other grammatical categories (e.g. tense and number) with gender. As suggested above, if the pinpointed split of morpheme-encoded gender is a key factor for gender translation, LMVR's lower level of granularity explains its reduced gender accuracy. Working on character' sequences, instead, the isolation of the gender unit is always attained.

	en-fr			en-it		
	M-C	M-SHE	Avg.	M-C	M-SHE	Avg.
BPE	**30.7**	25.9	28.3	21.4	21.8	21.6
Char	29.5	24.2	26.9	21.3	20.7	21.0
BPE&Char	30.4	**26.5**	**28.5**	22.1	22.6	22.3

Table 7: SacreBLEU scores on MuST-C tst-COMMON (M-C) and MuST-SHE (M-SHE) on en-fr and en-it.

	en-fr				
	ALL	1F	1M	2F	2M
BPE	65.18	37.17	**75.44**	61.20	80.80
Char	**68.85**	**48.21**	74.78	65.89	81.03
BPE&Char	68.04	40.61	75.11	**67.01**	**81.45**
	en-it				
BPE	67.47	33.17	**88.50**	60.26	81.82
Char	**71.69**	48.33	85.07	**64.65**	**84.33**
BPE&Char	70.05	**52.23**	84.19	59.60	81.37

Table 8: Gender accuracy (%) for MuST-SHE Overall (ALL), Category 1 and 2 on en-fr and en-it.

6 Beyond the Quality-Gender Trade-off

Informed by our experiments and analysis (§5), we conclude this study by proposing a model that combines BPE overall translation quality and Char's ability to translate gender. To this aim, we train a multi-decoder approach that exploits both segmentations to draw on their corresponding advantages.

In the context of ST, several multi-decoder architectures have been proposed, usually to jointly produce both transcripts and translations with a single model. Among those in which both decoders access the encoder output, here we consider the best performing architectures according to Sperber et al. (2020). As such, we consider: *i) Multitask direct*, a model with one encoder and two decoders, both exclusively attending the encoder output as proposed by Weiss et al. (2017), and *ii)* the *Triangle* model (Anastasopoulos and Chiang, 2018), in which the second decoder attends the output of both the encoder and the first decoder.

For the triangle model, we used a first BPE-based decoder and a second Char decoder. With this order, we aimed to enrich BPE high quality translation with a refinement for gender translation, performed by the Char-based decoder. However, the results were negative: the second decoder seems to excessively rely on the output of the first one, thus suffering from a severe *exposure bias* (Ranzato et al., 2016) at inference time. Hence, we do not report the results of these experiments.

Instead, the *Multitask direct* has one BPE-based and one Char-based decoder. The system requires a training time increase of only 10% and 20% compared to, respectively, Char and BPE models. At inference phase, instead, running time and size are the same of a BPE model. We report overall translation quality (Table 7) and gender accuracy (Table 8) of the BPE output (BPE&Char).[15] Starting with gender accuracy, the Multitask model's overall gender translation ability (ALL) is still lower, although very close, to that of the Char-based model. Nevertheless, feminine translation improvements are present on Category 2F for en-fr and, with a larger gain, on 1F for en-it. We believe that the presence of the Char-based decoder is beneficial to capture into the encoder output gender information, which is then also exploited by the BPE-based decoder. As the encoder outputs are richer, overall translation quality is also slightly improved (Table 7). This finding is in line with other work (Costa-jussà et al., 2020b), which proved a strict relation between gender accuracy and the amount of gender information retained in the intermediate representations (encoder outputs).

Overall, following these considerations, we posit that target segmentation can directly influence the gender information captured in the encoder output. In fact, since the Char and BPE decoders do not interact with each other in the Multitask model, the gender accuracy gains of the BPE decoder cannot be attributed to a better ability of a segmentation method in rendering the gender information present in the encoder output into the translation.

Our results pave the way for future research on the creation of richer encoder outputs, disclosing the importance of target segmentation in extracting gender-related knowledge. With this work, we have taken a step forward in ST for English-French and English-Italian, pointing at plenty of new ground to cover concerning *how to* split for different language typologies. As the motivations of this inquiry clearly concern MT as well, we invite novel studies to start from our discoveries and explore how they apply under such conditions, as well as their combination with other bias mitigating strategies.

7 Conclusion

As the old IT saying goes: *garbage in, garbage out*. This assumption underlies most of current attempts to address gender bias in language technologies. Instead, in this work we explored whether technical choices can exacerbate gender bias by focusing on

---

[15]The Char scores are not reported, as they are not enhanced compared to the base Char encoder-decoder model.

the influence of word segmentation on gender translation in ST. To this aim, we compared several word segmentation approaches on the target side of ST systems for English-French and English-Italian, in light of the linguistic gender features of the two target languages. Our results show that tokenization does affect gender translation, and that the higher BLEU scores of state-of-the-art BPE-based models come at cost of lower gender accuracy. Moreover, first analyses on the behaviour of segmentation techniques found that improved generation of gender forms could be linked to the proper isolation of the morpheme that encodes gender information, a feature which is attained by character-level split. Finally, we propose a multi-decoder approach to leverage the qualities of both BPE and character splitting, improving both gender accuracy and BLEU score, while keeping computational costs under control.

Acknowledgments

8 Ethic statement[17]

In compliance with ACL norms of ethics, we wish to elaborate on *i)* characteristics of the dataset used in our experiments, *ii)* the study of gender as a variable, and *iii)* the harms potentially arising from real-word deployment of direct ST technology.

As already stated, in our experiments we rely on the training data from the TED-based MuST-C corpus[18] and its derived evaluation benchmark, MuST-SHE. Although precise information about various sociodemographic groups represented in the data are not fully available, based on impressionistic overview and prior knowledge about the nature of TED talks it is expected that the speakers are almost exclusively adults (over 20), with different geographical backgrounds. Thus, such data are likely to allow for modeling a range of English varieties of both native and non-native speakers.

As regards gender, from the data statements (Bender and Friedman, 2018) of the used corpora, we know that MuST-C training data are manually annotated with speakers' gender information[19] based on the personal pronouns found in their publicly available personal TED profile. As reported in its release page,[20] the same annotation process applies to MuST-SHE as well, with the additional check that the indicated (English) linguistic gender forms are rendered in the gold standard translations. Hence, information about speakers' preferred linguistic expressions of gender are transparently validated and disclosed. Overall, MuST-C exhibits a gender imbalance: 70% vs. 30% of the speakers referred by means of *he/she* pronoun, respectively. Instead, allowing for a proper cross-gender comparison, they are equally distributed in MuST-SHE.

Accordingly, when working on the evaluation of speaker-related gender translation for MuST-SHE category (1), we proceed by solely focusing on the rendering of their linguistic gender expressions. As per (Larson, 2017) guidelines, no assumptions about speakers' self determined identity (GLAAD, 2007) – which cannot be directly mapped from pronoun usage (Cao and Daumé III, 2020; Ackerman, 2019) – has been made. Unfortunately, our experiments only account for the binary linguistic forms represented in the used data. To the best of our knowledge, ST natural language corpora going beyond binarism do not yet exist,[21] also due to the fact that gender-neutralization strategies are still object of debate and challenging to fully implement in languages with grammatical gender (Gabriel et al., 2018; Lessinger, 2020). Nonetheless, we support the rise of alternative neutral expressions for both languages (Shroy, 2016; Gheno, 2019) and point towards the development of non-binary inclusive technology.

Lastly, we endorse the point made by Gaido et al. (2020). Namely, direct ST systems leveraging speaker's vocal biometric features as a gender cue have the capability to bring real-world dangers, like the categorization of individuals by means of biological essentialist frameworks (Zimman, 2020). This is particularly harmful to transgender individuals, as it can lead to misgendering (Stryker, 2008) and diminish their personal identity. More generally, it can reduce gender to stereotypical ex-

---

[17]Extra space after the 8th page allowed for ethical considerations – see https://2021.aclweb.org/calls/papers/

[18]https://ict.fbk.eu/must-c/

[19]https://ict.fbk.eu/must-speakers/
[20]https://ict.fbk.eu/must-she/
[21]In the whole MuST-C training set, only one speaker with *they* pronouns is represented.

pectations about how masculine or feminine voices should sound. Note that, we do not advocate for the deployment of ST technologies *as is*. Rather, we experimented with unmodified models for the sake of hypothesis testing without adding variability. However, our results suggest that, if certain word segmentation techniques better capture correlations from the received input, such capability could be exploited to redirect ST attention away from speakers' vocal characteristics by means of other information provided.

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a Journal of General linguistics*, 4(1).

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana.

Duygu Ataman and Marcello Federico. 2018. An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–110, Boston, MA. Association for Machine Translation in the Americas.

Duygu Ataman, Orhan Firat, Mattia A. Di Gangi, Marcello Federico, and Alexandra Birch. 2019. On the Importance of Word Boundaries in Character-level Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 187–193, Hong Kong. Association for Computational Linguistics.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108:331–342.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On Using SpecAugment for End-to-End Speech Translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, China.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the Linguistic Representational Power of Neural Machine Translation Models. *Computational Linguistics*, 46(1):1–52.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Ruha Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

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

Yang T. 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.

Roldano Cattoni, Mattia A. Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. MuST-C: A Multilingual Corpus for end-to-end Speech Translation. Computer Speech & Language Journal. Doi: https://doi.org/10.1016/j.csl.2020.101155.

Marina Chini. 1995. *Genere grammaticale e acquisizione: Aspetti della morfologia nominale in italiano L2*. Franco Angeli.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

John W. Chotlos. 1944. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany. Association for Computational Linguistics.

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

Greville G. Corbett. 2013. *The Expression of Gender*. De Gruyter.

Marta R. Costa-jussà. 2019. An analysis of Gender Bias studies in Natural Language Processing. *Nature Machine Intelligence*, 1:495–496.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2020a. Evaluating Gender Bias in Speech Translation. *arXiv preprint arXiv:2010.14465*.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2020b. Gender Bias in Multilingual Neural Machine Translation: The Architecture Matters. *arXiv preprint arXiv:2012.13176*.

Marta R. Costa-jussà and José A. R. Fonollosa. 2016. Character-based Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany. Association for Computational Linguistics.

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.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*, 17(2):94–100.

Kate Crawford. 2017. The Trouble with Bias. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*, Long Beach, California.

Mathias Creutz and Krista Lagus. 2005. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. *International Symposium on Computer and Information Sciences*.

Caroline Criado-Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Penguin Random House, London, UK.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Online. Association for Computational Linguistics.

Mattia A. Di Gangi, Marco Gaido, Matteo Negri, and Marco Turchi. 2020. On Target Segmentation for Direct Speech Translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 137–150, Online. Association for Machine Translation in the Americas.

Mattia A. Di Gangi, Matteo Negri, Roldano Cattoni, Dessi Roberto, and Marco Turchi. 2019. Enhancing transformer for end-to-end speech-to-text translation. In *Machine Translation Summit XVII*, pages 21–31, Dublin, Ireland. European Association for Machine Translation.

Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism*. MIT Press, London, UK.

Mohamed Elaraby and Ahmed Zahran. 2019. A Character Level Convolutional BiLSTM for Arabic Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 274–278, Florence, Italy. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. From Bilingual to Multilingual Neural Machine Translation by Incremental Training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 236–242, Florence, Italy. Association for Computational Linguistics.

Ute Gabriel, Pascal M. Gygax, and Elisabeth A. Kuhn. 2018. Neutralising linguistic sexism: Promising but cumbersome? *Group Processes & Intergroup Relations*, 21(5):844–858.

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.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, page 3–9, New York, NY, USA. Association for Computing Machinery.

Vera Gheno. 2019. *Femminili Singolari: il femminismo è nelle parole*. Effequ.

GLAAD. 2007. Media Reference Guide - Transgender.

Hila Gonen and Kellie Webster. 2020. Automatically Identifying Gender Issues in Machine Translation using Perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Pascal M. Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, 10:1604.

Christian Hardmeier, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. 2021. How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP. *arXiv preprint arXiv:2104.03026*.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3042–3051, Online. Association for Computational Linguistics.

Marlis Hellinger and Heiko Motschenbacher. 2015. *Gender Across Languages. The Linguistic Representation of Women and Men*, volume IV. John Benjamins, Amsterdam, the Netherlands.

Charles F. Hockett. 1958. *A Course in Modern Linguistics*. Macmillan, New York, New York.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Brian Larson. 2017. Gender as a variable in Natural-Language Processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Enora Lessinger. 2020. The Challenges of Translating Gender in UN texts. In Luise von Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*. Routledge, New York, NY, USA.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. Character-based Neural Machine Translation. *arXiv preprint arXiv:1511.04586*.

Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 92–105, New York, NY, USA. PMLR.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, et al. 2018. XNMT: The eXtensible Neural Machine Translation Toolkit. In *Proceedings of AMTA 2018*, pages 185–192, Boston, MA.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech 2019*, pages 2613–2617, Graz, Austria.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Tomasz Potapczyk and Pawel Przybysz. 2020. SR-POL's System for the IWSLT 2020 End-to-End Speech Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 89–94, Online. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2020. Assessing Gender Bias in Machine Translation: a Case Study with Google Translate. *Neural Computing and Applications*, 32(10):6363–6381.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of 4th International Conference on Learning Representations ICLR*, San Juan, Puerto Rico.

Argentina A. Rescigno, Johanna Monti, Andy Way, and Eva Vanmassenhove. 2020. A Case Study of Natural Gender Phenomena in Translation: A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. In *Proceedings of the Workshop on the Impact of Machine Translation (iMpacT 2020)*, pages 62–90, Online. Association for Machine Translation in the Americas.

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C. Lipton. 2020. Decoding and Diversity in Machine Translation. In *Proceedings of the Resistance AI Workshop at 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada.

Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Danielle Saunders, 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, Online. 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*.

Elmar Schafroth. 2003. *Gender in French: Structural properties, incongruences and asymmetries*. John Benjamins Publishing, Amsterdam, The Netherlands.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, 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: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Deven S. Shah, Hansen A. Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Alyx J. Shroy. 2016. Innovations in gender-neutral French: Language practices of nonbinary French speakers on Twitter. *Ms., University of California, Davis*.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Consistent Transcription and Translation of Speech. *Transactions of the Association for Computational Linguistics (TACL)*, 8:695–709.

Artūrs Stafanovičs, Mārcis Pinnis, and Toms Bergmanis. 2020. Mitigating Gender Bias in Machine Translation with Target Gender Annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

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

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.

Susan Stryker. 2008. Transgender history, homonormativity, and disciplinarity. *Radical History Review*, 2008(100):145–157.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada, United States.

Mildred C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press.

Nneka Umera-Okeke. 2012. Linguistic sexism: an overview of the English language in everyday discourse. *AFRREV LALIGENS: An International Journal of Language, Literature and Gender Studies*, 1(1):1–17.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS)*, pages 5998–6008, Long Beach, California.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proceedings of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden.

Lal Zimman. 2020. Transgender language, transgender moment: Toward a trans linguistics. In Kira Hall and Rusty Barrett, editors, *The Oxford Handbook of Language and Sexuality*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Models

Our direct ST models are built with the Fairseq library (Ott et al., 2019) and are based on the Transformer (Vaswani et al., 2017). They have 11 encoder and 4 decoder layers (Potapczyk and Przybysz, 2020). The encoder layers are preceded by 2

	en-fr	en-it
BPE	60M	60M
Char	52M	52M
DPE	60M	60M
Morfessor	79M	76M
LMVR	74M	72M
BPE&Char	77M	77M

Table 9: Number of parameters (in millions). For BPE&Char, the reported number is the total of the training parameters, but at inference time only one decoder is used, so the size is the same of BPE.

3x3 convolutional layers with 64 filters that reduce the input sequence length by a factor of 4 and their attention weights are added a logarithmic distance penalty (Di Gangi et al., 2019). The models are optimized on label smoothed cross-entropy (Szegedy et al., 2016) – the smoothing factor is 0.1 – with Adam using β_1=0.9, β_2=0.98 and the learning rate is linearly increased during the warm-up phase (4k iterations) up to the maximum value 5×10^{-3}, followed by decay with inverse square root policy. The dropout is set to 0.2 and each mini-batch consists of 8 sentences, while the update frequency is 8. The source audio is augmented with SpecAugment (Park et al., 2019; Bahar et al., 2019) that is applied with probability 0.5 by masking two bands on the frequency axis (with 13 as maximum mask length) and two on the time axis (with 20 as maximum mask length).

The systems are trained on MuST-C (Cattoni et al., 2020). We filtered from the training set all the samples whose audio length is higher than 20s. So to avoid rewarding models' potentially biased behaviour, as a validation set we rely on the MuST-C gender-balanced dev set (Gaido et al., 2020). The target text was tokenized with Moses.[22] We normalized audio per-speaker and extracted 40 features with 25ms windows sliding by 10ms with XNMT[23] (Neubig et al., 2018).

Trainings are stopped after 5 epochs without improvements on the validation loss and we average 5 checkpoints around the best on the validation set. They were performed on 8 K80 GPUs and lasted 2-3 days.

---

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

[23]https://github.com/neulab/xnmt